

Laura Valtonen

**UNSUPERVISED MACHINE LEARNING  
FOR EVENT CATEGORIZATION IN  
BUSINESS INTELLIGENCE**

Faculty of Engineering and Natural Sciences  
Master of Science Thesis  
May 2019

# ABSTRACT

Laura Valtonen: Unsupervised Machine Learning for Event Categorization in Business Intelligence  
Master of Science Thesis  
Tampere University  
Industrial Engineering and Management  
May 2019

---

The data and information available for business intelligence purposes is increasing rapidly in the world. Data quality and quantity are important for making the correct business decisions, but the amount of data is becoming difficult to process. Different machine learning methods are becoming an increasingly powerful tool to deal with the amount of data. One such machine learning approach is the automatic annotation and location of business intelligence relevant actions and events in news data.

While studying the literature of this field, it however became clear, that there exists little standardization and objectivity regarding what types of categories these events and actions are sorted into. This was often done in subjective, arduous manners. The goal of this thesis is to provide information and recommendations on how to create more objective, less time consuming initial categorizations of actions and events by studying different common unsupervised learning methods for this task.

The relevant literature and theory to understand the followed research and methodology is studied. The context and evolution of business intelligence to today is considered, and specially its relationship to the big data problem of today is studied. This again relates to the fields of machine learning, artificial intelligence, and especially natural language programming. The relevant methods of these fields are covered to understand the taken steps to achieve the goal of this thesis. All approaches aided in understanding the behaviour of unsupervised learning methods, and how it should taken into account in the categorization creation.

Different natural language preprocessing steps are combined with different text vectorization methods. Specifically, three different text tokenization methods, plain, N-gram, and chunk tokenizations are tested with two popular vectorization methods: bag-of-words and term frequency inverse document frequency vectorizations. Two types of unsupervised methods are tested for these vectorizations: Clustering is a more traditional data subcategorization process, and topic modelling is a fuzzy, probability based method for the same task. Out of both learning methods, three different algorithms are studied by the interpretability and categorization value of their top cluster or topic representative terms. The top term representations are also compared to the true contents of these topics or clusters via content analysis.

Out of the studied methods, plain and chunk tokenization methods yielded the most comprehensible results to a human reader. Vectorization made no major difference regarding top term interpretability or contents and top term congruence. Out of the methods studied, K-means clustering and Latent Dirichlet Allocation were deemed the most useful in event and action categorization creation. K-means clustering created a good basis for an initial categorization framework with congruent result top terms to the contents of the clusters, and Latent Dirichlet Allocation found latent topics in the text documents that provided serendipitous, fruitful insights for a category creator to take into account.

Keywords: business intelligence, information extraction, automated content analysis, clustering, topic modelling

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

# TIIVISTELMÄ

Laura Valtonen: Ohjaamaton koneoppiminen tapahtumakategorisoinnissa liiketoimintatiedon hyödyntämisessä  
Diplomityö  
Tampereen yliopisto  
Tuotantotalous  
Toukokuu 2019

---

Datan määrä, josta on johdettavissa liiketoiminnassa hyödynnettävää tietoa, kasvaa maailmassa tällä hetkellä nopeasti. Tämän datan laajuus ja laatu ovat tärkeitä oikeiden päätösten tekemisen kannalta, mutta kasvavaa määrää on yhä hankalampi käsitellä ihmisten tai tietokoneiden toimesta. Erilaiset koneoppimismenetelmät ovat kehityksessä tehokkaiksi työkaluiksi toimia tämän suuren datamäärän kanssa. Yksi tällainen tapa on automaattinen tapahtumien löytäminen ja kategorisointi uutisdatasta.

Alan kirjallisuutta tutkiessa kävi selväksi, että tällaisen tapahtumakategorisoinnin tekeminen ei ole mitenkään standardisoitua, vaan usein työläs ja subjektiivinen prosessi. Tämän opinnäytetyön tavoite on tarjota tietoa ja suosituksia siitä, kuinka objektiivisempia ja vähemmän työläisiä alustavia kategorisointeja voitaisiin luoda ohjaamattomia koneoppimismenetelmiä hyödyntäen.

Työssä käydään läpi tarvittava teoriaa ja kirjallisuutta liiketoimintatiedon hyödyntämisen evoluutiota ja tilannetta tänään arvioidaan erityisesti datan määrän kasvun valossa. Datan määrä taas liittyy olennaisesti koneoppimisen, tekoälyn, ja erityisesti luonnollisen kielen kanssa tietokoneellisesti toimimisen aloihin. Tehdyn työn kannalta, näiden alojen oleelliset käsitteet ja menetelmät käydään läpi. Luonnollisen kielen tietokoneella käsittelyn perusteet avataan, sekä tutustutaan tarkemmin muutamaa eri suosittuun aihehallinnus- ja klusterointimetodiin: aihehallinnuksesta Latent Dirichlet Allocation, Latent Semantic Indexing, Hierarchical Dirichlet Process, ja klusterointista K-means klusterointi, Affinity Propagation, sekä Mean Shift -algoritmit.

Tutkituista ohjaamattoman koneoppimisen menetelmistä K-means -klusterointi ja LDA-aihehallinnus tulivat hyödyllisimmiksi tapahtumakategorisoinnin luonnissa. K-means klusterointi loi hyviä alustavia kategorisointeja, joissa metodin termiesitykset olivat yhdenmukaisia tulosten sisällön kanssa. LDA puolestaan tuotti yllättävämpiä termiesityksiä, joiden anti oli herättää lukijassa ymmärryksen hetkiä siitä, miten kategorisointia voisi kehittää. Molemmat tavat tuottivat ymmärrystä ohjaamattoman oppimisen käytöksestä, ja siitä, kuinka se tulisi huomioida kategorisointia tehdessä. Yksi tutkituista tekstin osittelutavoista tuotti vaikeammin tulkittavia tuloksia, mutta numerien esitysten välillä ei esiintynyt merkittävää eroavaisuutta.

Avainsanat: liiketoimintatiedon hyödyntäminen, tiedon eristäminen, automatisoitu sisällönanalyysi, klusterianalyysi, aihehallinnus

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -ohjelmalla.

## PREFACE

The premise for this thesis was set by a larger research goal regarding the automated information retrieval, event and action extraction, described in this thesis. However, while pursuing this, it was noticed that the steps taken in this thesis were necessary for me to begin to pursue the previously set goals. I am grateful for the learning opportunity of this thesis, and hope to be able to employ everything I have now learnt to the described, possibly larger goals. I am grateful for the guidance from my instructor Prof. Saku Mäkinen. I am especially grateful of the patience regarding the difficulties of determining an appropriate scope for this thesis.

Writing this thesis was not an easy task. When I began my studies as a mathematics student, I did not imagine I would end up writing a thesis overlapping the fields of machine learning and industrial engineering and management. Especially machine learning was very much a new field of science for me. I am, of course, very grateful to my fiancé for all of his support and for putting up with me coding late into the nights at times when this thesis was really putting up a fight.

Of course, as is custom, I thank my family and friends for support and always believing in me. On a more concrete level, I am thankful to the students and teachers of the Challenge Based Innovation -course for being interested and inspired by the progress of this thesis at times when I was feeling I was making no progress. Especially I would like to thank my co-assistant Helinä, for making sure I ate lunch on most days.

Tampere, 26th May 2019

Laura Valtonen

# CONTENTS

1	Introduction . . . . .	1
2	Theoretical Background . . . . .	7
2.1	Business Intelligence . . . . .	7
2.2	The Big Data Problem . . . . .	11
2.2.1	Content Analysis . . . . .	11
2.2.2	The Situation of Data Today . . . . .	12
2.3	Navigating the Methodology of Business Intelligence . . . . .	14
2.4	Machine Learning . . . . .	15
2.5	Classification . . . . .	16
2.5.1	Evaluating Goodness . . . . .	18
2.6	Clustering . . . . .	19
2.6.1	K-Means Clustering . . . . .	20
2.6.2	Affinity Propagation . . . . .	21
2.6.3	Mean Shift Clustering . . . . .	22
2.6.4	Evaluating Goodness . . . . .	23
2.7	Natural Language Processing and Computational Linguistics . . . . .	24
2.7.1	The Text Preprocessing Process . . . . .	25
2.7.2	Vectorization . . . . .	28
2.8	Topic Modelling . . . . .	31
2.8.1	Latent Semantic Indexing . . . . .	32
2.8.2	Latent Dirichlet Allocation . . . . .	33
2.8.3	Hierarchical Dirichlet Process . . . . .	34
2.8.4	Evaluating Goodness . . . . .	35
2.8.5	Perplexity . . . . .	35
2.9	The Research Problem . . . . .	37
3	Data and Methodology . . . . .	38
3.1	Data Gathering . . . . .	38
3.2	Data Clean-up . . . . .	39
3.3	Categorization of News Data . . . . .	40
3.4	Unsupervised methods . . . . .	41
3.4.1	Data Clean-up for Clustering and Topic Modelling . . . . .	42
3.4.2	N-grams, Chunks, and Vectorization . . . . .	43
3.4.3	Clustering . . . . .	44
3.4.4	Topic Modelling . . . . .	44
3.4.5	Evaluation Methods . . . . .	45
3.4.6	Content Analysis . . . . .	45

4	Results . . . . .	47
4.1	Dimensionality and Run Times . . . . .	47
4.2	Coherences . . . . .	47
4.3	Topics and Clusters - Top Terms . . . . .	51
4.3.1	Plain Tokenizer with Bag-of-words Vectorization . . . . .	51
4.3.2	Plain Tokenizer with TF-IDF Vectorization . . . . .	52
4.3.3	N-gram Tokenizer with Bag-of-words Vectorization . . . . .	54
4.3.4	N-gram Tokenizer with TF-IDF Vectorization . . . . .	55
4.3.5	Chunk Tokenizer with Bag-of-words Vectorization . . . . .	58
4.3.6	Chunk Tokenizer with TF-IDF Vectorization . . . . .	59
4.3.7	Content Analysis . . . . .	60
5	Discussion and Conclusions . . . . .	64
5.1	Reliability and Validity . . . . .	64
5.2	Discussions . . . . .	65
5.3	Conclusions . . . . .	69
5.4	Limitations and Future Research . . . . .	72
	References . . . . .	74
	Appendix A Top Term and Content Analysis Congruence Evaluations . . . . .	81

## LIST OF FIGURES

2.1	Artificial intelligence and its subfields . . . . .	15
2.2	A deep neural network with four representation levels . . . . .	17
2.3	Part-of-speech tag examples . . . . .	27
4.1	Coherence values for algorithms for different tokenizer and vectorizer combinations - Sample 1 . . . . .	48
4.2	Coherence values for algorithms for different tokenizer and vectorizer combinations - Sample 2 . . . . .	49
4.3	Coherence values for algorithms for different tokenizer and vectorizer combinations - Sample 3 . . . . .	50

## LIST OF TABLES

2.1	The knowledge discovery in databases (KDD) process . . . . .	13
2.2	A confusion matrix example . . . . .	19
2.3	K-means clustering algorithm . . . . .	21
2.4	Affinity Propagation clustering algorithm . . . . .	22
2.5	Mean Shift clustering algorithm . . . . .	23
2.6	Short sentences for use in text vectorization examples . . . . .	29
2.7	Bag-of-words representation of example sentences . . . . .	30
2.8	TF-IDF representation of example sentences . . . . .	31
3.1	Keywords used for data collection . . . . .	39
3.2	Data events and actions as interpreted by the annotator . . . . .	40
4.1	Plain tokenizer with count vectorizer. Top 15 clusters and topics . . . . .	53
4.2	Plain tokenizer with TF-IDF vectorizer. Top 15 clusters and topics. . . . .	55
4.3	N-gram tokenizer with count vectorizer. Top 15 clusters and topics . . . . .	56
4.4	N-gram tokenizer with TF-IDF vectorizer. Top 15 clusters and topics. . . . .	57
4.5	Chunk tokenizer with count vectorizer. Top 15 clusters and topics . . . . .	58
4.6	Chunk tokenizer with TF-IDF vectorizer. Top 15 clusters and topics . . . . .	60
4.7	Congruence of top terms to content analysis per algorithm . . . . .	61
4.8	Amounts of overlapping topics and clusters . . . . .	62
5.1	Relative time requirements and potential suitable use cases of the different clustering and topic modelling approaches - a conclusion . . . . .	71
A.1	K-means content analysis and clustering top terms congruence with bag-of-words vectorization . . . . .	82
A.2	K-means content analysis and clustering top terms congruence with TF-IDF vectorization . . . . .	90
A.3	LDA content analysis and clustering top terms congruence with bag-of-words vectorization . . . . .	100
A.4	LDA content analysis and clustering top terms congruence with TF-IDF vectorization . . . . .	109

## LIST OF SYMBOLS AND ABBREVIATIONS

$\alpha$	positive scalar
<b>a</b>	availability in Affinity Propagation
<b>AI</b>	artificial intelligence
<b>AP</b>	Affinity Propagation
$\beta_n$	$n^{th}$ topic in Hierarchical Dirichlet Process
<b>BI</b>	business intelligence
<b>BOW</b>	bag-of-words
$c(x_i)$	cluster containing datapoint $x_i$
$\text{coh}_{u\_mass}$	coherence measure
$\text{coh}_{UCI}$	coherence measure
<b>CI</b>	competitive intelligence
<b>CL</b>	computational linguistics
<b>CNN</b>	convolutional neural network
$D$	corpus
$D$	singular vector in SVD for LSI
$d$	a document consisting of tokens
<b>DM</b>	data mining
<b>DP</b>	Dirichlet process
$\epsilon$	small positive scalar value
<b>ESS</b>	squared euclidean distance
$F_n$	false negatives in supervised machine learning
$F_p$	false positives in supervised machine learning
$f$	frequency
$G_0$	base topic distribution for Hierarchical Dirichlet Processing
$G_d$	topic distribution for Hierarchical Dirichlet Processing
$\gamma$	positive scalar
$H$	symmetric Dirichlet distribution
<b>HDP</b>	Hierarchical Dirichlet Process
$i$	index marker

<i>idf</i>	inverse document frequency
IE	information extraction
IR	information retrieval
<i>j</i>	index marker for topics
<i>K</i>	kernel function in Mean Shift
<i>k</i>	reduced rank of matrix
<i>K</i>	number of clusters and topics
KDD	knowledge discovery in databases
KM	K-means clustering
<i>L</i>	set of datapoints
$\lambda$	damping factor in Affinity Propagation
LDA	Latent Dirichlet Allocation
LSI	Latent Semantic Indexing
<i>M</i>	index marker
max	maximum
min	minimum
ML	machine learning
MS	Mean Shift
<i>N</i>	index marker
<i>N</i>	full rank of matrix
<i>N</i>	neighbourhood of datapoint in Mean Shift
<i>n<sub>d</sub></i>	number of documents
<i>n</i>	index marker for topics
NLP	natural language processing
NN	neural network
$\phi$	word distribution
<i>P</i>	precision
<i>p</i>	probability distribution
<i>P</i>	probability
PMI	pointwise mutual information
POS	part-of-speech
PULS	Pattern-based Understanding and Learning System
<i>r</i>	responsibility in Affinity Propagation
<i>R</i>	recall

RNN	recurrent neural network
ROI	return of investment
$S$	singular matrix in SVD for LSI
$s$	similarity
SIE-OBI	streaming information extraction platform for operational business intelligence
SVD	matrix singular value decomposition
$T$	singular vector in SVD for LSI
$k$	index marker
$t$	iteration index
$\theta$	topic distribution
$tf$	token frequency
$t$	a token, such as a singular word or a chunk of words
$T$	transpose
$T_n$	true negatives in supervised machine learning
$T_p$	true positives in supervised machine learning
TF-IDF	term frequency inverse document frequency
TV	television
$V$	vocabulary, made of words
VSM	vector space model
$w$	word
WSD	word sense disambiguation
$x_k$	centroid candidate datapoint in Affinity Propagation
$\bar{x}_k$	centroid of cluster $k$
$x_i$	$i^{th}$ datapoint in set
$X$	full rank data matrix
$\hat{X}$	reduced rank data matrix
$z_j$	$j^{th}$ topic

# 1 INTRODUCTION

Today, the world is overflowing with data, be it structured data from an organization's inner processes and sensors, or an unstructured mixed form of data from the internet and social media. The quantity is increasingly intimidating to delve into, yet it is all the more important to do so. In a fast paced environment, it becomes vital for a company's survival to be able to turn this available data into actionable, quality intelligence. In other words, making use of all this easily available, gargantuan, fast paced data is a critical business intelligence task.

Business Intelligence is not a clearly defined term, and various differing definitions are found in literature. Most, however, at their core, define business intelligence to be a field of data review and analytic methods geared towards making decision making more effective in organizations. The field borrows from a variety of other fields, most of which can be considered as principles of their own, but act as tools for business intelligence. (Cvitaš 2010) Wani and Jabin (2018) summarize, in their brief review, different business intelligence methods to be statistical analysis, data mining and analytics, predictive modelling and analysis, and big data analytics and text analytics.

The data available for business intelligence purposes is becoming very large and unstructured, due to emerging big data (Wani and Jabin 2018), which has made rule based approaches to data analytics nearly obsolete. Mostly, nowadays, the go-to approach for business intelligence data analytics is some form of machine learning. (Cvitaš 2010)

Machine learning is generally split into supervised and unsupervised methods, out of which supervised methods require human processed data, and are therefore more time-consuming than unsupervised methods, which do not require such pre-processing of data (Friedman, Hastie and Tibshirani 2001). While unsupervised learning methods are only one of many options on how to approach business intelligence, they are becoming an increasingly appealing method due to the lack of required manual labour (Cvitaš 2010). One reason for this may be that the timeliness of business intelligence information is very important to decision makers (Wani and Jabin 2018), and the use of unsupervised methods is faster due to the lack of the slow human processing phase.

A clear example of a commonly utilized unsupervised learning method is clustering, which is listed (Chen, Chiang and Storey 2012) as a foundational technology for big data analytics. Clustering can be used to segment groups of data differing from other groups, such as users, customers, technologies, product review sentiments based on vocabulary, and

many more. Other foundational technologies listed are genetic algorithms for data analytics, and topic models for text analytics - a more specific field of data analytics, but increasingly important since the majority of unstructured data is in text format (Chen, Chiang and Storey 2012). Furthermore, very different unsupervised learning method applications for business intelligence purposes exist – for example from the very general-purpose data variable association algorithms for decision support (Orriols-Puig et al. 2013), to more specific purposes, such as the political risk analysis method for choosing business locations (Herrero, Corchado and Jiménez 2011).

One way to go about obtaining business intelligence data and information is called information extraction or retrieval. It refers to methods that, as the name might suggest, go through data and pick out or turn it into relevant intelligence for a company. Today, there exist systems available for organizations to do this. Examples include systems such as PULS (Du, Pivovarova and Yangarber 2016) and SIE-OBI (Castellanos et al. 2012).

PULS extracts business activity - events - information from online news, press releases, and social media from the following categories: acquisition, investment, order, marketing, product launch, merger, leadership change, bankruptcy, lawsuit, business closing, layoff, product recall, and accident (Du, Pivovarova and Yangarber 2016). Also other information regarding for example the parties involved, location, and industry is extracted and presented. What is interesting however, is that there is provided no reasoning whatsoever for these categories for having been chosen. Were they determined to be exhaustive by literature, or perhaps by experts? What about patent grants or applications? (Du, Pivovarova and Yangarber 2016) Clearly, there exists some more work to do regarding basing or referencing these event categories on something.

SIE-OBI allows for the comparison of external news feed events and social media reactions to an organization's inner data and information to locate relevant events in real time. Example events such as political instability, currency fluctuations, acquisitions, mergers, changes in law, and natural disasters are provided. A very different set compared to the PULS categorization. In this case however, the set of interesting document categories or entities has to be predefined per user of the system before use. This is a very heavy manual labor annotation task that requires hours of work. (Castellanos et al. 2012)

On top of requiring a large amount of work compared to the built-in categories of PULS, SIE-OBI is prone to human error and forgetfulness in determining the document categories of interest. There exists clearly a trade-off of customizability and required effort, but still both approaches have their obstacles. Now these have been only two examples of approaching this type of an event related information extraction task, but lack of clarity regarding the creation the event categorization did not end here.

On the spring 2019 implementation of the research methodology course at Tampere University, a possible course assignment was an event study of a company, in which a group was supposed to go through a company's history and news and categorize events, and from these events, draw conclusions about a company's strategy. As a reference frame-

work for categorization was the resource based categorization of events (Morgan and Hunt 1999) into financial, legal, physical, human, organizational, relational, and informational resources. The groups who did this assignment expressed their frustration of not finding other clear reference frameworks for categorization. Groups presenting at the course seminar had chosen the previous recommended framework. (Litovuo 2019) Such a broad categorization would not benefit event and action extraction and categorization on a more detailed level, for example compared to the PULS "accident" category.

Google Scholar, Scopus, and Web of Science were searched in hopes of other event categorization frameworks for comparison, but the theme of generality persisted. While event definition was very much unambiguous in the studied literature - "something that happens at a certain time and place" – the types of events and their categories varied very wildly (Wei and Lee 2004; Zhou, Chen and He 2015). For example, very general event categories such as "government", "law", "business", "music", "sports", and "TC" were used at times (Zhou, Chen and He 2015), while others used categories such as "domestic business", "domestic arts and education", "foreign affairs", "domestic finance", "domestic health", "Taiwan local news", "Taiwan sports", "domestic military", "domestic politics", "Taiwan stock markets", "domestic travel", and "weather report" (Wu, Tsai, Hsu et al. 2003). Clearly, the categorizations are not compatible. However, what was similar was that both cases stated that the categorization was "chosen" or "empirically set". (Wu, Tsai, Hsu et al. 2003; Zhou, Chen and He 2015)

Some references do exist for doing this event categorization, but as the previous examples demonstrate, they are not systematically followed, and seem too general and abstract for event and action categorization. One such categorization framework with literary backup, is by the Linguistic Data Consortium: elections, scandals, legal cases, natural disasters, accidents, violence or war, science discoveries, financial news, laws, sports news, political meetings, celebrity and human interest news, and miscellaneous news. (Cordeiro and Gama 2016)

Wei and Lee (2004) had come to the same conclusion that traditional event extraction techniques do not appear to support direct event categorization, and proposed a methodology where similar event groups are created based on event news with known event topics. These event topics could be used in the future for categorizing new events. The results included event topics such as "airplane crash", "adjustment of interest rate", "business merger", "business partnership", and "computer virus". These are obviously more detailed and informative regarding the nature of the event than the previous categorizations. The methodology implies that the event topics were human annotated from a dataset, which makes it a subjective supervised learning approach. This approach is similar to what is called content analysis, which is explained in chapter 2.2.1. These literature observations further reinforced the considerations raised by PULS and SIE-OBI: possibly some objective framework could be beneficial for creating event categorizations quicker with a computerized unsupervised approach. Some attempts have been made at this automated content analysis (Altaweel, Bone and Abrams 2019), but they are more

focused on specialized events with specific actors rather than an overarching categorization applicable to multiple different actors (X. Liu et al. 2016), and are not set in a business intelligence context.

In the case of text data, basically this type of an approach would mean inputting a large collection of the data available to an algorithm that creates smaller subsets of the data based on the used language - an unsupervised learning approach. The algorithm may, for instance, find that one subcategory often includes the term "application", and another one "appointment", hinting at different sorts of events and actions found in news related to business. How this all works is covered in detail in chapters 2 and 3 of this thesis.

Now of course, this may not be a perfect solution to determining the event categories to use, but it does provide additional information to human evaluation. The categories found by a computer can help expand the set of the events determined by PULS (Du, Pivovarova and Yangarber 2016) for example, by finding events outside of them, enriching the categorization and providing some transparency. This is only one possibility on how unsupervised learning might enrich categorizations made by humans.

In the case of SIE-OBI (Castellanos et al. 2012), to ease the required human effort, the categorization done by a machine could help act as a base to build the categorization further to suit the means. It can also act as a reference in the case humans forget something, as they are prone to do. Furthermore, machines are not people and do not do things in the same way. This kind of a clustering test can help provide insights how a computer processes these types of documents and events. Information derived in such a way may come in useful when developing these systems of event information extraction. Some events that seem different to humans may seem similar to machines and vice versa. Unsupervised methods may help answer how this should be taken into account in these information extraction systems. Moreover, how does the way the data is processed and handled affect the way computers interpret it? There exist many unknowns regarding the utilization possibilities of unsupervised learning in event and action categorization creation. These points raise the two main research questions of this thesis:

- How can unsupervised machine learning methods be exploited in the creation of an action and event categorization framework for business intelligence purposes?
- How do these exploitation possibilities differ in different, common unsupervised method approaches?

The former question uses the term "exploit", because it allows for a broad range of possible findings. The term is used over "can", because "exploit" implies that some value and benefit is expected from the use. The typical methodology for subsectioning data without too much human intervention, unsupervised learning, yields sometimes these unforeseen serendipitous findings that require consideration (Ziegler 2012). Some utilization opportunities might come as a surprise, and should be taken into account. This question leaves room for these possibilities.

The latter question is in response to Denny and Spirling 2018, who emphasize that there

exists little research into how these unsupervised learning method results are affected by taken data preprocessing measures, especially since they differ by field. Therefore, to be able to answer the former question, the latter needs to be addressed in the context of business intelligence. The term "common" is included, since there exist various different methods, and it is not possible to study all of them in the context of this thesis, which is why some refining is needed.

To add to the potential benefits of subcategorizing with unsupervised approaches, they may help in preventing the issue of overfitting when constructing an event information extraction system for business intelligence. To clarify, unknown data can be very unevenly distributed in the sense that a set could have 80% events on patent applications and registrations. An information extraction system may learn to always predict a patent event and still display an accuracy of said 80%. Subcategorizing with unsupervised methods and studying the data with these means beforehand might help notice that event categories are of different representation sizes, which can be then taken into account appropriately.

Now this is all very exciting, but the task of subcategorization is not very straightforward. For instance, "this is all very exiting" could just now have been interpreted as sarcastic or not. A computer does not understand sarcasm, it would without a doubt categorize that sentence as of "positive sentiment" - correctly in this case. The typical approach to analyzing texts with a computer means splitting the document into a set of words, and a computer can not determine a difference between "letter" as in an alphabetical unit, or a written paper letter, without more context, as neither can a human. Dealing with these peculiarities of natural language by computational means is as a field of study called natural language processing. Natural language processing methods are important in the preprocessing choices made regarding datasets and their influences on the outcome of subcategorization.

This subcategorization can be approached from different angles, and in this thesis some common unsupervised text analysis methods, clustering and topic modelling to be exact, are considered; open source, easily implementable Python programming libraries for machine learning and topic modelling are applied. All methods were implemented as they are "out-of-the-box", and not optimized for the data specific to this thesis for reasons of generalisability. The outcomes of these approaches are difficult to evaluate, and meaningful results are typically a result of heuristic argumentation based on the outcomes of clustering or topic modelling (Friedman, Hastie and Tibshirani 2001). This is also the case in this thesis.

The two largest contributors to the limitations on reliability and validity in this thesis are that this thesis naturally can not cover all different possible approaches to answering the posed research questions, and that the very nature of evaluating unsupervised learning results is very subjective. Further limitations and the set scope for this thesis is considered in detail in the discussion and conclusions chapter 5. Regardless, it is clearly found in this thesis that unsupervised learning approaches can indeed aid in creating a slightly more objective, and potentially faster event categorization framework for a human

decision maker for business intelligence event and action extraction, and that different approaches can offer various different types of value into creating the framework. This thesis also makes recommendations based on the tested approaches on what one may wish to use for a specific outcome.

In the following chapters, the theoretical context and important concepts for understanding the research done in this thesis are presented next, in chapter 2. First, the context of business intelligence is explained further along with the state of the data available for business intelligence processes. Then the methodology relevant for business intelligence is explained in broader terms, after which the considered data subcategorization tasks of clustering and topic modelling, and their required preprocessing choices related to natural language processing are explained in greater detail. Evaluation methods of unsupervised learning methods are discussed after each relevant chapter to understand their complex nature.

In chapter three the acquired data and applied research methodology is explained in further detail: How the studied data was acquired, how it was studied and processed, how it was used for answering the research questions, what choices were made and why. After, the results of the done research are presented as they are, with created tables and literal explanations. Some limitations of the results are discussed also. Only after that their implications and intricacies are discussed further in the discussion and conclusions chapter, which summarizes the done research and its main findings and contributions.

## 2 THEORETICAL BACKGROUND

This chapter delves into the theoretical background required for understanding the research of this thesis. First, the context is set by defining business intelligence, the big data problem, and their relationship with each other. Afterwards, the methodology of these fields that is relevant for this thesis, is considered. The considered methodology is split into roughly three main parts: machine learning, natural language processing, and topic modelling. The machine learning section will study the differences of supervised machine learning methods and unsupervised machine learning methods, and how both types are applied in this thesis. Clustering, as a form of unsupervised learning, will be considered in even more detail, as it is central to the research done, and three used algorithms are explained in detail: K-means clustering, Affinity Propagation, and Mean Shift. Afterwards, the evaluation methods regarding clustering results are discussed, as they are not exactly straightforward.

After this, natural language processing and its methods such as tokenization and vectorization, along with other preprocessing choices associated with natural language processing, will be considered, to ensure the reader will be able to understand the premise for the choices made later on. Following this, topic modelling is covered, as it is closely related to natural language processing and often considered comparable to clustering (Zhou, Chen and He 2015; Pourvali, Orlando and Omidvarborna 2019). Similarly to clustering, three used methods, Latent Semantic Indexing, Latent Dirichlet Allocation, and Hierarchical Dirichlet Processing, are covered in detail, and the evaluation of topic models is discussed. Before moving on from this chapter to the detailed explanation of methodology, the research problem is reinstated - now better understandable due to the covered theory.

### 2.1 Business Intelligence

The basic techniques and methods implemented and analyzed in this thesis may be relevant in a very similar manner over a wide range of disciplines - data analytics tend to be interesting to all fields of research, but the focus and main point of view coming to this piece of research is that of business intelligence. As a term, business intelligence seems to evolve and vary in its exact meaning in relevant literature (Gibson et al. 2004). Jourdan, Rainer and Marshall (2008) use the definition originated by Vedder et al. (1999), according to which business intelligence (BI) is both a process and a product, in which

the process is comprised of methods and technologies used by organizations to develop information that aids organizations to “thrive in the global economy”, and the product is the developed information. This information is supposed to help predict the behavior of “competitors, suppliers, customers, technologies, acquisitions, markets, products and services, and the general business environment” (Vedder et al. 1999). Some journal articles only consider the “process” part of the previous definition as business intelligence, and the goal of BI not to be predictability, but long term sustainable competitive advantage, Atriwal et al. (2016) for instance complies with this view.

At times, “business intelligence” is used interchangeably with “competitive intelligence” (Vedder et al. 1999; Atriwal et al. 2016), and sometimes “competitive intelligence” is seen as a subcategory of business intelligence (Zheng, Fader and Padmanabhan 2012). During searches for relevant information and articles for this thesis, both terms were associated with relevant literature. Sassi et al. (2015) studied literature and found that no specific definition for competitive intelligence is likely to be found. For instance, Zheng, Fader and Padmanabhan (2012) describe competitive intelligence (CI) as the part of BI that is focused on gaining actionable knowledge about an organizations external competitive environment. Sassi et al. (2015) define competitive intelligence as “a systematic and ethical program for gathering, analyzing, and managing external information that can affect a company’s plans, decisions and operations”. This definition was originally coined by the Society of Competitive Intelligence Professionals (Sassi et al. 2015). These competitive intelligence definitions seem to be more in line with each other than the considered BI definitions.

Moreover, some articles, Chen, Chiang and Storey (2012) for instance, address the field as “business intelligence and analytics”. The term “analytics” came to describe the analytical side of business intelligence in the 2000’s (Davenport et al. 2006). This division of terms makes sense with the “both process and a product” view of BI. Overall, the clear definition of business intelligence, its relation to competitive intelligence, and what it comprises seems to be a “pick and mix” one. The most in line with the thought process throughout this thesis is this one from Wani and Jabin (2018): “Business intelligence relates to a technology-oriented process for analyzing data and presenting actionable information to help scientists, corporate executives, business managers, and other end users make more informed business decisions.”

This definition highlights the data analysis focused process to create not any, but actionable information and knowledge for any organization – not just a company. This thesis considers competitive intelligence to concern business intelligence processes that aim to produce actionable information on the competitive landscape of an organization, rather than “in house” information.

Business intelligence has become an increasingly interesting field of study for both researchers and practitioners, which can be seen in the increasing number of papers published every year on the topic of business intelligence, even if the terminology is not always uniform (Jourdan, Rainer and Marshall 2008; Chen, Chiang and Storey 2012).

Alpar and Shulz (2016) concluded that as BI gains interest and sees increasing possibilities due to advances in methods and data availability, business intelligence know-how and professionals have even become a bottleneck in operations, which has created a need for BI tools available and comprehensible for “casual” users and organization members, without any BI expertise. These possibilities pop up increasingly, but it appears that many have yet to take advantage of them (Alpar and Schulz 2016). So, why is business intelligence such a lucrative concept for organizations, what are the benefits in engaging in BI processes?

Watson and Haley (1998) summarized the key benefits of data warehousing - and older term used frequently interchangeably with business intelligence (Trieu 2017) - to be saved time, improved information quality and quantity, improved business processes, and support for accomplishing strategic objectives. Trieu (2017) summarizes in their review that BI processes can improve customer targeting and offering development in addition to refining organizational intelligence. All reasons and benefits mentioned are intangible and difficult to measure. Hence, calculating a ROI for investments in business intelligence systems is a nontrivial task (Watson and Haley 1998), but some attempts made have shown that BI investments are a high risk – high reward –endeavor. It is estimated that nearly half of all attempts fail (Kelly 2007), yet successes might lead to returns (ROI) of greater than 1000% (Watson, Wixom et al. 2006).

In an empirical study of how data analytics competences correlate to the decision making accuracy and effectiveness in organizations, it was found that the considered data analytics competences (bigness of data, data quality, analytical skills, domain knowledge, and tools sophistication) all contribute towards decision making accuracy, and all, except data bigness, contribute towards decision making efficiency (Ghasemaghaei, Ebrahimi and Hassanein 2018). Therefore, the quality of data analytics processes related to business intelligence competences are important for making fast paced, correct strategic decisions. In today’s fast moving business environment, where fast and accurate strategic decision-making can make you or break you, it is no wonder BI is gaining traction as a field of study.

The most obvious application of BI methods is e-commerce and market intelligence, which generates most of the “hype” surrounding BI (Chen, Chiang and Storey 2012). Ziegler (2012) goes as far as to suggest that the “primary purpose of CI methods is to find out how a given brand or company is perceived by the public, and how this relates to a brand’s or company’s competitors” – market intelligence in other words. These types of applications take on forms such as recommender systems and targeted marketing, social media monitoring and analysis, which enable long-tail marketing, increased sales and customer satisfaction (Chen, Chiang and Storey 2012).

Of course, business intelligence processes have applications outside of e-commerce and market intelligence. Chen, Chiang and Storey (2012) raise government and politics, science and technology, health and well-being, and security and public safety as other major areas of application. In science and technology, the applications of BI processes

present themselves as knowledge discovery, hypothesis testing, and innovation methods with goals in scientific advances. In healthcare there are applications such as decision support, and community analysis with the goals of improved care quality, and patient empowerment. Improvements in public safety may be achieved with BI application such as crime analysis and informatics. In politics, tools such as opinion mining, network analysis, and text analytics may be used to keep track on the political developments in both official data and public opinion – a kind of political “market analysis”. (Chen, Chiang and Storey 2012)

Jourdan, Rainer and Marshall (2008) summarize in their review that the research on business intelligence covers a wide range of fields of expertise, from computer science to artificial intelligence, to user experience, to marketing, to information management. The most commonly mentioned methods over fields of BI applications in the studied literature are text and web analytics and mining, sentiment analysis, network analysis, data mining. Other methods considered in the literature include clustering, statistical analysis, predictive modelling, citation counting, and trend detection, which might be argued to be a form of predictive modelling. (Chen, Chiang and Storey 2012; Ziegler 2012; Wani and Jabin 2018) While, most BI related tasks are automated with text analytics, natural language processing, and information retrieval methods, some are still done by hand (Ziegler 2012).

Data mining refers to a search of valuable, new information in large volumes of data via cooperation of humans and machines to achieve a descriptive or predictive goal (Kantardzic 2011). Web mining can be defined as data mining methods applied to “large web repositories” (Mobasher et al. 1996). Stavrianou, Andritsos and Nicoloyannis (2007) summarize in their review text analytics/mining as knowledge discovery in text archives, which could be defined as data mining applied to text data per the two former definitions. Sentiment analytics refer to methodology to analyze and extract subjective information and meaning from opinions and attitudes in language - typically used to determine how opinions are polarized between “positive” and “negative” sentiment (Dave, Lawrence and Pennock 2003; Indurkha and Damerau 2010). Network analytics refers to examining and interpreting relationships and discovering patterns among individuals and groups – typically via mathematical methods (Scott 2017).

These methodologies are yet again vague regarding terminology – overlapping and ambiguous. For example, in the definitions text and web mining are just application of data mining. Furthermore, in this thesis clustering is used as a form of text analytics. The most relevant methodologies for this thesis will be studied in more detail later on in this chapter. Firstly, the context for method relevancy needs to be set. The next chapter studies the big data problem and how its current situation plays into business intelligence.

## 2.2 The Big Data Problem

To begin to assess the situation of what is today called "the big data problem" some historical context for text and data mining may need to be set. The next part of this thesis explains the research methodology called content analysis, which may be seen as the most basic form for all of the more automated data analysis methods explained further on in this thesis.

### 2.2.1 Content Analysis

Content analysis is defined as a research method for analyzing communication messages of verbal, written, or visual nature (Cole 1988). The main goal of content analysis is simply to "provide knowledge and understanding of the phenomenon under study" (Downe-Wamboldt 1992). Since its birth, content analysis has been used to analyze communication from hymns to news articles, from advertisements to speeches (Harwood and Garry 2003). It has a wide range of fields of applications - from nursery and health opinion mining (Chew and Eysenbach 2010) to ecological disturbance study (Altaweel, Bone and Abrams 2019).

There exist different visions on types of content analysis. For instance (Elo and Kyngäs 2008) propose there exist two types of content analysis: inductive and deductive. Inductive works from "data up", meaning that categorization of the data is created based on the studied data (Elo and Kyngäs 2008). Deductive content analysis, on the other hand, uses an external categorization based on previous knowledge (Elo and Kyngäs 2008). This split is well in line with the different types of content analysis proposed by (Hsieh and Shannon 2005), who call these two previously described approaches conventional and direct content analysis respectively. Hsieh and Shannon (2008) also describe a summative content analysis, in which the focus on is on discovering keywords or contents and their usage within certain contexts.

In the conventional, or inductive, form of content analysis the data is studied and coded – categorized – in order to create understanding of a previously unknown phenomenon (Hsieh and Shannon 2005). This categorization can be hierarchical with different levels of abstraction (Elo and Kyngäs 2008). The content analysis process begins with the researcher delving into the data and simply reading it through (Tesch 2013). After this, the data is labelled and coded, grouping data together in order to create meaningful clusters of it (Coffey and Atkinson 1996; Patton 1990).

Hsieh and Shannon (2005) describe the setting for direct, or deductive as by Elo and Kyngäs (2008), content analysis as one where prior information exists about a phenomenon under study, but which might benefit from further study and description. The goal being to extend or validate the external framework and previous research (Hsieh and Shannon 2005).

The quality of the content analysis largely depends on the researcher's level of immersion in the data. Failing to understand the data through and through may lead to misunderstanding the key categories and concepts in the data. (Hsieh and Shannon 2005) The reliability and validity of conventional content analysis can be bettered by increasing researcher interaction with the data, either by longer engagement with it, or by the data being studied by multiple researchers (Lincoln 1985; Manning 1997). Moreover, in all content analysis the data studied should be representative of the studied phenomena. Often the data in content analysis gets very large, and a smaller sample is taken. In these cases the sample should retain its representativeness as well. (Duncan 1989) In quantitative research fields, content analysis is sometimes seen as too simplistic a method, while qualitative fields may deem it not qualitative enough (Morgan 1993).

## 2.2.2 The Situation of Data Today

Chen, Chiang and Storey (2012) studied the development of business intelligence, and noticed a transition of focus from organizations' internal structured databases for BI into unstructured web-based data, and further on into mobile and sensor based data for BI. This development trend increases the amount of data organizations are able to use for business intelligence processes, but at the same time, it becomes an unfathomable amount of data to convert into actionable intelligence. Instead of clean internal company databases, the data of the whole world becomes one's oyster.

This leads into the definition of the big data problem, given by Wani and Jabin (2018): the amount of data is increasing faster than the abilities to process it. Every second 1.9 million emails are sent, there are 50 million tweets per day, and 3.5 billion Google searches every day. This amounts to exa- zetta- and yottabytes of data. These types of numbers of data are simply not anymore feasible to content analysis, where a human is required to familiarize themselves with the data to derive information - automation is required. This automation can be achieved via machine learning methods, which will be covered in more detail later on.

This web based data is mostly unstructured, i.e. text, video, or sound. Data can also be semi-structured, such as XML files, or structured, meaning preprocessed databases and tables. This unstructured data, or at least the sets of the data important to BI, are scattered around the internet and often miss parts of important information. Sometimes data might even be biased and misleading on purpose. One ought to be careful when choosing which data to incorporate into analyses. (Wani and Jabin 2018)

To know what kind of data to collect there exists a need for knowledge discovery in databases (KDD) and data mining tools. Allahyari et al. (2017) use a definition for KDD and data mining as follows: knowledge discovery in databases is the whole process of locating actionable and useful knowledge in data, and data mining is a specific step in this process. Izenman (2008) determines KDD and data mining aim at descriptive results,

**Table 2.1.** *The knowledge discovery in databases (KDD) process (Izenman 2008)*

1.	Selecting target data
2.	Data cleaning – removing noise, locating outliers, and dealing with missing data
3.	Data preprocessing – transforming data into appropriate format
4.	Determining data mining tasks – regression, classification, clustering etc.
5.	Analyzing the data
6.	Interpreting and assessing the knowledge acquired from the analytics

and that the process is composed of the activities listed in table 2.1.

To try to explain how incomprehensible a process this is, try to imagine you download the whole of Twitter and Reddit into your brain and try to locate sensible knowledge there within to make important choices. You would most likely end up eating soap for a “meme” rather than gaining important knowledge on what business moves to make. Naturally, data analysis follows data mining, which is considered in this thesis as by Ghasemaghaei, Ebrahimi and Hassanein (2018), who combined two different definitions (Russom 2011; Ertemel 2015) into: “combination of processes and tools, including those based on predictive analysis, statistics, data mining, artificial intelligence, and natural language processing, often applied to large and possibly disperse datasets for gaining invaluable insights to improve firm decision making”.

Data analysis is difficult to get right: only 27% of companies that have invested into data analytics have noticed any benefits (Colas et al. 2014). The quality of the data used may be the most crucial obstacle to overcome to see benefits from data analytics (Hazen et al. 2014). Hence, we have arrived at the cornerstone principle of machine learning - “garbage in, garbage out”. So what makes data quality good? In this thesis defined data quality is defined as “quality of raw facts that reflect the characteristics of an entity or an event” (Detlor et al. 2013), and it is comprised of the data correctness and objectivity, data completeness, timeliness, and relevance, and how easy the data is to understand, interpret, present, and obtain (Wang and Strong 1996). However, even the top quality data is useless if there are no people to understand it, or means to process it into information. This is why employee domain knowledge, analytical skills, and the sophistication of data analysis tools also heavily influence how well data analysis can benefit an organization’s decision-making (Ghasemaghaei, Ebrahimi and Hassanein 2018). To be able to tap into the potential benefits of data analytics, organizations need to acquire quality data, human resources to understand what data is good, and how to turn it into actionable intelligence.

The most interesting form of data for this thesis is textual data, probably the most common form of unstructured data relevant for business intelligence processes (Allahyari et al. 2017). Text analytics, or text mining, refers to the discovery of knowledge that can be found in text archives (Stavrianou, Andritsos and Nicoloyannis 2007). This field has received much attention due to its wide application as a multi-purpose tool, borrowing techniques from Natural Language Processing (NLP), Data Mining (DM), Machine Learning

(ML), Information Retrieval (IR) and more. (Hu and H. Liu 2012)

From here on out it can be assumed that all considered data, in this thesis, is of unstructured textual form. Data analytics are considered from this point of view. As illustrated by Wani and Jabin (2018), the amount on data in the world is increasing at a nearly incomprehensible rate, and most of it is textual and unstructured. Natural humans could not possibly manually analyze the increasing amount of text data in the world, and it is reaching amounts that are difficult for computers as well. This “brute force” method for analyzing all textual data in databases is not sustainable. Therefore, there is a need for filtering and extracting only important, good quality, relevant information and patterns from text. This process is a part of the field called text mining, and its methods and processes have been gaining a lot of attention in recent times. Text mining is, of course, related to data mining, along with the other previously mentioned fields of study and methodologies. (Allahyari et al. 2017)

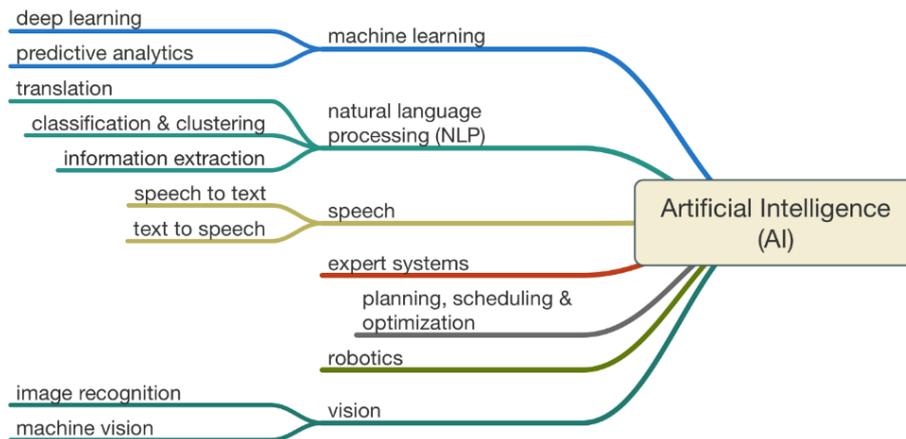
Luckily, knowledge discovery, and text and data mining have been able to see major advances due to progress and improvements in both software and hardware. As stated before, loads of disciplines are interested in improving data analytics, and this interdisciplinary effort has paid off. Data mining and analytics have improved with advances in fields such as machine learning, statistics, and artificial intelligence. (Allahyari et al. 2017) As mentioned before, text analytics borrow a lot of methods and techniques from various fields of study. Let us now begin to explore the nuances of the incredibly exiting world of computer science subfields and data analytics.

## 2.3 Navigating the Methodology of Business Intelligence

As there exists lots of overlap in the use of the terms used for business intelligence, competitive intelligence, there exists for data and text mining, knowledge discovery, information retrieval, data and text analysis, machine learning, artificial intelligence, natural language processing, computational linguists, and statistics to mention of few of the common disciplines mentioned in relation to BI. The following part of the thesis maps and explains the frameworks and methodologies used in this thesis, and their relationships with each other in the research to follow.

Loads of data mining and processing algorithms are heavily reliant on statistical principles and theories, which is logical seeing that statistics can be defined as a mathematical science dealing with the collecting, interpreting, explaining, and presenting data (Allahyari et al. 2017). Therefore, statistics are heavily interwoven into all the following considered fields of study. This thesis roughly follows the logic presented in the lecture series for the “Introduction to Pattern Recognition and Machine Learning” –course (Kämäräinen 2018), presented in figure 2.1, for subcategorizing the field of artificial intelligence.

In the context of this thesis, artificial intelligence (AI) is defined as by Izenman (2008) as a subfield of computer science focused on making machines able to think in a rational



**Figure 2.1.** Artificial intelligence and its subfields (Kämäräinen 2018)

manner and solve problems similarly to natural humans. The AI subfields most relevant to this thesis are natural language processing (NLP) – due to the textual nature of most BI data – and machine learning (ML). The following chapters explain these subfields, as they are relevant for the following research in the thesis – from the focus point of text data in BI contexts. Both NLP and ML are large fields with many methodologies, and only those required for understanding the research process of this thesis will be covered in detail.

## 2.4 Machine Learning

Machine learning (ML), a subfield evolved out of AI, concerns the ability of computers to learn from previous experience, and create and define approaches to locate predictable patterns in data by studying methods and algorithms for automatic information extraction. Machine intelligence is not possible without machine learning, which is why ML plays a very dominant role in the field of AI. While goal of KDD was descriptive, the goals of ML tend to be more of the predictive nature. (Izenman 2008; Allahyari et al. 2017) Machine learning is most often divided into two separate subfields: supervised learning and unsupervised learning (Friedman, Hastie and Tibshirani 2001).

In supervised learning, a machine learning algorithm is given input variables along with correct matching output labels appointed by an external entity. The learning algorithm attempts to formulate a function, typically referred to as a classifier, to map the input variables to output variables to make predictions on future input data. Variables may be continuous or categorical. A continuous output variable problem creates a regression problem, while a categorical output variable creates a categorization problem. Supervised methods are applicable in cases of known trends and topics. They include various methods such as nearest neighbour classifiers, decision trees, rule based and probabilistic classifiers, neural networks, and support vector machines. (Izenman 2008; Ziegler 2012; Allahyari et al. 2017)

In unsupervised learning, the external entity does not exist. While supervised learning explores the relationships between the in and out variables, unsupervised learning studies only the features and characteristics and “hidden” structure of the input variables in data. These methods are applicable to all kinds of data, and do not require a manual effort and a heavy labelling and teaching phase like supervised methods do. Since there exists no a priori knowledge, unsupervised learning methods provide an opportunity for serendipitous findings. Unsupervised learning methods include methods such as cluster analysis, joint probability density, proximity maps, and outlier location. (Izenman 2008; Ziegler 2012; Allahyari et al. 2017)

## 2.5 Classification

While supervised learning methods are not the focus point of this thesis, they are the most common form of machine learning (LeCun, Bengio and Hinton 2015) and are relevant to the data preprocessing and clean-up of this thesis. Therefore, some understanding of them is needed for understanding the methodology, and hence they are spared a brief overview now.

As stated before, some common supervised learning methods include nearest neighbour classifiers, decision trees, rule based and probabilistic classifiers, neural networks, and support vector machines. (Izenman 2008; Ziegler 2012; Allahyari et al. 2017) Neural networks will be considered in little more detail now, but for more information on other methods see, for example, Friedman, Hastie and Tibshirani (2001).

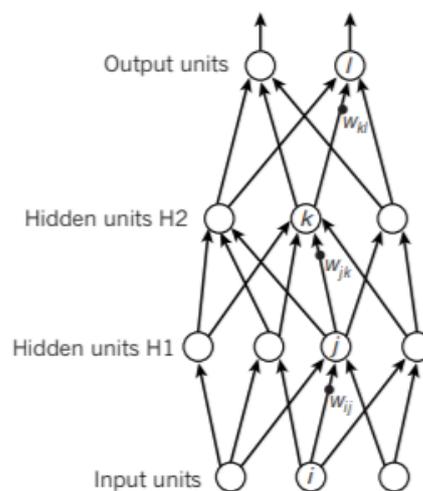
### Convolutional and Recurrent Neural Networks

Neural networks, and especially deep learning, have been beating other machine learning methods left and right in various applications in recent years - also in different natural language tasks - summarize LeCun, Bengio and Hinton (2015) in their review. Neural networks are not especially mysterious, while the literature and hype surrounding them may make it seem so, but a rather simple statistical nonlinear models that aim to model linear combinations of feature inputs to nonlinear functions representing the target output (Friedman, Hastie and Tibshirani 2001). The mysterious aspect of neural networks is what is called representation learning, in which a machine is fed raw data and automatically discovers data representations required for machine learning classification or categorization tasks - without humans interfering with the raw data (LeCun, Bengio and Hinton 2015). In other words, the machine itself decides what is important and what is not for the task it has been assigned.

How this works, is that a large set of data is collected and labelled correctly according to class. This data is given to a machine learning algorithm to train it for the classification task. The algorithm has an objective function, which measures the distance - sometimes

called error or loss - between the predicted output class labels and desired class labels. This is sometimes also called the loss function (Friedman, Hastie and Tibshirani 2001). The algorithm iteratively adjusts its own parameters - weights - to minimize this distance. This minimization is done with optimization methods, most often stochastic gradient descent. (LeCun, Bengio and Hinton 2015)

"Vanilla" neural networks as Friedman, Hastie and Tibshirani (2001) like to call them, have a single layer of representation, but when more layers - to transform the representation - are added, we are dealing with deep neural networks, and machine learning becomes deep learning (LeCun, Bengio and Hinton 2015). A deep neural network can be visualized as a web of the representation levels of nodes connected by the parameter weights, as shown in figure 2.2.



**Figure 2.2.** A deep neural network with four representation levels, and weights as  $w_{nm}$  (LeCun, Bengio and Hinton 2015)

When dealing with data that comes in multiple arrays, for instance a colour photo can be treated as three arrays per each colour RGB channel, convolutional neural networks are a go-to. They have seen great many successes and bested their competitors in various cases of applications in recent years. Convolutional neural networks, or CNNs, have two types of layers: convolutional, and pooling layers. The role of a convolutional layer is to locate "motifs" - local conjunctions of features, which is done by analyzing smaller, local patches of features from previous layers. Pooling layers are meant to merge semantically similar features, which is done by conserving the maximum values of local patches and reducing the dimensionality to create invariance to minor distortions and shifts in the raw data. (LeCun, Bengio and Hinton 2015)

When it comes to sequential input data, recurrent neural networks or RNNs are the man of the hour. As it happens, natural language is sequential data. RNNs process sequential input data single element at a time, while maintaining information about the previous elements of a sequence. For example, they are good at predicting the next word in a sentence, and can even output "meaning" vectors of images that can be formulated

as sentences used for captioning. These types of neural networks are under a lot of research and have had already many interesting successful use cases related to dealing with language. (LeCun, Bengio and Hinton 2015)

### 2.5.1 Evaluating Goodness

While deep learning is deemed so very promising and great in literature, it may now be the correct moment to consider what makes a supervised machine learning method good or bad. The simple answer is generalisability (Friedman, Hastie and Tibshirani 2001). Evaluating supervised learning models is done by testing the trained model on similarly labelled data the model has not seen before - the test set (LeCun, Bengio and Hinton 2015). This can be done for example by splitting the labelled data into a training and a testing set by taking a certain percentage of the data into training and leaving the rest for testing. If there is an abundance of data, the best option would be to have three sets: training (50%), validation (25%), and test data (25%). Validation data is used to estimate error, while test data is reserved only for testing model generalisability. (Friedman, Hastie and Tibshirani 2001)

Loss functions measure the error between the true label of the data and the label a trained model predicts for the input. Two typical loss measurements are squared error and absolute error, but depending on the use case others are used as well. They can be calculated for both training and testing data separately - the goodness of a model is of course assessed by its accuracy on test data. Typically the more a model is trained and the more complex it gets, the smaller the training data loss values. However, this results in overfitting, in which the variance of the test data label prediction increases and the model loses generalisability to data outside the training set - definitely something to avoid. There exists an optimal level of model complexity for test error. (Friedman, Hastie and Tibshirani 2001) Therefore, a model can be evaluated by the development of the loss function value on test data, resulting in information on how much the model should be trained and how models compare to each other.

Another straightforward way to assess the model performance is accuracy - or precision. This is simply the percentage of the labels the model predicted correctly. Sometimes it is useful to present what is called a confusion matrix, that takes into account true and false positives, as well as true and false negatives. An example is presented in table 2.2. (Friedman, Hastie and Tibshirani 2001) However, a critical touch must be maintained to this assessment. For instance, imagine an imbalanced dataset of negative (80%) and positive (20%) sentiment sentences. In such a case, even if the model only predicted a negative sentiment for all data, it would receive an accuracy score of 80%.

Other possible measurements used for assessing model performance are precision, recall, and F-measure. These can be seen in programming language libraries and applications such as Scikit-learn (SKLEARN n.d.(c)) and Prodigy (Montani and Honnibal 2018)

**Table 2.2.** Confusion matrix with 5.5% error (Friedman, Hastie and Tibshirani 2001)

True Class	Predicted Class	
	Class 1	Class 2
Class 1	58.3%	2.5%
Class 2	3.0%	36.3%

for Python. These are especially useful in the cases where data is indeed imbalanced. High precision implies low false positives, and high recall implies low false negatives. The higher the scores for both the better. Precision ( $P$ ) can be calculated as

$$P = \frac{T_p}{T_p + F_p}, \quad (2.1)$$

where  $T_p$  is the number of true positives and  $F_p$  is the number of false positives. Recall ( $R$ ) is calculated by

$$R = \frac{T_p}{T_p + F_n}, \quad (2.2)$$

where  $T_p$  is again true positives and  $F_n$  is the number of false negatives. Sometimes an F-measure is calculated which is the harmonic mean ( $2 \frac{P \cdot R}{P + R}$ ) of precision and recall. (SKLEARN n.d.(c)) These have been only a few methods to analyze classification model quality to give a brief overview into assessing supervised machine learning. The presented general ideas will be compared to assessing unsupervised learning later on in this thesis.

## 2.6 Clustering

In their article on deep learning, LeCun, Bengio and Hinton (2015) ponder on the future of machine learning, and conclude that while supervised learning methods have dominated the attention and successes thus far, unsupervised learning methods will likely become increasingly interesting and useful in the future. A carrying argument for this is the fact that most human and animal learning is unsupervised and undoubtedly efficient. Furthermore, with the increasing amount of data in the world, the manual annotation required with supervised learning is losing appeal (Zhou, Cheng and Zhang 2019).

Recall that unsupervised learning has no external entity to label data, or any other way to tell a machine the desired results and outputs. There is simply data to make sense out of. Unsupervised learning typically deals with higher dimensional data, and the interesting data properties tend to be more complex than in supervised learning. Taking all this into consideration, it can be said that despite the optimism of LeCun, Bengio and Hinton (2015), unsupervised learning deals with difficulties on a different level compared

to supervised learning. (Friedman, Hastie and Tibshirani 2001)

Different goals of unsupervised learning have different methods. For reducing data dimensionality or finding the most interesting latent variables in data, one might employ methods such as principal component analysis, self-organizing maps, multidimensional scaling, or principal curves. For binary high-dimensional data, association rules may be used to create descriptions of the data. For the research aims of this thesis as per the goal of categorizing business events, clustering is the most important method family. (Friedman, Hastie and Tibshirani 2001)

Clustering, or cluster analysis or data segmentation, aims to decipher whether, or how, a batch of data can be represented as smaller distinct and separate classes or categories of data - in other words subsets or clusters. This is done by creating clusters such that the distances of data points in a cluster are small, and the distances to data points in other clusters are large. In other words, clustering groups similar data-points together according to chosen measure of similarity. This measure of similarity is a fundamental element in all clustering methods. Hierarchical clustering, as the name implies, first creates larger clusters, which are split into smaller subclusters. (Friedman, Hastie and Tibshirani 2001) Some clustering methods require a specified number of clusters to be created, and some do not, example cases of both are presented shortly.

Clustering, and especially text clustering, is still an active area of research, and the common methods today will most likely evolve a lot in the future. This is possibly due to the increasing interest in unsupervised methods. (Zhou, Cheng and Zhang 2019) For instance, Zhou, Cheng and Zhang (2019) apply neural networks, typically a supervised method, to text clustering, hoping to create an end-to-end solution for text clustering, instead of having to split text clustering into separate process phases. These process phases will be explained the following parts of this thesis. Moreover, whether text clustering may benefit from a fusion method approach, for example with topic modelling, has been studied (Pourvali, Orlando and Omidvarborna 2019). Topic modelling will be covered later on this thesis.

### **2.6.1 K-Means Clustering**

Probably the most popular clustering algorithms used in text mining and other large scale clustering projects is the K-means clustering algorithm, due to its efficiency (Friedman, Hastie and Tibshirani 2001; Izenman 2008; Allahyari et al. 2017). It is a so called partition based algorithm, in which the centre of data points is considered the centre of the cluster of those data points (Xu and Tian 2015). Partition based clustering algorithms are typically computationally highly efficient, but relatively sensitive to outliers, and are not suited for all types of data, and need to be given a fixed number of clusters to create beforehand - which highly influences the quality of the clustering (Xu and Tian 2015).

The general functionality of the algorithm is of iterative nature: current cluster centres

**Table 2.3.** *The K-means clustering algorithm (Izenman 2008)*

- 
1. Input:
    - Set of datapoints  $L = \{x_i, i = 1, 2, \dots, n\}$
    - Number of clusters  $K$ .
  2. Do one of the following:
    - Form random initial clustering and for cluster  $k$  compute centroid  $\bar{x}_k, k = 1, 2, \dots, K$ .
    - Prespecify  $K$  cluster centroids  $\bar{x}_k, k = 1, 2, \dots, K$ .
  3. Compute squared euclidean distance of each item to its current cluster centroid:

$$ESS = \sum_{k=1}^K \sum_{c(i)=k} (x_i - \bar{x}_k)^T (x_i - \bar{x}_k),$$

where  $\bar{x}_k$  is the  $k^{\text{th}}$  cluster centroid and  $c(i)$  is the cluster containing  $x_i$ .

4. Reassign each item to its nearest cluster centroid so that ESS is reduces in magnitude. Update cluster centres after each reassignment.
  5. Repeat steps 3. and 4. until convergence.
- 

are updated based on the new points until convergence criteria is met (Xu and Tian 2015): The first set of possible cluster centres is a random guess, after which the closest cluster centre according to chosen distance measure is calculated for all points. Cluster centres are updated and replaced with the average "position" of all the data points that were closest to the centroid, until the convergence criteria is met. (Friedman, Hastie and Tibshirani 2001) A more detailed presentation of the algorithm is seen in table 2.3.

The required input is minimal in this presentation, but the algorithm may accept different types of parameters in practice, for instance as in the Python Scikit-learn machine learning library (SKLEARN n.d.(b)). However, there exists an even simpler algorithm inputwise for clustering that is very comparable to the K-means clustering called Affinity Propagation (Xu and Tian 2015).

## 2.6.2 Affinity Propagation

Affinity Propagation is also a partition based algorithm. However, it does not require a preset number of clusters to create. Despite its newness - proposed in 2007 - its meaning and contribution to clustering methodology is considered significant. Its advantages include the fact that the number of clusters is not preset, it is insensitive to outliers, and very simple and straightforward. However, it is very complex timewise, is sensitive to parameters, and is not very well suited for large datasets. (Xu and Tian 2015)

In its functionality, the algorithm regards every data point as a potential cluster center and the euclidean distance between data points as affinity. In practice this means, that

**Table 2.4.** The Affinity Propagation clustering algorithm (SKLEARN n.d.(b))

- 
1. Input:
    - Set of datapoints  $L = \{x_i, i = 1, 2, \dots, n\}$
  2. Set  $a$  and  $r$  values to zero:  $a$  is the *availability*, which means the evidence that datapoint  $x_i$  should choose point  $x_k$  to be its clusters centroid, and  $r$  is the *responsibility*, which means the evidence that point  $x_k$  should be cluster centroid of the cluster that contains point  $x_i$ . They can be calculated by

$$r(x_i, x_k) \leftarrow s(x_i, x_k) - \max[a(x_i, x'_k) + s(x_i, x'_k) \forall x'_k \neq x_k]$$

$$a(x_i, x_k) \leftarrow \min[0, r(x_k, x_k) + \sum_{i \text{ s.t. } i' \notin \{x_i, x_k\}} r(x'_i, x_k)],$$

in which  $s(x_i, x_k)$  is the similarity measure between points  $x_i$  and  $x_k$ .

3. Calculate

$$r_{t+1}(x_i, x_k) = \lambda r_t(x_i, x_k) + (1 - \lambda) r_{t+1}(x_i, x_k)$$

$$a_{t+1}(x_i, x_k) = \lambda a_t(x_i, x_k) + (1 - \lambda) a_{t+1}(x_i, x_k)$$

where  $t$  is the number of the iteration and  $\lambda$  is a damping factor.

4. Iterate step 3. over  $t$  until convergence.
- 

the higher the sum of all affinities of a data point, the higher the probability of the data point being a cluster centre. (Xu and Tian 2015) As can be seen from the mathematical representation of the algorithm in table 2.4, cluster centres are formed based on whether the centroids have enough similar points to them, and if they are "chosen representatives of themselves" by enough other data points (SKLEARN n.d.(b)).

### 2.6.3 Mean Shift Clustering

While K-Means and Affinity Propagation are so called partition algorithms, another popular type is a so called density based algorithm. The idea is simple to understand: In the vector space of the given data, the points closer to each other - a "dense" area in the space - is grouped together as a cluster. The advantage in these types of clustering algorithms compared to partition ones, is that they do not assume the shape of the clusters in the space, whereas K-means for example assumes that clusters are all of convex shape. However, these types of clustering methods are very sensitive to parameters and require a lot of memory. The most widely known density based clustering algorithm is DBSCAN, but in this thesis the Mean Shift algorithm will be studied more closely, as it is an iterative centroid based algorithm, and therefore more easily comparable to the infamous K-means algorithm. Simply put, the form of the Mean Shift algorithm in the Python Scikit-learn library is as presented in table 2.5. (Xu and Tian 2015; SKLEARN n.d.(b))

Now that a brief overview of what clustering algorithms are and how they work has been presented - along with three similar yet different types of algorithms studied in detail - it

**Table 2.5.** *The Mean Shift clustering algorithm (SKLEARN n.d.(b))*

- 
1. The algorithm determines the number of clusters to create based on the density and picks a possible center point for each cluster
  2. The possible next centroid candidate can be calculated from  $x_i^{t+1} = m(x_i^t)$ , where  $t$  is the number of the iteration and  $m(x_i) = \frac{\sum_{x_j \in N(x_i)} K(x_j - x_i)x_j}{\sum_{x_j \in N(x_i)} K(x_j - x_i)}$ . Here  $N(x_i)$  denotes the neighbourhood of allowed size of datapoint  $x_i$ , and  $K$  denotes a defined kernel function. Basically the new centroid candidate is the mean of samples in its neighbourhood.
  3. The algorithm converges when changes in iterations are small, after which centroid candidates are filtered to eliminate possible overlap.
- 

is time consider what makes a clustering result good or bad. In other words, what does it mean when Xu and Tian (2015) state that Affinity Propagation or Mean Shift are not suitable for large datasets - where do they fail?

## 2.6.4 Evaluating Goodness

As stated previously, different clustering methods are suitable for different situations - others are faster than others, and others are better suited for dealing with a certain type of data. These stem from the nature of the methods themselves, but these comparisons are mostly important before doing the clustering itself - they answer the question of what type of clustering algorithm to choose depending on one's data, and computational and time resources. The situation is different when assessing the result and outcome of the clustering applications.

While supervised learning evaluation had some possible difficulties, it is still a rather straightforward process with clear measures of success, such as loss or accuracy. When it comes to unsupervised learning, this is not the case - there exist no clear measures of success. The inferences drawn from unsupervised method applications are difficult to prove to be valid or of good quality, which often needs to be done with "heuristic arguments". Since the quality and effectiveness of unsupervised learning method outcomes is largely a matter of subjective opinion, a great variety of proposed methods for result assessment has sprung up. (Friedman, Hastie and Tibshirani 2001)

For clustering evaluation - considered the most important class of unsupervised learning methods (Xu and Tian 2015) - some metrics do exist. They are split into external and internal measures. External measures require an external "gold standard" classification the clustering result is compared to. However, these nearly never exist in real applications. Internal measures focus on measuring how "well defined" clusters are in terms of how similar data points are within a cluster compared to the similarity of data points between clusters. The more similar cluster points are in-cluster and the more dissimilar points in different clusters, the better the evaluation result. Some examples of external evaluation

measures are the Adjusted Rand index, Mutual Information based scores, homogeneity and completeness measures, Fowlkes-Mallows score, and the Contingency Matrix. Examples of internal measures are the Silhouette coefficient, the Calinski-Harabaz, and Davies-Bouldin indexes. (SKLEARN n.d.(b); Maulik and Bandyopadhyay 2002; Xu and Tian 2015)

A question may be raised on if - as stated by Friedman, Hastie and Tibshirani (2001) - unsupervised learning result quality and validity needs to be heuristically argued in most cases, what are the measures presented above useful for? For internal measures of clustering performance, a good score indicates the internal similarity of points in clusters and high dissimilarity of data points in different clusters. For clustering methods that require the number of clusters to be predetermined, these scores may be used to evaluate what is the best defined number of clusters to create if not known beforehand (Maulik and Bandyopadhyay 2002). Different clustering results may be created with different numbers of clusters, and the best defined one can be then evaluated heuristically.

## **2.7 Natural Language Processing and Computational Linguistics**

Natural language processing was categorized as a subfield of computer science, artificial intelligence, and linguistics by the literature studied by Allahyari et al. (2017). Computational linguistics are defined as the “study of computer systems for understanding and generating natural language” (Grishman 1986). However, as if there was not enough ambiguity in terminology before, the terms computational linguistics (CL) and natural language processing (NLP) are commonly used interchangeably, when they in fact do not refer to the same field of research despite being closely related. Their main difference is in the goals and objectives: CL is a more theoretical, formal field of study that attempts to discover how verbal and grammatical patterns correspond to meaning, while NLP is a more engineering type approach to computationally approach natural language. NLP focuses more on effective algorithms, referred to as parsers, to compute the structure and meaning of natural language. Some parsers are based on findings in the grammatically oriented CL, but the trend is developing towards probabilistic modelling, which focuses on attempting to find the most likely meaning and structure in text among many different possibilities. (Tsuji 2011) Both are subfields of computer science attempting to make sense of natural language regardless of the nuanced differences. However, considering these differences, it can be said that NLP is the more relevant field of study concerning this thesis.

Natural language has its own rather peculiar difficulties and obstacles. Computational linguistics deals with grammatical patterns corresponding to meaning, but these grammars are not a law of nature and do not always hold. As Sapir (2004) wisely concludes: "Unfortunately, or luckily, no language is tyrannically consistent". People often forego for-

mal grammar, especially in day-to-day conversational situations such as social media and emails. As stated before, this kind of data is nowadays important for BI processes. This is one compelling fact to turn towards NLP methodology over CL in this context. Moreover, grammar focused methods tend to be prone to failure in name recognition (Sapir 2004). Business intelligence systems unable to recognize the brand names of their customers or competitors may not be very useful.

If grammatical inconsistencies were not enough trouble, some situations have a different meaning per the exact same language used. The term "play" can be a verb, a noun, and even those may have different meanings depending on context: One can play a game or play a piece of music. By itself, analyzing the meaning of the term is difficult for people as well as computers. Language meaning may vary even depending on the human reader. For instance, one reading this thesis may have interpreted the sentence "let us now begin to explore the nuances of the incredibly exiting world of computer science subfields and data analytics" in chapter 2.2 as either sarcastic or not, depending on their subjective worldview, as well as their view on the writers view of the world. A computer does not have access to these interpretations by the reader, hence making the sentence difficult to analyze objectively - or subjectively - by computational means. This dilemma of various meanings of language, and figuring out the most probable meaning, is often referred to as WSD - word sense disambiguation. (Stavrianou, Andritsos and Nicoloyannis 2007)

The next part of this thesis covers typical NLP related preprocessing tasks and their functions and purposes. Preprocessing deals with turning raw text into a format understandable by computers - numbers. A dilemma to consider is the relationship of data quality and required computational effort - minimizing loss of information and maximizing efficiency. This is not an easy task, and has been studied very little when it comes to unsupervised learning methods. Scholars tend to like to go with whatever previous literature has done. This is not ideal, as the studied field may affect what preprocessing steps contribute to data and information quality, and what subtract from it. For example, including numerical data in text may be beneficial in studying juridical texts, but pointless in some applications. Moreover, there is no reason to assume preprocessing that benefits supervised learning methods is applicable to unsupervised methods. There exists little research into how text preprocessing affects unsupervised learning results. (Denny and Spirling 2018)

### **2.7.1 The Text Preprocessing Process**

This chapter presents a text preprocessing logic that roughly follows the combination of methods and steps considered by Denny and Spirling (2018), Allahyari et al. (2017), and Stavrianou, Andritsos and Nicoloyannis (2007). Few terminological points: corpus as a term refers to a large collection of text data. Corpora is the plural of a corpus.

## Tokenization

Tokenization refers to the process of splitting a string of text into smaller pieces, referred to as tokens. A sentence is a token of a paragraph, and words are tokens of sentences. Generating a list of tokens is essential for further text processing via computational methods. (Allahyari et al. 2017; Hardeniya 2016)

There are of course different methods to go about tokenization. Tokenization can be done with readily available tokenizer functions in programming language libraries, and there are many to choose from: Some tokenizing methods transform all tokens into lowercase, and some do not (Denny and Spirling 2018). Some tokenizers split a sentence into words in a way that creates tokens of punctuation, or a way that keeps punctuation as a part of the token. For example, "can't" could become "can't" or "can" and "'t" separately depending on the tokenizer. Should this not be enough options to choose from, one could create their own rule-based tokenizer. For example, one could specify that if a token has "n't" in it, it will be split into two tokens: one merging with the former token and other becoming "not", such as "can't" becoming "can" and "not"... Obviously this is not easy, as "isn't" would with the same logic become "isn" and "not". (Hardeniya 2016)

## Filtering and removing stop-words

This step deals with the computational effort versus data quality dilemma. A typical step in text preprocessing is filtering out stop-words. Stop-words can be defined as words that do not contribute much information, such as "the", "a", and "he". They can be defined by the preprocessor themselves, for example counting the 5% most common words in a corpus and ignoring them, or retrieved readily composed from an external source, such as a programming library. Similarly, words too rare to bring any informational gain may be removed. (Denny and Spirling 2018; Hardeniya 2016)

As stated before, in some cases the removal of numerical information from textual data may be meaningful and not so much in others. For some research cases, such as Twitter, the inclusion of hashtag characters may be informational, but irrelevant in others. The researcher must carefully determine what special characters to cleanse from data in order to conserve information. Moreover, some words may become stop-words in a certain context. Denny and Spirling (2018) use the example of "congress" being a stop-word in a political context. (Denny and Spirling 2018) One ought to be careful with stop-words, in some contexts "she was arrested" and "she arrested" mean two totally different events, and the distinction between them may very well be crucial (Stavrianou, Andritsos and Nicoloyannis 2007). These filtering effects should be taken into consideration when determining the preprocessing process for text analysis.

## Stemming and Lemmatization

Stemming refers to the process of simply chopping words to their "stems" in a rule based manner, usually referring to simply removing endings like "-s", "-es", "-ed", or "-ing". This is considered a "crude" method, with an accuracy of around 70%. Lemmatization refers to a more sophisticated methodology of reducing words to their basic format. It takes into account context and the part of speech in question - more on this soon - and normalizes the word to its "root". For example, in stemming, "eaten" and "eating" become "eat", but stemming is incapable of turning "ate" into "eat", which lemmatization does. (Hardeniya 2016) Similarly to stop-word removal, these processes may affect the text analysis results, and should be considered carefully before implementation.

## Part-of-speech Tagging

Part-of-speech (POS) refers to whether a word, as a part of a sentence, is a verb, a noun, an adjective, and so forth. The words - tokens - are assigned a tag to go along with their predicted POS value in the preprocessing step referred to as POS tagging. State of the art POS tagging algorithms, usually included in programming language NLP libraries, can predict the part-of-speech of a word with an accuracy of around 97%. (Hardeniya 2016) Figure 2.3 presents examples of used POS tags.

POS	DESCRIPTION	EXAMPLES	POS	DESCRIPTION	EXAMPLES
ADJ	adjective	<i>big, old, green, incomprehensible, first</i>	PART	particle	<i>'s, not</i>
ADP	adposition	<i>in, to, during</i>	PRON	pronoun	<i>I, you, he, she, myself, themselves, somebody</i>
ADV	adverb	<i>very, tomorrow, down, where, there</i>	PROPN	proper noun	<i>Mary, John, London, NATO, HBO</i>
AUX	auxiliary	<i>is, has (done), will (do), should (do)</i>	PUNCT	punctuation	<i>., (, ), ?</i>
CONJ	conjunction	<i>and, or, but</i>	SCONJ	subordinating conjunction	<i>if, while, that</i>
CCONJ	coordinating conjunction	<i>and, or, but</i>	SYM	symbol	<i>%, \$, ©, +, -, ×, ÷, =, :, 😊</i>
DET	determiner	<i>a, an, the</i>	VERB	verb	<i>run, runs, running, eat, ate, eating</i>
INTJ	interjection	<i>psst, ouch, bravo, hello</i>	X	other	<i>sfpksdpsxmsa</i>
NOUN	noun	<i>girl, cat, tree, air, beauty</i>	SPACE	space	
NUM	numeral	<i>1, 2017, one, seventy-seven, IV, MMXIV</i>			

**Figure 2.3.** Part-of-speech tag examples (Honnibal et al. 2019)

POS-tagging comes in handy when someone has an interest in a certain parts of speech in a text, or combinations of them. For example, someone may be interested in finding all the nouns in a text, or all adjectives followed by a noun, combinations, such as "green forest" and "white house". These kinds of POS patterns are sometimes referred to as chunks. (Hardeniya 2016)

## Chunking and N-grams

Extracting chunks, sometimes referred to as "partial parsing", is done by specifying interesting modified regular expressions (Hardeniya 2016). Regular expressions are a kind of mini programming language intended to parse through text, matching occurrences to a specified pattern. Few examples: in regular expressions "." marks any character, "[0-9]" marks any numeral, and "VB.", for example, would match "VBC", "VB?", "VBx", and so forth. Therefore, "<ADJ>\*<NOUN>\*" in regular expressions would be used to match a POS-tagged piece of text corresponding to any number of adjectives followed by any amount of nouns, such as "long yellow tape roll", since in regular expressions "\*" marks that any number of repetitions of the pattern is taken into account (Friedl 2002). If "roll" can correctly be interpreted as a noun instead of a verb in this case.

While a single word is the most commonly used token unit in NLP, sometimes it is worthwhile to consider word sequences - N-grams, since some words hold little meaning by themselves. As an example, "house" as a unit of a sentence containing "white house" loses meaning. A sequence of "the", "white", and "house", could be considered as a trigram "the white house" or bigrams "the white" and "white house". "The white house" may be extracted as a noun chunk with chunking methods, the difference in using N-grams is that they take into account all "contiguous sequences of tokens". (Denny and Spirling 2018)

## Named entity recognition

Aside from chunks and N-grams, so called named entities may be especially interesting. The used example "the White House" may in some instances be recognized a named entity. Names entities may be persons, locations, or organizations, to name a few possibilities. Models to recognize named entities are readily available in programming language libraries, but sometimes there exists incentive to create one from scratch for a specific purpose. The names entities are assigned tags, such as "PERSON", "ORG", or "TIME", in a similar manner to part-of-speech tagging. (Hardeniya 2016) Notice that previous choices regarding lowercasing for instance may become important later on in recognizing interesting data.

### 2.7.2 Vectorization

To be able to make use out of any of the preprocessing steps described before, text needs to be transformed into a form understandable for computers. This does not mean simply a ".txt" file format, since words and sentences do not mean anything to a binary machine. Therefore, they need to be transformed into a numerical representation. (SKLEARN n.d.(a)) This presentation in NLP tasks most commonly takes the form of vectors, referred

**Table 2.6.** *Short sentences for use in text vectorization examples*

1.	This is a sentence.
2.	This is another sentence.
3.	The second sentence follows the first sentence.

to as a vector space model (VSM) (Coelho, Peng and Murphy 2010). The form a VSM takes is a matrix format, in which documents are represented as the words of the corpus as dimensions. In these representations tokens can be assigned a value of importance - a weight - regarding a document (Allahyari et al. 2017; Benedetti et al. 2019). This can be done with vectorizers, of which there are different kinds (Pedregosa et al. 2011). Two common, but very basic and buildable (Zhang, Jin and Zhou 2010; Allahyari et al. 2017) logics to perform vectorization are presented next.

Both are based on the assumptions that terms are the more important to a document the more they appear in it, and that rare terms in a corpus are more meaningful than common ones (Salton and Buckley 1988; Zobel and Moffat 1998; C. Manning, Raghavan and Schütze 2010). Improving vectorization is an active area of research today, and alternatives to the two common methods presented next are being created (Nikolenko, Koltcov and Koltsova 2017; Aryal et al. 2019; Xu, Harzallah and Guillet 2019).

### **Bag-of-words**

A simple way to perform text document vectorization is called the bag-of-words, or BOW, representation. It has been around since the 1970's and has stayed popular (Salton, Wong and Yang 1975). The result is a numerical representation of tokens that does not take into account the relations or positions of the tokens - simply the count of tokens themselves. These tokens may be singular words, N-grams, or chunks of words. Taking N-grams or chunks into consideration brings some word relation information into the method, as words appearing in close proximity to each other are grouped together. (SKLEARN n.d.(a))

The bag-of-words representation is generated by first tokenizing the documents and assigning integer identification numbers for each token. The tokenization can be of course done by any means deemed most useful: creating tokens per word, N-gram, or chunk. After this, the number of times each token appears per document is counted. After, these counts are normalized. For a document, each token occurrence frequency is considered as a feature, and these features make up a multivariate sample vector for each document. As a result, a document corpus can be represented as a matrix, for example with the document feature vectors as rows and tokens as columns. (SKLEARN n.d.(a)) Tables 2.6 and 2.7 present a simple example of how a BOW representation is formed for a corpus.

Table 2.7 represents the output of a bag-of-words model where the sentences are con-

**Table 2.7.** *The bag-of-words representation of the example sentences.*

	this	is	a	sentence	another	the	second	follows	first
1.	1	1	1	1	0	0	0	0	0
2.	1	1	0	1	1	0	0	0	0
3.	0	0	0	2	0	2	1	1	1

sidered as documents. In this example case, the documents have been submitted to a preprocessing process of tokenizing and lowercasing without punctuation. Stop-words have not been removed and nothing has been stemmed or lemmatized.

A conclusion can be drawn, that a large corpus will have many zero values in the matrix as there are many tokens that will not appear in all texts. Also, that there will be very high counts for common words if stopwords are not filtered out in preprocessing.

## TF-IDF

Term Frequency Inverse Document Frequency, or more commonly TD-IDF, is a statistical weighing methodology for locating most relevant tokens in a text document in relation to a corpus. As the name of the term implies, the method determines the relevant frequency of tokens in individual documents in comparison to the inverse frequency over all documents in a corpus. The result is an estimate on how relevant a token is to a certain document. (Ramos et al. 2003) It tackles the potential issues of the straightforward count measures in BOW representations. For example, while in BOW a document may have a rare word appear frequently, it is not considered any more important to the document than a common term appearing as many times. A very sophisticated level of stop-word removal would be required to prevent common terms from overpowering the rare yet interesting terms. This is not the case in TF-IDF, which is the most used term weighting method (Salton and Buckley 1988 ;C. Manning, Raghavan and Schütze 2010).

The TF-IDF weight for a token is calculated by

$$tfidf(t, d) = tf(t, d) \cdot idf(t), \quad (2.3)$$

in which  $tf(d, f)$  represents the token frequency in a document. In other words, the number of times a token appears in a document, which is multiplied by a value for each token ( $t$ ) referred to as *inverse document frequency*

$$idf(t) = \log\left(\frac{n_d}{df(d, t)}\right), \quad (2.4)$$

where  $n_d$  is the total number of documents in a corpus, and  $df(d, t)$  is the number of documents containing token  $t$ . (Berger et al. 2000; Allahyari et al. 2017; SKLEARN

n.d.(d)) Table 2.8 presents an example TF-IDF transformation done on the BOW model in tables 2.6 and 2.7.

	this	is	a	sentence	another	the	second	follows	first
1.	0,46	0,46	0,69	0,33	0	0	0	0	0
2.	0,46	0,46	0	0,33	0,69	0	0	0	0
3.	0	0	0	0,34	0	0,71	0,36	0,36	0,36

**Table 2.8.** TF-IDF representation of the example sentences.

While the above theoretical representation is seen in literature, some different variations of calculating the TF-IDF can appear in programming applications. For example, the Python machine learning library Scikit-learn, used for the example in table 2.8, uses two different methods to calculate  $idf$ -values. Namely,  $idf(t) = \log\left(\frac{1+n_d}{1+df(d,t)}\right) + 1$  or  $idf(t) = \log\left(\frac{n_d}{df(d,t)}\right) + 1$  depending on given parameters. In table 2.8 the latter is used. Moreover, the resulting weight vectors for tokens are normalized by the Euclidean norm in the library's method. (SKLEARN n.d.(d))

It can be seen, that the TF-IDF model prioritizes the most characteristic tokens to a document. For example, while in the bag-of-words model (2.7) both tokens "sentence" and "the" receive a score of 2 for document number 3, the token "sentence" is more common across all documents and receives a lower weight in the TF-IDF model than token "the", which is unique to the document in question. While TF-IDF has this advantage and sophistication, it still does only count words in a document without regard for further semantics, and can therefore still be considered a form of a BOW representation. Moreover, it has been noticed that the results of TF-IDF may be misleading when assessing document similarity (Aryal et al. 2019).

## 2.8 Topic Modelling

In addition to clustering, a commonly used unsupervised learning method for text mining is called topic modelling. (Nikolenko, Koltcov and Koltsova 2017; Pourvali, Orlando and Omidvarborna 2019) It is a compatible model family to NLP, and when combined allows for better interpretation of topics and text meanings. Topic modelling is a relatively new text mining method that has received increased attention and acknowledgement in various fields of study. Topic models are unsupervised models to split a corpus of texts by content into "substantively meaningful categories" referred to as topics. These topics are represented by a bag-of-words distribution of words marking a specific topic. (Mohr and Bogdanov 2013) Topic modelling aims at discovering abstract latent topics that exist within a text dataset, rather than simply split the data into separate groups (Blei, Carin and Dunson 2010).

The bag-of-words representation receives often critique for ignoring relations or words in text. This may be thought to imply that while topic modelling is ideal for locating topics

and themes in bodies of texts, the method may be ill suited for studying other interesting textual information such as narratives. However, topic modelling has proven surprisingly useful in cases where it may have been thought to not be the ideal analysis method, such as poetry analysis. (Mohr and Bogdanov 2013)

In their functionality, topic models typically accept documents and the number of topics to create as input, and returns the probabilities of words - tokens - coming up in discussion related to a certain topic - the bag-of-words representation. Topic models also return the distribution of the found topics across the studied corpus. The main difference to clustering is in the fact that topic modelling is a probabilistic method, meaning that all words found in a corpus have a probability of belonging to a topic, and documents have a certain probability of belonging to all topics, whereas clustering results are "black and white" - a document belongs to a certain cluster or does not. (Mohr and Bogdanov 2013)

Mohr and Bogdanov (2013) describe the main benefits of topic modelling as shifting the required interpretative work done on text corpora from pre-analysis to post-analysis - the interpretative work required will be explained in more detail in section 2.8.4. This is increasingly important as sets of text data grow larger and become difficult to interpret as they are. Analysis via topic modelling is faster, more efficient and objective than traditional content analysis methods. Topic models create a "macroscopic lens" to view the corpora through, which allows viewing the corpora in different lights - providing new insights and clarity for further analysis. Topic models are scalable and create different types of "lenses" for different projects. (Mohr and Bogdanov 2013)

As demonstrated by a review on trends in topic modelling (Mulunda, Wagacha and Muchemi 2019), the application areas of topic modelling are plenty. When searching online for text analytics in a marketing setting, often sentiment analysis comes up. Some topic modelling is specially made for, for instance, opinion mining on Twitter (Lim, Chen and Buntine 2016). On top of opinion mining, topic modelling on Twitter can be used for conversation trend analysis online (Lansley and Longley 2016; Särkiö et al. 2019). Other promising applications are health and education (Mulunda, Wagacha and Muchemi 2019), for instance, topic modelling may help in the identification of depression and neuroticism in students based on their texts (Resnik, Garron and Resnik 2013), which would aid in early intervention. Topic modelling is sometimes applicable also outside text data analytics. Examples of these types of cases are satellite image annotation (Lienou, Maitre and Datcu 2009) and gene function modelling (Liu et al. 2010).

### **2.8.1 Latent Semantic Indexing**

Latent Semantic Indexing (LSI) - sometimes used interchangeably with Latent Semantic Analysis (LSA) - is an older method for topic modelling, first proposed by Deerwester et al. (1990) for overcoming the issues of polysemic and synonymous terms - word sense disambiguity - regarding keyword document retrieval. Deerwester et al. (1990) created

this method believing that text clustering methods were too limited for the richness of text semantics, due to their incapability to cross classify. In its general functionality, the method creates a high-dimensional matrix of term-document relation data and creates a "semantic space" - usually of Euclidean metrics - via matrix singular value decomposition (SVD), where terms and documents are assigned a point in the space - allowing semantically similar terms and documents to be closely related in the space even when not necessarily sharing the exact same terms. (Deerwester et al. 1990)

The process begins with a matrix of terms per document, the bag-of-words representation of the corpus, analyzed by SVD. In the resulting matrix decomposition, some components may be very small and insignificant, allowing for dimensionality reduction, which in turn enables the generalization of the document-term matrix for the LSI model to be of any use. In the reduced model term to term, term to document, and document to document similarities are approximated, and all terms and documents are represented by factor vectors in the reduced dimension. (Deerwester et al. 1990)

To provide a more mathematical explanation, the bag-of-words matrix  $X$ , with terms as rows and documents as columns in this case, is calculated its SVD:

$$X = T_N S_N D^T_N \text{ and } \hat{X} = T_k S_k D^T_k,$$

in which  $N$  is the original full rank of the matrices and  $k$  is the reduced rank.  $S$  represents a diagonal matrix with singular values,  $T$  and  $D$  matrices represent left and right singular vectors. Term similarity can be computed as the dot product of two row vectors in  $\hat{X}$  representing the term occurrence patterns. Similarly document similarity can be estimated via the dot product of column vectors of  $\hat{X}$ . For comparing document and term relation, the single matrix cell value is of importance. LSI, in its time, was a promising new approach to document retrieval, and provided a good ground for further advances in topic modelling. (Deerwester et al. 1990)

## 2.8.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is the most used, "state-of-the" art topic modelling algorithm (Mohr and Bogdanov 2013). It was created by Blei, Ng and Jordan in 2003 by extending the LSI model and its derivative probabilistic LSI (Blei, Ng and Jordan 2003; Alahyari et al. 2017). The method is more easily applied to more complex and rich models, such as N-grams (Blei, Ng and Jordan 2003). The idea, simply put, is that documents can be represented by random mixtures of latent topics, which are in turn represented by a probability distribution over terms. The set of  $M$  documents  $D = \{d_1, d_2, \dots, d_M\}$  is the corpus and the set of  $N$  terms  $V = \{w_1, w_2, \dots, w_N\}$  is the vocabulary of the corpus. Topics  $z_j$   $j \in \{1, 2, \dots, K\}$  can be represented as multinomial probability distributions  $p$  over  $V$ , such that  $\sum_i^N p(w_n|z_j) = 1$ . The term distribution over a document  $d$  can be expressed as  $p(w_n|d_m) = \sum_{j=1}^K p(w_n|z_j)p(z_j|d_m)$ , in which  $K$  is the number of predeter-

mined topics. (Blei, Ng and Jordan 2003; Blei, Carin and Dunson 2010; Allahyari et al. 2017)

The generative process for the corpus in LDA can be presented as follows (Blei, Ng and Jordan 2003; Blei, Carin and Dunson 2010; Allahyari et al. 2017):

1. For each topic  $z_j, j \in \{1, 2, \dots, K\}$  sample a word distribution  $\Phi_j \sim \text{Dir}(\beta)$
2. For each document  $d_m, i \in \{1, 2, \dots, M\}$  sample a topic distribution  $\theta_d \sim \text{Dir}(\alpha)$ , and for each word in the document sample
  - a topic  $z_j \sim \text{Multinomial}(\theta_d)$
  - a word  $w_n \sim \text{Multinomial}(\phi_d)$

A joint distribution for the latent variables - topics, topic proportions, and topic assignments - and observed variables - terms, is generated by this process. The posterior distribution of the latent variables can be expressed as  $p(\phi_{1:K}, \theta_{1:M}, z_{1:M,1:N} | w_{1:M,1:N})$ , in which  $\phi_{1:K}$  represents the topics - patterns of terms over the whole corpus,  $\theta_{1:M}$  represents topic proportions and  $\theta_d$  represents how document  $d$  expresses topic patterns. Topic assignments are represented by  $z_{1:M,1:N}$  in which  $z_{m,n}$  expresses which topic the  $n$ th term of document  $d$  is associated with. (Blei, Carin and Dunson 2010; Allahyari et al. 2017)

The process of "reversing" the generative process in order to find the posterior distribution that generated the corpus in question is called *posterior inference* (Blei, Carin and Dunson 2010). Computing the posterior exactly requires too much effort to be a sensible thing to do (Blei, Ng and Jordan 2003), which is why several estimation methods are used such as variational inference, Gibbs sampling, or a Markov Chain (Allahyari et al. 2017). Calculating the posterior is the main computational task in LDA, and is under a lot of research in order to improve its efficiency (Blei, Carin and Dunson 2010).

### 2.8.3 Hierarchical Dirichlet Process

The previous topic modelling methods are so called finite-dimensional parametric topic models, which require a preset number of topics to create - a serious limitation. A newer development in topic modelling is the emergence of models that can adapt to a growing number of possible topics to create, namely Bayesian non-parametric models based on the Hierarchical Dirichlet Process (HDP). (Blei, Carin and Dunson 2010)

HDP is rather similar to LDA, but allows the number of topics to be determined by the model and has the possibility of new documents to trigger the emergence of new topics. These methods place a priori of infinite on the number of topics to model and the number of topics then becomes a part of the posterior distribution.

HDP draws a corpus from a process that looks similar to the LDA process:

1. Sample the topic base distribution  $G_0$  from  $G_0 \sim \text{DP}(\gamma, H)$ , in which  $H$  is a symmetric Dirichlet distribution over the vocabulary and  $\gamma$  is a positive scalar.
2. For each document  $d_m, i \in \{1, 2, \dots, M\}$  sample a topic distribution over topics  $G_d \sim \text{DP}(\alpha, G_0)$ , and for each word  $w_n$  in the document sample
  - the topic  $\beta_{d,n} \sim G_d$
  - the word  $w_{d,n} \sim \text{Multinomial}(\beta_{d,n})$ ,

in which  $\beta_n$  is a topic and DP is the Dirichlet Process. In the Dirichlet process of the form  $G \sim \text{DP}(\alpha, G_0)$   $G$  is a random distribution,  $G_0$  denotes the known base distribution over the same space as  $G$ . Samples drawn from DP are discrete with positive probabilities placed at these discrete "atoms". The positive scalar  $\alpha$  determines the distribution of probabilities over the atoms - the larger the value the more evenly distributed probabilities. The posterior inference is in this case also arduous to compute exactly, and the same estimators as for LDA are typically used. (Blei, Carin and Dunson 2010)

## 2.8.4 Evaluating Goodness

As with all unsupervised learning methods, the assessment and evaluation of topic modelling results is not a straightforward task and result validation requires well informed holistic argumentation. This of course applies to topic models as well. The researchers will need to interpret the results according to their subjective knowledge of the corpus and relative phenomena, which requires a profound familiarization with the field, and texts and discourse in question. The researcher will need to determine whether the topic modelling results are proper and not misleading.

## 2.8.5 Perplexity

One persisting question on the quality of topic modelling, is how many created topics is the optimal to give for the parametric topic models. This may be done iteratively to determine when the number of topics generated seems the most valuable regarding interpretation. (Mohr and Bogdanov 2013) This type of iteration is still an arduous trial and error task, and ways to avoid it are being researched. (Zhao et al. 2015)

Perplexity is a numerical measurement used to assess the quality of the whole topic modelling result (Nikolenko, Koltcov and Koltsova 2017). However, it appears that the determination of topic model topic number determination could benefit from perplexity evaluation (Zhao et al. 2015). The approach for topic number determination proposed by Zhao et al. (2015) calculates a perplexity based value for different topic number results, and the number that yielded the best perplexity value is suggested as an appropriate number of topics. However, perplexity is not the only measurement of topic modelling quality.

## Coherence

There is a need to efficiently evaluate topic models without the need for expensive gold standard human evaluation or external reference material that almost never exists. Topic coherence is a numerical metric that can estimate rather well the topic model evaluations made by experts compared to perplexity (Nikolenko, Koltcov and Koltsova 2017). The output of topic models for expert evaluation is typically a list of the top five to twenty most representative words of a topic in descending order, based on which experts can evaluate whether the terms represent an understandable, coherent topic. (Mimno et al. 2011; Stevens et al. 2012; Röder, Both and Hinneburg 2015)

Coherence, compared to perplexity, focuses on the coherence of the individual topics compared to the larger picture evaluated by perplexity (Nikolenko, Koltcov and Koltsova 2017). Coherence measures take the top  $N$  representative topic terms and calculate a confirmation measure of a sum over all word pair combinations. However, there exist different ways to calculate this confirmation measure, and therefore different coherence measures. Coherences based on PMI, pointwise mutual information, are the most similar to how a human would evaluate the model result. This coherence measure can be calculated by

$$\text{Coh}_{UCI} = \frac{2}{N \cdot (N - 1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \text{PMI}(w_i, w_j), \quad (2.5)$$

in which

$$\text{PMI}(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)} \quad (2.6)$$

and the probabilities  $P$  are estimated based on word co-occurrence counts derived from generating documents from a corpus by sliding a "window" over it. While  $\text{Coh}_{UCI}$  is the most alike to human evaluation, it is quite slow to compute compared to the fastest coherence measure  $\text{Coh}_{UMass}$  proposed by Mimno et al. (2011), which can be calculated by

$$\text{Coh}_{UMass} = \frac{2}{N \cdot (N - 1)} \sum_{i=2}^N \sum_{j=1}^{i-1} \log \frac{P(w_i, w_j) + \epsilon}{P(w_j)}, \quad (2.7)$$

in which the word probabilities are estimated from document frequencies of the original corpus of documents. The  $\epsilon$  is added to avoid a logarithm value of zero, and both measures perform better when its value is small, compared to the original publications, in which its value is 1. There exist other measures as well, but for this thesis the fastest and the most "accurate" ones are of most interest. (Mimno et al. 2011; Stevens et al. 2012; Röder, Both and Hinneburg 2015)

## 2.9 The Research Problem

To recap, the relevant research questions are as follows:

- How can unsupervised machine learning methods be exploited in the creation of an action and event categorization framework for business intelligence purposes?
- How do these exploitation possibilities differ in different, common unsupervised method approaches?

In the light of the studied literature and presented theory, the details and requirements for answering these questions can be considered further. Clustering and topic modelling can provide similar or different kinds of value to the categorization compared to each other, and both or clearly the other one may be more valuable. The results should therefore be compared and evaluated by a human familiar with the dataset.

How about different algorithms of topic modelling and clustering? Do they provide different information and value and how useful is it? For instance, does the approach of Affinity Propagation reveal useful information about the amount of topics in a dataset or not, how about Mean Shift? Might using LSI with N-grams be more fruitful than using LDA with unigrams, how about extracted noun chunks? Possibly, on top of evaluating what type of clustering may be the most useful, there may be something interesting to notice about the behavioral differences of these methods.

In the tests done in this thesis, the goal is to try different types of approaches and approach combinations to clustering and topic modelling and comparing them in order to determine how different choices in preprocessing and methods affect the result. This is to provide information and recommendations on what one may wish to do when determining the event and action categories of documents to use in information extraction via clustering or topic modelling. Possibly only a single approach is clearly better than the others, or maybe all have something different to contribute. If they do, the goal is to find out what it is that they have to offer.

The next chapter sheds light on the data used for clustering and topic modelling - how it was acquired, what does it look like, how it was processed. After this, the process of the preprocessing, clustering, and topic modelling is explained in further detail. Such as, why the algorithms used are from open source libraries, and what parameters were used and why.

## 3 DATA AND METHODOLOGY

The given open ended research questions for this thesis were inspired from reading relevant literature on business intelligence related event extraction, and often times not finding any reasoning for presented event categorization. Furthermore, the interest in these questions was reinforced when at a research methodology course seminar at Tampere University in spring 2019, for an event study assignment presentation (Litovuo 2019), the presenting groups stated that categorizing events was the most difficult part of the assignment due to lack of standard.

This kind of a situation is a typical premise for exploratory research, in which open questions are asked to gain insights and understanding of a topic of interest when there exists uncertainty regarding the nature of the problem. Exploratory studies tend to be unstructured and adaptable to change. (Saunders, Lewis and Thornhill 2009) This is also the case in this thesis, and many different approaches were tried and changed since they did not contribute to answering the research questions. The following chapter presents the final workflow of the research steps executed in order to answer the posed questions. In this chapter, the code and its functionality is described on a general level.

### 3.1 Data Gathering

The data was collected from LexisNexis by keywords. The used keywords being companies that manufacture digital cameras. These keywords are presented in table 3.1. The data consists of text news articles from various publications. The keywords were restricted to a specific field of technology companies to avoid too wide a scope of news, which would result in a large amount of data not related to business intelligence, and in order to attempt to mitigate the effects of unnecessary word sense disambiguation - terms of a specific field of business are more likely to hold the same meaning than terms in differing fields.

It was also thought, that these companies are varied and large enough in their operations, that there exist a sufficient amount of data, considering both timeframe and quantity, for answering the research questions. Moreover, the focus on a certain field of business - digital cameras - may reflect a more realistic use case for business intelligence, since organizations are assumed to be more interested in the events closest to their own field of actions.

**Table 3.1.** *Used keywords for data gathering*

Acer	Agfa	Benq	Canon	Casio
Contax	DXO	Epson	Exakta	Fujifilm
Fujitsu-Siemens	Hasselblad	Hewlett	Hewlett-Packard	Holga
HP	Jenoptik	JVC	Kodak	Konica
Konica-Minolta	Kyocera	Leica	Lenovo	Logitech
Lytro	Medion	Minolta	Minox	Mustek
Nikon	Nokia	Olympus	Panasonic	Pentax
Polaroid	Praktica	Pretec	Revue	Ricoh
Rollei	Samsung	Sanyo	Sipix	Sony
Toshiba	Umax	Vivitar	Voigtlander	Yakumo
Yashica	Yi			

## 3.2 Data Clean-up

The collected data was then observed and skimmed over in the original text format, and it was noticed that many of the files contained diverse news not related to the company name that had been used as a keyword. For instance, news retrieved with the keyword "canon" contained many news related to religious organizations and biblical canon. Moreover, many sports and entertainment industry news were included due to companies sponsoring arenas and venues. Sales, raffles, and robbery announcements with technology from the companies of interest were included in the news. Furthermore, people with names matching the keywords resulted in news being collected over a great distribution of topics from local gardening clubs to motorcycle gangs in Australia. It was clear at this point that some cleaning and categorization was necessary to be able to collect the news related to business events, since this is the main interest and adding data required more computational capacity, which was limited to begin with.

In addition to this variety of topics, the collected data also contained news in the form weekly and monthly summaries that contained multiple reported events related to the companies of interest. Due to this reason, the following clean up was decided to be done per sentence instead of news article, since it would be rather impossible to categorize a summary article as either a "patent application" or "product launch" if the text contained both events and their related vocabulary.

This was done by importing all the text files of news into Python and splitting them into singular articles via regular expressions, since every news unit began with a numerical title, for example "DOCUMENT 2 OF 5602", which was split with regex syntax "[0-9]+ OF [0-9] DOCUMENTS". This resulted in roughly 350 000 different articles, which were further split into sentences with the NLP library Spacy for Python with the provided `en_core_web_sm` model sentence tokenizer.

Another complicating detail was the uneven distribution of news: patent news are abun-

**Table 3.2.** *Data events and actions as interpreted by the annotator*

Market share announcement	Layoffs	Price predictions
Appointment of personnel	Target announcement	Target achievement results
Negotiations	Pricing changes	Politics and scandals
New facilities	New jobs and positions	Filed for patent
Investments	Secured deals	Granted patent
Lawsuit	Product recalls	Market reactions
Distribution channel changes	Discontinuing products	New product launches
Price predictions	Product changes	Charity events
Acquisitions	Mergers	Events (hackathons etc.)
Complementary products released	Competitions	Geographical availability changes
Training schemes for personnel - current and potential	Accidents	Spin-offs

dant whereas organizational changes and mergers, for instance, are more rare, which was reflected in the number of these news. How these obstacles were dealt with is explained in the following sections.

### 3.3 Categorization of News Data

This initial clean-up was executed via supervised machine learning methods. The collected data was naturally initially unlabelled for the future intentions of using unsupervised learning, which made this step nontrivial. There were millions of extracted sentences, which were not possible to clean by manual human labor in the scope of this thesis. Therefore, active learning methods were utilized in the form of Prodigy (Montani and Honnibal 2018). Prodigy is a web based application for labelling items and teaching a neural network simultaneously. In this thesis, the extracted sentences were labelled whether they were business related or totally irrelevant via the text categorization method of Prodigy, which trained a convolutional neural network model based on a unigram bag-of-words model to assess whether a sentence was business related or not. In other words, whether a data point was interesting for this thesis or not. This enabled reducing the number of sentences from over eight million to just over a million and avoiding creating clusters or topics that are very likely to be irrelevant to business intelligence such as "gardening".

The language model used for categorization was Spacy's `en_vectors_web_lg` language

model. Altogether 5824 sentences, including some few hundred whole news for variety to avoid over-fitting, were annotated "business" or "not business". Annotation was stopped after a while after the suggested texts for annotation displayed relatively accurate estimates of the "business or not" categorization. For example, patent news were recognized as 100% business and family weekend stories were less than 10% probability of being business. Prodigy displays the texts the active learning model is most uncertain of to the annotator, meaning that when it has learned that patent news are most likely of interest, it will stop displaying them for the user of the interface. This process was of course very subjective, and while this was being done, a list of possibly interesting business events was collected by the annotator for possible future reference. The results can be seen in table 3.2.

The convolutional neural network trained with the annotated data displayed an accuracy of 86% when 80% of the data was used for training and 20% for testing. Furthermore, the precision value was 0.86, recall value 0.87, and F-score was 0.86. These results were surprisingly even considering the noticeably disproportionate amount of patent grant and application news events.

### 3.4 Unsupervised methods

To recall yet again the research questions:

- How can unsupervised machine learning methods be exploited in the creation of an action and event categorization framework for business intelligence purposes?
- How do these exploitation possibilities differ in different, common unsupervised method approaches?

The questions are broadly termed, and it is simply impossible to exhaust all the possible different method combinations to use in answering them. For this reason, this thesis aims for a broader perspective that would be available for also any casual reader and user looking to overview the structure business texts for a general picture or to bring more information to creating possible event and action categorization frameworks.

For this reason, it makes sense to gravitate towards open source implementations of clustering and topic modelling that work "out of the box", without having to look too much into all possible parameters and having to tweak them with knowledge on what these parameters do. This is possible with Python and the Scikit-learn and Gensim libraries. Scikit-learn (Pedregosa et al. 2011) is for more straightforward machine learning, whilst Gensim (Řehůřek and Sojka 2010) is simply for topic modelling. Both libraries have methods and algorithms one can import and use straight away within the same code, which was deemed as a suitable approach here.

The broadly termed questions allows deliberation outside of strictly the clustering and topic modelling results themselves, such as the observation that data may be very un-

evenly distributed, or that a learning model may seem to have a hard time differentiating between certain types of texts. Some of these observations may be of interest to the person analyzing and interpreting the data, and some may not. But as they can be interesting, they are a crucial part of answering the presented research questions, which is why these types of observations will also be discussed briefly in the results section later on.

The standard sentence and word tokenizers from Spacy were deemed sufficient - no additional value was thought to emerge from different tokenization methods such as retaining punctuation or upper-casing. Secondly, the resulting token words per business sentence were lemmatized. This choice was made since it was thought that the event vocabulary does not change significantly per event. For instance, "was arrested" and "arrest", "eat" and "ate", still refer to a same type of event. The same logic was applied to stopword removal, and at first the most common 10% and the rarest words appearing only in one document were removed in the preprocessing. However, while testing the preprocessing it was noticed that these parameters did not function intuitively and greatly reduced the vocabulary. Therefore, according to the "minimum parameter tinkering required" logic followed, stopwords were not removed according to frequency.

### **3.4.1 Data Clean-up for Clustering and Topic Modelling**

On top of the typical NLP related preprocessing choices made, some very specific to the clustering and topic modelling were required as well. The aim of these preprocessing steps was to ensure the clustering and topic modelling was done per business related event instead of, for example, per company or industry. For this same reason retaining uppercasing, usually seen in named entities, was deemed counterproductive.

First, the sentences related to business events were filtered out from the rest. This was done by creating a new dataset consisting of the sentences the Prodigy trained CNN model categorized as 85% more more likely to be business related. This percentage was subjectively determined based on empirical notices when studying the sentences along with the percentages while annotating the data and afterwards. News business label probabilities under 70% were often very clearly not related to business events, and sentences with probabilities from 70% to 85% were often ambiguous and impossible to determine the topic based on the sentence alone. A possibility, but not at all a certainty, is that the terms in these sentences are general and vague, but often appear also in business event related texts. From 85% and over the sentences tagged were clearly business event related for a human interpreter from the considered 1000 test sentences - not included in annotation process - categorized and studied.

After this the generated clusters potentially unrelated to business intelligence interests will be fewer, but the issue of possible clustering per company or industry still prevails. It was dealt with censoring tokens that are likely irrelevant to events, but which may result

in clustering per organization or industry: tokens that are proper nouns - "PROPN" POS-tag in Spacy, special characters, persons and named entities, numbers, and emails. All these tokens were located by their Spacy annotated tags and simply deleted from the document. This list is obviously not exhaustive, but this step is thought to contribute to finding more event based clusters than actor or entity based clusters.

These censored documents were then given as data to clustering and topic modelling methods. Due to computational power and time limits, all the over million datapoints were not possible to run all at once. This was dealt with taking a random sample of 29067 texts three times, 2,5% of the data, and comparing them to ensure that the results were reasonably similar in all three cases. This was noticed to be a lot faster than running a larger number of datapoints, since the time complexity of methods like Affinity Propagation increases with sample size significantly (Xu and Tian 2015). Both topic modelling and clustering are used to answer the first research question. Both are different approaches to the same task and can bring differing insights, or similar results, which again would be a good thing to know when planning on how to approach a dataset.

### **3.4.2 N-grams, Chunks, and Vectorization**

As discussed in chapter 2, there exist different approaches to vectorization and data processing on top of the typical preprocessing steps of NLP. TF-IDF is a slightly less straightforward bag-of-words model, but is there a noticeable difference in result interpretability or time requirements depending on the chosen approach? And does this differ by clustering or topic modelling method? In this thesis, both TF-IDF and bag-of-words tokenizers were imported from the Python Scikit-learn library and used as they are "out of the box".

Similar questions apply to N-gram and chunk use. For instance, consider the example of "white" and "house" and "the white house". They hold different meaning whether used as separated tokens or as a noun chunk token. What might be the differences in interpretability and time requirements depending on what type of approach is used? In the scope of this thesis, simple unigrams - "white" and "house" separately - are compared with bigrams - "white house" - and whole noun chunks - "the white house".

The standard unigram tokenization was executed as described next. The text is given to the Spacy language model and the token lemmas from the model output are returned if they are not Spacy determined stopwords or punctuation. For the N-gram tokenization the Python library Textacy, which is built upon Spacy, was utilized. The library automatically extracts N-grams of specified length from the text, processed with the used Spacy model, with an importable method. This thesis used bigrams, N-grams of two tokens, for the tokenization. Similarly, for tokenizing chunks, the Textacy library was used. Textacy allows the automatic extraction of regex patterns via an importable method in a similar manner to the N-gram extraction. Verb chunks were extracted according to the pattern

<VERB>\*<ADV>\*<PART>\*<VERB>+<PART>\*, in addition to which noun chunks were extracted as they are from the Spacy model itself. In both cases, tokens that were not a part of an N-gram or a chunk, were included as they are as lemmas, if they were not stopwords or punctuation according to the Spacy model.

### 3.4.3 Clustering

Different clustering methods were chosen for studying the dataset with. The criteria for choosing was that the chosen set of clustering methods is as different as possible, but that the results are possible to evaluate in a similar manner in order to avoid differences caused by methodological differences instead of clustering algorithm based ones. Centroid based clustering algorithm methods had a the most difference among them, but were comparable with a similar approach. Two methods that do not require a set number of clusters as a parameter and one that requires were chosen out of these.

The K-means clustering algorithm is very popular and well known, which also suits the criteria of "easy to approach". Moreover, a very much comparable algorithm is the Affinity Propagation algorithm, which does not require a number of clusters as a parameter. However, as Affinity Propagation and K-means are similar, another type of algorithm could be compared to them for more information on how different approaches function. The Mean Shift algorithm was chosen as a density based method, that can be comparable by cluster centroids to the two previous partition based algorithms.

All algorithms are available from Scikit-learn as simple import commands to Python, which fit the set criteria: Minimum tinkering with parameters. Only Affinity Propagation needed a provided affinity, which was set as Euclidean distance - comparable to the default in K-means. Otherwise all parameters were as set as default by the library.

### 3.4.4 Topic Modelling

Out of topic modelling methods for comparing, LDA is a popular and referred to as the "state of the art" choice (Mohr and Bogdanov 2013), and is therefore included in the comparisons. It is also at times directly compared to K-means (Zhao et al. 2015; Zhao et al. 2015). HDP is proposed by Blei, Carin and Dunson (2010) to be an improved approach from LDA and LSI that does not require a set number of topics to create, and is compared to LDA with a similar intention as Affinity Propagation is compared to K-means. LSI is also a well known and older (Mohr and Bogdanov 2013) topic modelling approach, and is compared to the previously mentioned two, as it is more distinct from the other two mathematically considered. Parallel, however not directly comparable, to how Mean Shift is mathematically slightly more distinct than K-means and Affinity Propagation in their partition base, compared to Mean Shift's density base.

All topic model algorithms are possible to simply import from the Gensim topic modelling library to Python, and similarly to clustering methods, the minimum of parameter tinkering is done on them after being imported and therefore use the default parameters provided by the library. This defaultness resulted in HDP being limited to a maximum of 150 topics, which it always creates in order of topic importance.

### 3.4.5 Evaluation Methods

These models are run for all the previously explained vectorization and chunk or N-gram combinations with a Python script utilizing the library imported algorithms. The vectorized data is fit for clustering with the Scikit-learn fit method, which is then converted into a required corpus and a dictionary that are required for topic modelling via Gensim's own importable methods "Sparse2Corpus" and "Dictionary.from\_corpus".

For the methods that require a set number of clusters of topics to create, different numbers of this parameter is given beginning from one topic or cluster to 300 with an interval of 15 topics or clusters. The coherences of the created models are plotted and compared using the Gensim imported "u-mass" coherence, since it is the fastest and there is no great seen additional value in using the slower coherence models most alike human judgment, since the results will be evaluated by a human interpreter as well as explained next.

When the models have been fitted as described in the former paragraph, the top ten terms for each topic and cluster, in order of importance, will be printed into a csv-file, where they are compared by human interpretability. For models requiring the number of topics or clusters parameter the minimum number of clusters or topics created by the other models will be the one used for comparison and analysis.

Moreover, the time and computation power required for the methods may be of interest to one looking for a suitable approach. For this reason, random samples of news data are presented to the clustering and topic modelling methods in increasing size and the runtimes per each approach are documented for future comparison.

### 3.4.6 Content Analysis

It is in line with the goals of this thesis to assess how well the done interpretations of the top terms correspond to the actual content of the cluster or topic. In other words, how reliable topic and cluster content determination is based on the top terms. This method is embraced since the goal is not to have to read every item in a cluster or topic for determining the event or theme of the cluster or topic every time after clustering, but rather be able to interpret the clustering or topic modelling results based on the top terms. Otherwise, the whole point of unsupervised learning methods is lost and manual

annotation and labour is required to discouraging extents.

However, in order to assess this top term - content -interpretability, said manual labour and analysis of cluster contents is required. The process follows the logic of direct content analysis (Hsieh and Shannon 2005), in which the clustering and topic modelling results are used as the external framework or theory to be validated or extended. A sample of text documents is ran for the studied clustering and topic modelling algorithms with the same tokenizers as previously for a comparable number of topics and clusters to previous tests. The results for each of the resulting clustering and topic modelling outcomes are sorted by cluster or topic. The text documents in each cluster are read through and the reader codes the cluster as per their interpretation of its contents.

The same process is followed for the topic modelling results, except that due to the probabilistic nature of topic modelling, the documents in a topic are sorted in a way that the ones most representative documents of a topics, according to percentages of belonging, are given more weight in the coding process by the reader. These topics were again read through and coded as by how the reader interpreted the contents of the documents in the topic. Documents with small percentages for any topic were considered to be vague and difficult to assign to any topic for certain.

Since the content analysis is of the direct form, the clusters and topics were used as an external frame of reference. However, the top terms per topic and cluster are not studied while reading the contents of each topic and cluster. They are independently given a content description according to how they were interpreted by the reader.

Afterwards, the coded and named clusters and topics as by their contents are compared to the top terms and words representing the cluster or topic as by the algorithm results. It is then assessed whether the contents of a cluster or a topic are in congruence with the top terms, and an evaluation measure for this congruence is given. The scale used for this was from "in congruence" to "not in congruence", with special attention paid to clear business events and actions. The amounts of these congruence evaluations are compared between the different approaches.

## 4 RESULTS

This chapter reports the results of the previously explained methodology and goes further into the details of each result. The analysis and synthesis regarding the bigger picture of the results is presented in the discussion and conclusions chapter 5 after.

### 4.1 Dimensionality and Run Times

The focus of the results of this thesis is in the interpretability and categorization utility of clustering and topic modelling methods on business intelligence related news. However, observations were made outside of the results themselves, especially regarding the functionality and time requirements of the different approaches to clustering and topic modelling. The most significant of these types of observations are presented here at the beginning of the chapter.

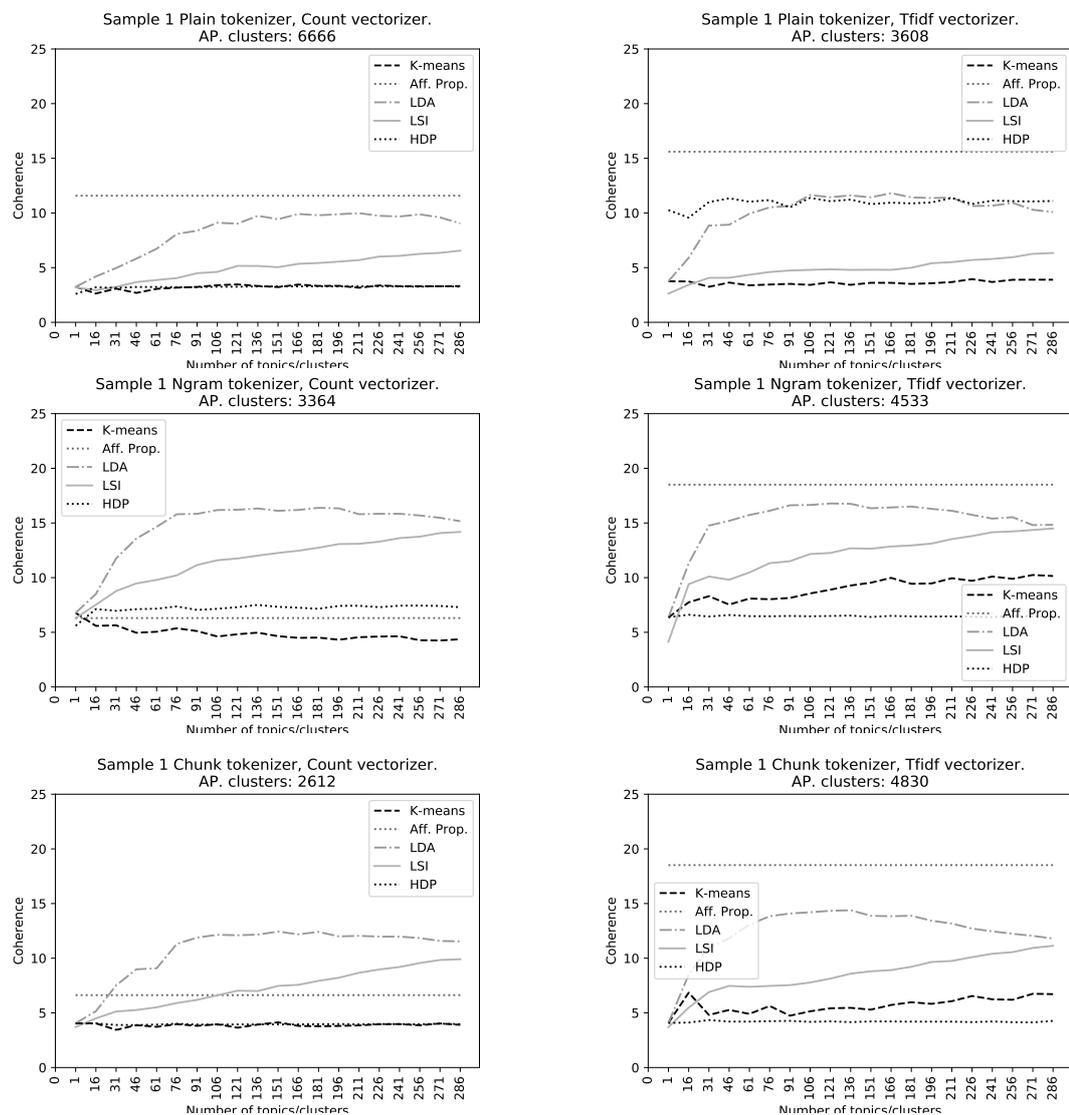
While studying the algorithm run times, it was noticed that the Mean Shift algorithm was outrageously slow. Some runs that took at maximum 15 minutes for the second slowest method, Affinity Propagation, took up to 69 hours with Mean shift. A smaller sample of text documents was tested with Mean Shift, and it did not outperform the other methods in any sense. For this reason, it was concluded that Mean Shift was simply not a feasible method to consider further in this thesis, with larger data sample sizes, due to its time requirements. The behaviors of the other tested algorithms stayed similar throughout all runs. The descriptions and comparisons of these behaviors are summarized in chapter 5 in table 5.1. Moreover, slight time requirement differences existed between different tokenization and vectorization methods. To summarize them, N-grams were typically slowest due to high data dimensionality, followed by chunk tokenization and plain tokenization respectively. TF-IDF vectorization was generally slightly slower than BOW vectorization.

### 4.2 Coherences

Figures 4.1 through 4.3 present the u-mass coherence values for the run algorithms over the range of one through 300 created clusters or topics, with an interval of 15, for three different random samples of 29067 text documents. This sampling was done in order to make sure the coherence values act similarly regardless of sample. Moreover, for the

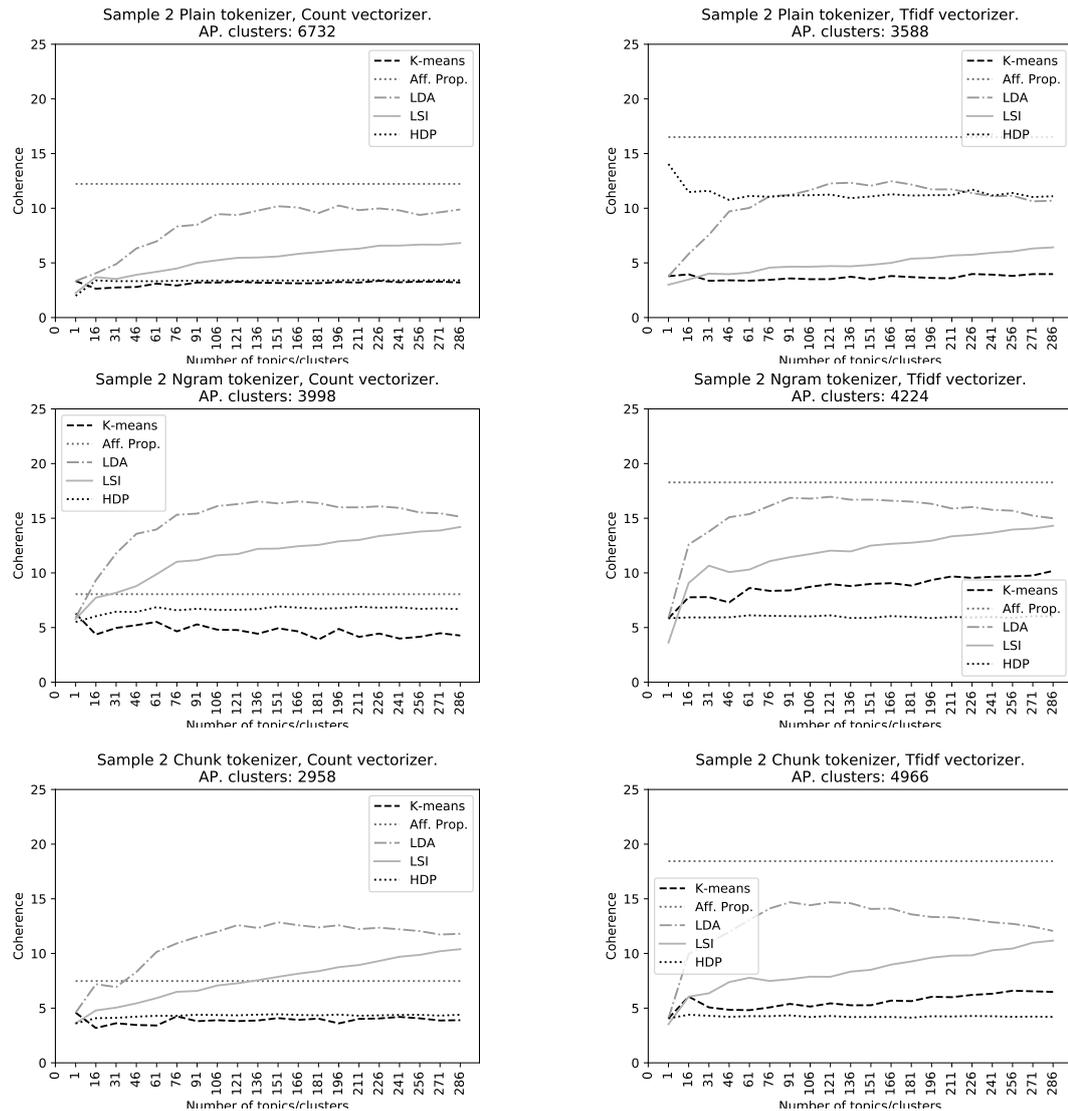
smaller sample of 5813 data points, the MS coherence value was also studied and it was the lowest for all combinations of approaches.

Overall, for the three different random samples, the graphs look very similar. The only major difference being that for sample three, the coherence value for Affinity Propagation is noticeably lower than for the other two samples with plain tokenization. In the figures, HDP was run with the optional number of topics parameter corresponding to the ones given to K-means, LDA, and LSI, for better comparison. For topic numbers after 150 the HDP coherence corresponds to the coherence measure it would give if not given the parameter, so both cases can be studied at the same time. Therefore, Affinity Propagation and HDP can be told apart in the figures by Affinity Propagation being a straight line throughout the figure, while HDP coherence has some slight variation.



**Figure 4.1.** Coherence values for algorithms for different tokenizer and vectorizer combinations - Sample 1

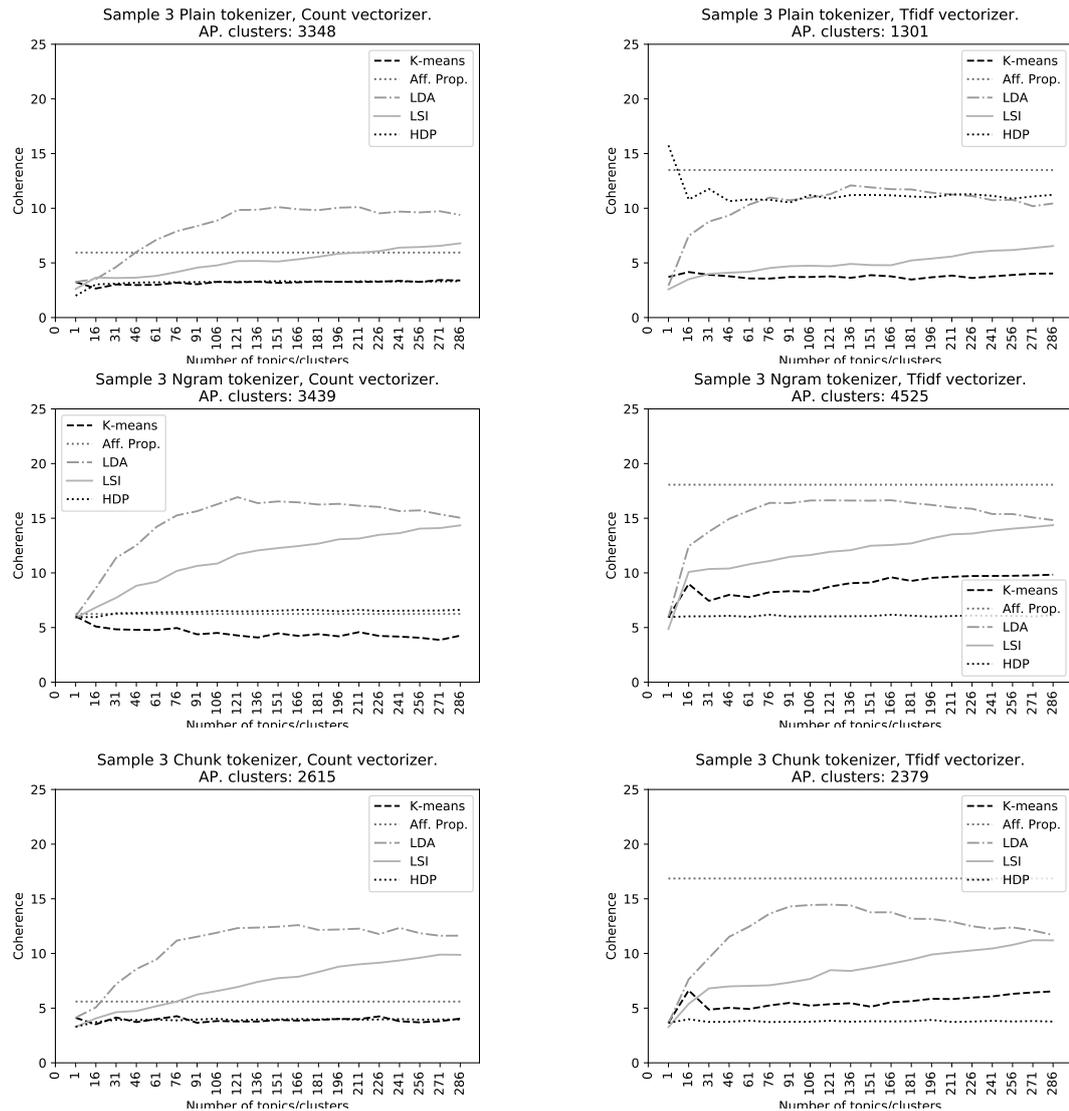
Some general remarks are that HDP stays relatively stable throughout the figures, with an exception with the BOW and TF-IDF vectorizers. An exception to this notice is the



**Figure 4.2.** Coherence values for algorithms for different tokenizer and vectorizer combinations - Sample 2

plain tokenizer combined with the TF-IDF vectorizer. Similarly K-means coherence stays very stable with the exception of the TF-IDF vectorizer combined with N-gram or chunk tokenizer, where the coherence value begins to climb with the number of clusters. LSI coherence invariably always climbs slightly in a rather linear fashion, while always staying above the K-means value and below the LDA coherence value, which in every approach increases faster, but flattens out quite fast as a function of the number of topics. The LSI coherence value stays noticeably smaller with the plain tokenizer.

For every other method than HDP and Affinity Propagation, the coherence values for the BOW vectorizer are smaller than for TF-IDF. For Affinity Propagation in particular, the TF-IDF vectorization coherences are remarkably higher. For the Affinity Propagation cluster numbers, the minimum over the samples is 1301 for plain tokenizer with TF-IDF vectorizer in sample three, and the largest is 6732 for plain tokenizer with BOW vectorizer in sample two. For the BOW vectorizer, the number of Affinity Propagation clusters decreases in



**Figure 4.3.** Coherence values for algorithms for different tokenizer and vectorizer combinations - Sample 3

order from plain to N-gram to chunk tokenizer, whereas the trend is opposite for the TF-IDF vectorizer. The plain tokenizer case for sample three is an exception in this, as it seems to have a much lower number of clusters compared to the other two samples in general.

Overall, K-means typically gives the lowest - best - coherence values, with the exception of HDP in some cases, which is usually the second best, and the best in the TF-IDF cases of N-gram and chunk tokenizers. An exception to HDP giving the second best or best coherence values is the case of TF-IDF vectorizer combined with the plain tokenizer, in which the HDP coherence behaves similarly to the LDA coherence, which is usually the highest value excepting the cases of Affinity Propagation with TF-IDF vectorizer. In the bigger picture, while all figures generally appear similar, it seems that all coherence values are slightly lower for the plain tokenizer than for the chunk tokenizer, whereas N-gram has the largest values.

## 4.3 Topics and Clusters - Top Terms

This section reports the results of a human interpreter examining the top terms per each cluster or topic created by the different methods with differing approaches. As stated before, the number of created topics for analysis was determined by the methods that did not require the number of topics to be a set parameter. Out of these, the minimum number of created clusters or topics was given as a parameter to the methods that required the parameter. Since Mean Shift was discarded due to its slowness, and Affinity Propagation cluster numbers were in the thousands, the number of topics studied per algorithm was the 150 given by HDP.

The same three random samples were taken for this analysis, corresponding to the coherence values. All six algorithms results for the 150 top clusters or topics were read through for their top 10 terms, and evaluated by their comprehensibility for all three samples. Altogether 202 pages of printed Excel tables were read and studied, but since that all would make very little sense to visualize in this thesis, the top 15 topics or clusters for each method are presented for sample three, which seemed to make the most sense, in tables 4.1 through 4.6. However, the written out observations for each approach are addressed for the whole 150 clusters or topics, over all the three samples. In the following tables, "KM" indicates K-means, "MS" indicates Mean Shift, "AP" indicates Affinity Propagation, "C" indicates chunk tokenization, "N" indicates N-gram tokenization, "P" indicates plain tokenization, "T" indicates TF-IDF vectorization, and "C" indicates BOW tokenization. For instance, "CT" stands for chunk tokenization with TF-IDF vectorization.

### 4.3.1 Plain Tokenizer with Bag-of-words Vectorization

For the plain tokenizer combined with BOW vectorization, the top terms per top 15 topics for each considered algorithm are presented in table 4.1. Compared to the other algorithms, in this case, K-means clustering found a noticeable amount of technology specific tokens and terms, and made many clusters based on these technology specific terms combined with business event related terms, and in addition it found a few clear business event clusters. Patent related events were split into a few different clusters based on the terminology used in the news, for example "application", "grant", and "registration" tokens are in clearly patent related, yet different, clusters.

The LDA algorithm created topics which were noticeably less riddled with technology related terms and were more generalisable business events. Obviously, this was not the case for all of the created topics, and there were many nonsensical or technology specific ones. However, when a business event topic was distinguishable as such, it was rather clear, such as declines or rises in sales or stock prices.

The LSI algorithm included loads of repetition in the topic terms compared to the previous methods. For instance, the top eight topics involve patent news, without clear distinction

on what types of patent news events might be involved in each separate topic, and in the latter half of the created topics, the term "medical" began to repeat often, making a disproportionate amount of topics seeming to concern medical technology. Moreover, as in K-means, often technology specific terms were mixed with general event related terms.

HDP seemed to capture the fact that the portion of patent news in the data was rather large. The first topic is a more technology term heavy approach to a patent application or grant, and the second appears to concern market share. All of the following topics are very repetitive versions of patent news, and are nearly totally useless.

Affinity Propagation naturally, due to its significantly larger number of created topics, captured more of singular news and their terms. Meaning, that a cluster's top terms were clearly all from the same piece of text, and unlikely to reoccur in any other text in the dataset. This resulted in the Affinity Propagation clusters looking a lot different from the other algorithms. Outside of these captured very ungeneralisable small clusters, the created clusters were repetitive in their terms, and it was obvious that the algorithm created various clusters of a singular datapoint.

Over all the samples, LSI, HDP, and Affinity Propagation behaved similarly, with the exception of LSI not iterating over the "medical" word, and for sample three, Affinity Propagation found slightly more differing, unrepeatitive clusters compared to the other two samples. K-means behaved similarly in all samples, as well as LDA, but over all the samples, it was clear that the topics LDA found were more sample specific than K-means, in which the terms per cluster were more similar to each other than those of LDA over the three samples.

### 4.3.2 Plain Tokenizer with TF-IDF Vectorization

For the plain tokenizer combined with TF-IDF vectorization, the top terms per top 15 topics for each considered algorithm are presented in table 4.2. Here there was no difference in the behaviour of the HDP algorithm compared to the previous case, and Affinity Propagation behaved even worse - simply repeating the same nonsensical topic over.

In this approach, K-means algorithm created more totally nonsensical, ungeneralisable topics, but simultaneously located now more clear BI event clusters that were not technology reliant. Different types of events such as investments, loans, shipments, development plans, patent lawsuits in addition to patent applications and grants, and even the exploding Samsung phones were distinguishable as event categories.

What was surprising, was that LDA behaved worse in terms of understandability with the TF-IDF vectorizer. It now singled out singular news, and the topics more rarely made sense compared to the BOW vectorization, such as a topic with the tokens "copier", "retire", and "temperature". A possible explanation for this, is that rarer tokens hold more weight, and therefore single out certain news more easily. Moreover, it seemed that K-

**Table 4.1.** Plain tokenizer with count vectorizer. Top 15 clusters and topics.

K-means1	blade	rotor	steam	turbine	fan	have	portion	height	leakage	fix
K-means2	grant	word	mark	trademark	title	trade	pto	kabhiki	plus	contract
K-means3	grant	patent	title	word	method	system	plus	display	manufacture	include
K-means4	patent	title	publish	registration	device	method	word	apparat	apparatus	system
K-means5	image	accord	unit	abstract	release	process	apparatus	title	display	capture
K-means6	light	source	mold	reflector	dispose	encapsulation	surface	reflection	frame	include
K-means7	computer	software	content	video	mobile	phone	download	digital	audio	share
K-means8	word	technology	user	have	announce	software	sell	device	smartphone	phone
K-means9	lens	group	power	refractive	have	positive	optical	surface	component	negative
K-means10	software	computer	datum	management	network	storage	application	hardware	information	database
K-means11	image	frame	object	set	predetermine	process	learn	section	recognize	late
K-means12	image	unit	capture	datum	configure	light	information	control	base	region
K-means13	communication	apparatus	wireless	multiplex	network	time	receive	division	beacon	period
K-means14	share	stantial	buy	\$	sell	earnings	onarket	director	transfer	compute
K-means15	apparatus	apparat	receive	title	word	pto	publish	method	kabhiki	dateline
LDA1	operate	system	sector	public	connection	income	trillion	economic	proprietary	say
LDA2	video	record	disclose	direct	block	box	camera	attention	act	land
LDA3	increase	see	statement	say	international	market	question	decrease	company	race
LDA4	value	current	to"s	overseas	amid	family	conglomerate	phablet	market	individual
LDA5	range	pcs	standard	laptop	maintain	new	trend	wide	suffer	sale
LDA6	percent	revenue	rise	quarter	shipment	attempt	2	company	previous	takeover
LDA7	s	news	package	word	government	opportunity	company	campaign	say	technology
LDA8	lead	internet	tv	key	great	bank	nextgeneration	war	companys	word
LDA9	set	introduce	point	warn	curve	index	word	new	expense	analyze
LDA10	focus	demand	effort	car	cartridge	ongoing	colour	peak	business	depend
LDA11	monitor	own	object	directly	share	web	environment	embed	computer	browser
LDA12	solution	allow	target	go	create	huge	cover	new	engineer	business
LDA13	surface	pattern	extend	recognition	wire	applicant	bond	cover	tender	have
LDA14	manufacture	method	thereof	device	word	rank	title	analysis	grant	patent
LDA15	apparatus	program	method	image	process	control	code	title	medium	multimedia
LSI1	patent	grant	title	word	device	apparatus	method	application	image	computer
LSI2	computer	software	patent	grant	datum	title	word	information	network	management
LSI3	apparatus	computer	software	camera	instrument	unit	image	patent	measure	machine
LSI4	apparatus	image	device	unit	computer	grant	patent	software	instrument	camera
LSI5	word	application	patent	image	company	receive	market	say	unit	file
LSI6	application	company	market	say	have	receive	grant	title	patent	\$
LSI7	file	grant	patent	word	application	publish	image	registration	device	title
LSI8	device	image	application	unit	electronic	word	file	receive	grant	apply
LSI9	\$	company	have	market	say	share	phone	new	mobile	light
LSI10	information	image	light	unit	device	process	datum	communication	computer	software
LSI11	light	image	power	device	unit	include	electric	information	lens	have
LSI12	phone	company	publish	title	file	registration	mobile	datum	word	have
LSI13	electric	machine	power	light	image	publish	registration	camera	tool	datum
LSI14	light	market	information	publish	file	registration	phone	title	device	unit
LSI15	market	company	information	share	unit	have	datum	light	process	phone
HDP1	device	image	unit	title	include	apparatus	information	patent	accord	abstract
HDP2	company	market	say	word	have	new	\$	sale	share	business
HDP3	title	patent	application	word	publish	receive	device	grant	method	file
HDP4	patent	grant	title	word	device	file	method	application	publish	apparatus
HDP5	patent	word	title	grant	file	device	company	say	application	market
HDP6	patent	word	title	company	computer	grant	device	market	have	software
HDP7	patent	word	title	market	grant	company	say	have	file	device
HDP8	patent	word	title	file	grant	application	device	company	method	publish
HDP9	patent	word	title	grant	device	file	company	say	application	new
HDP10	patent	word	title	grant	device	company	file	application	have	market
HDP11	patent	word	title	grant	company	device	file	market	say	application
HDP12	word	patent	title	company	say	grant	market	new	application	file
HDP13	patent	word	title	company	market	grant	say	application	file	device
HDP14	patent	word	title	company	say	grant	market	have	file	device
HDP15	patent	word	title	grant	company	say	device	file	market	application
AP1	new	shoot	pixel	software	support	design	call	pc	get	imagery
AP2	terminal	video	network	configure	mobile	information	phone	unit	communication	controller
AP3	portion	end	portiontobefixed	blade	drum	contact	fix	include	member	outside
AP4	enterprise	build	directory	solution	mobile	resource	print	map	deploy	environment
AP5	boost	camera	say	tourism	leverage	maker	word	sale	boom	ecommerce
AP6	nm	wavelength	record	method	medium	accord	range	absorbance	information	release
AP7	production	help	horticulture	boost	word	section	project	development	show	decide
AP8	communication	system	path	node	supervise	comprise	solution	method	+	send
AP9	band	multifrequency	electronic	device	support	response	operate	release	provide	title
AP10	process	grant	title	apparatus	image	word	method	patent	exabytes	ews
AP11	company	business	acquire	maker	intelligence	opportunity	artificial	future	jail	de
AP12	of"s	?&A?Z?e?è	ewallets	examiner	examine	examination	exam	exainer	exaggerate	exactly
AP13	column	value	sense	sensor	element	provide	system	release	row	device
AP14	monitor	transmission	status	unit	relay	frequency	information	probability	record	device
AP15	maker	near	accord	purchase	familiar	steak	word	person	sauce	deal

means and LDA flipped over roles now, since LDA now incorporated more technical terms in its topics than K-means in its clusters. In this case also for LDA, there was a difference in topics regarding the terms used for patent grants and applications.

LSI behaved somewhat similarly now as for the BOW vectorizer, creating a smooth transition from topic area to another, without the topics being distinguishable from each other in these topic areas. To explain, it seemed that the first ten topics dealt with patents, which transitioned into product launch topics, which again slowly transitioned into sales events. Topic top terms again included many technology specific terms. However, again, there seemed to be no clear event categories in these results - just the general topics of discussion, which is exactly what topic modelling is set out to do.

Over all the samples LSI, HDP, and Affinity Propagation behaved very similarly. K-means seemed to adopt LDA behavior from the BOW case, since it captured the different sample now better, by including more specific event categories not seen in the other samples, such as a tablet launch, share purchases, a licensing agreement, and a software breach. LDA behaved similarly across all samples, but seemed to get slightly more coherent with the second and third samples, often having sensible topics, but at times incorporating very specific terms such as "dud" and "nimble".

### 4.3.3 N-gram Tokenizer with Bag-of-words Vectorization

For the N-gram tokenizer combined with bag-of-words vectorization, the top terms per top 15 topics for each considered algorithm are presented in table 4.3. For the N-gram tokenization, it was more difficult to find any difference in interpretability between the K-means and LDA algorithms. For instance, both involved a roughly similar amount of technology specific terms.

Altogether, both seemed to make less sense compared to the plain tokenizer, in the way that both incorporated more seemingly irrelevant terms in the clusters and topics, which replaced the previously separate tokens in the cluster or topic, that had now become a combined N-gram. This made some technology and business related terms clearer, such as "market\_share", "launch\_event", "volume\_leadership", and "smartphone\_lineup", which might be more difficult to interpret as separate tokens, but in the bigger picture this did not aid the interpretability of the method. As an interesting sidenote, K-means managed to create a cluster of news in German.

The LSI and HDP algorithms behaved very similarly as in the case for the plain tokenizer. LSI incorporated the odd sensible N-gram in the topics, whereas HDP recognized "grant\_patent" and "company\_have" as N-grams in its repetitive patent topic. Affinity Propagation on the other hand included a significant number of N-grams in its clusters compared to the other methods. Some clusters contained multiple N-grams with the other half of the bigram being always the same one, for reference see cluster AP6 in table 4.3. Otherwise it acted similarly to the case of plain tokenizer combined with the BOW vectorizer - creating very many clusters of ungeneralisable singular text documents, often technology specific.

**Table 4.2.** Plain tokenizer with TF-IDF vectorizer. Top 15 clusters and topics.

K-means1	user	allow	device	interface	access	let	say	application	mobile	provide
K-means2	grant	plus	patent	title	word	device	method	apparatus	kabhiki	image
K-means3	file	patent	evoting	examine	examination	exam	exainer	exaggerate	exactly	exact
K-means4	apply	receive	application	title	word	device	method	length	thereof	memory
K-means5	invention	title	container	p	present	embodiment	material	3d	harmonic	content
K-means6	registration	publish	title	patent	method	system	memory	thereof	kabhiki	
K-means7	world	have	maker	large	big	mobile	smartphone	phone	market	company
K-means8	application	file	patent	publish	word	fuel	cell	ews	ex	ex2
K-means9	open	store	word	new	office	market	have	company	say	source
K-means10	machine	wash	appliance	word	sell	say	refrigerator	new	home	product
K-means11	grant	patent	title	word	device	method	system	communication	control	display
K-means12	announce	word	new	addition	company	availability	release	offer	partnership	launch
K-means13	sell	company	smartphones	say	share	worldwide	expect	right	pcs	tvs
K-means14	head	say	business	company	have	of's	market	unit	division	mobile
K-means15	technology	company	say	product	new	use	word	market	patent	giant
LDA1	make	component	g	fact	download	colour	correct	error	arrive	
LDA2	local	begin	term	government	surface	overtake	market	have	flexible	applicant
LDA3	price	fall	low	drop	share	competitor	beat	life	cause	market
LDA4	apparat	ing	kabhiki	enable	application	publish	image	pickup	include	pto
LDA5	note	upgrade	reporter	speaker	dedicate	analyst	remark	smartphone	ouster	dualsim
LDA6	news	contract	office	slow	wave	contain	contrast	sit	collaborate	trail
LDA7	site	cell	exchange	web	appear	yen	refrigerator	round	sponsorship	participate
LDA8	turn	recall	strategic	copy	engineer	cite	use's	reduction	global	
LDA9	conference	tell	fiscal	press	conduct	touchscreen	save	house	half	year
LDA10	partner	smart	follow	online	leader	claim	aim	try	market	vendor
LDA11	subject	widely	difficult	prove	prior	relevant	apparent	bar	exhibit	irregularity
LDA12	+	generation	number	website	globally	currently	factory	hand	send	filter
LDA13	end	member	high	complete	part	jump	cartridge	thin	collaboration	year
LDA14	advance	function	oled	will	effectively	fuel	attention	display	successor	gas
LDA15	dateline	kabhiki	module	live	concentrate	pad	augment	system	reward	spark
LSI1	file	patent	application	grant	title	publish	registration	word	device	method
LSI2	title	grant	file	word	publish	application	registration	receive	device	apply
LSI3	grant	application	receive	apply	patent	publish	pto	apparat	apparatus	file
LSI4	publish	registration	grant	application	word	receive	apply	patent	title	file
LSI5	company	application	market	device	say	share	grant	have	publish	pto
LSI6	application	apply	publish	pto	title	receive	grant	registration	market	company
LSI7	process	method	device	apply	information	image	apparatus	plus	word	application
LSI8	device	apparat	process	apparatus	application	method	information	image	publish	memory
LSI9	registration	publish	application	word	share	device	pto	plus	market	title
LSI10	share	device	market	registration	new	company	\$	cent	own	publish
LSI11	apparat	apparatus	process	plus	pto	information	device	grant	method	application
LSI12	plus	apparatus	apparat	process	word	information	title	image	pto	patent
LSI13	market	company	profit	\$	launch	new	phone	sale	net	smartphone
LSI14	image	plus	information	process	apparatus	form	apparat	system	device	unit
LSI15	company	sale	share	market	profit	have	own	unit	expect	percent
HDP1	patent	file	grant	title	application	word	publish	apply	language	videoediting
HDP2	patent	file	grant	title	word	application	bt34	discern	browser	tinkerers
HDP3	patent	file	grant	title	word	application	publish	method	device	architecturebased
HDP4	patent	file	grant	title	application	word	publish	connection	offshore	device
HDP5	patent	file	grant	title	application	word	publish	new	clarity	atent
HDP6	patent	file	grant	word	title	trader	application	wo201201578	publish	aircontrol
HDP7	patent	file	application	title	kabhiki	grant	word	publish	unbaked	apply
HDP8	patent	file	grant	word	title	application	everchanging	interceptive	sale	biconvex
HDP9	patent	file	grant	title	kingpin	application	word	asymco	publish	exclusionary
HDP10	patent	file	grant	word	application	grant	depress	publish	counter	receive
HDP11	patent	file	grant	title	word	application	asked's	device	biness	publish
HDP12	patent	file	title	grant	application	word	measurable	nanostructure	jobless	publish
HDP13	patent	file	title	grant	word	application	publish	capable	registration	of's
HDP14	patent	file	grant	title	turnkey	application	suggest's	neighborhoodthe	word	elaborate
HDP15	patent	file	grant	title	application	word	solargenerated	globe	publish	device
AP1	registration	publish	title	patent	ewaste	ews	ex	ex2	exabytes	
AP2	portion	portiontobefixed	end	drum	blade	contact	fix	bend	include	respect
AP3	enterprise	directory	rich	tag	deploy	map	resource	environment	build	printer
AP4	registration	publish	title	patent	ewaste	ews	ex	ex2	exabytes	
AP5	column	sense	element	sensor	value	row	sum	array	touch	multiple
AP6	steak	maker	sauce	unite	person	familiar	matt	near	purchase	deal
AP7	supplychain	analysis	account	make	revenue	of's	customer	large	accord	have
AP8	registration	publish	title	patent	ewaste	ews	ex	ex2	exabytes	
AP9	launch		evoting	examination	exam	exainer	exaggerate	exactly	exact	exabytes
AP10	registration	publish	title	patent	ewaste	ews	ex	ex2	exabytes	
AP11	publish	patent	evoting	examination	exam	exainer	exaggerate	exactly	exact	
AP12	spin-dry	exit	tv	personal	computer	unit	business	exact	ex	ex2
AP13	recordingreproducing	apply	receive	application	device	title	word	evident	ex2	example
AP14	trademark	grant	mark	trade	title	word	ex	ex2	exabytes	exact
AP15	registration	publish	title	patent	ewaste	ews	ex	ex2	exabytes	

### 4.3.4 N-gram Tokenizer with TF-IDF Vectorization

For the N-gram tokenizer combined with TF-IDF vectorization, the top terms per top 15 topics for each considered algorithm are presented in table 4.4. Between the BOW and

**Table 4.3.** N-gram tokenizer with count vectorizer. Top 15 clusters and topics.

K-means1	image_apparat	publish	registration	patent	title	feed_device	open	cassette	computer_readable_storage	storage_medium
K-means2	image	device	display	picture	memory	title	acquire	accord	response	user
K-means3	file	patent	unit	plurality	abstract_publish	abstract_release	apply	method	include	present_invention
K-means4	possibility_achievable	reverse_texture	demand_special	application_create	include_unique	gloss_effect	eyecatching_gloss	engage_application	atfs_booth	effect_matt
K-means5	publish_patent	pto_publish	patent_application	title	title_word	word	sony_title	length_word	olympus_title	device_word
K-means6	company	corfirm	company_have	mega_job	walk	change	originalise	smartphone	campus	honor
K-means7	authentication	perform	information_correspond	select_external	biometric_authentication	relation	authentication_base	processor	receive	transmit
K-means8	grant	grant_patent	title	system_word	communication_system	apparatus	method	wireless_communication	device	memory_system
K-means9	page	say	voltage	block	activate	count	apply	number	plurality	gate
K-means10	movable	guide	fix	si_accord	shape	camera_module	open	optic_axis	form	ro_gu
K-means11	appar	title	receive	application	apply	apparatus	method	output_system	manufacture	plus_grant
K-means12	grant_patent	title	method_word	process_method	information_process	process_apparatus	program_word	process_system	process_device	method_word
K-means13	receive	title	pto_publish	publish_application	apparatus	word_dateline	method	length_word	word	diode_display
K-means14	grant	manufacture	title	grant_patent	method	word	memory_device	light_emit	package	execute
K-means15	apparatus	abstract_release	title	update	determine	perform	response	provide	apparatus_include	execute
LDA1	use	list	prevent	unit_have	digital	equipment	world_leader	computer_industry	simple	apparatus_information
LDA2	enable	mainland	cooperate	acquire_image	qualify	control_method	image_pickup	image_control	process_apparatus	launch
LDA3	sell	result	claim	lose	division	represent	trillion_win	image_apparatus	accordance	detail
LDA4	country	research_firm	market_research	firm_say	prove	scandal	big_corporate	server_market	of's	handset_unit
LDA5	customer	group	vision	innovation	investigate	celebrate	vendor	new_tablet	of's_sale	-
LDA6	company	consumer	leave	leader	split	share_price	emerge	choice	say	highend_market
LDA7	surface	user_interface	abstract_release	pickup_apparatus	callphone_maker	inventor_applicant	bond	determination	expose	light_receive
LDA8	operate_system	release	increase	price	platform	stock	reach_unit	action_involve	wait	family
LDA9	firm_have	abstract_release	unit_configure	title	plurality	view	power_supply	type	chance	assemble
LDA10	base	subsidiary	service_provider	surge	gear	patent_portfolio	recover	export	expect_sale	strike
LDA11	market_share	make	decide	workstation	source_say	spend	factory	dvd_recorder	profit_margin	proth
LDA12	method_word	process_method	information_process	process_apparatus	process_device	title	rank	pc_unit	process_image	new_printer
LDA13	record_medium	perform	request	software_company	inventor	block	3d	renew	information_record	prior
LDA14	profit	easy	core	fiscal	overall_market	startup	emphasis	cloud	stock_close	enjoy
LDA15	abstract_release	title	abstract_release	accord	website	abstract_publish	method	stake	supply	#NAME?
LS1	title	grant	grant_patent	patent	receive	application	registration	word	file	publish_device
LS2	patent	grant	grant_patent	patent	publish	registration	application	title	receive	device
LS3	grant	receive	title	application	file	apply	grant_patent	patent	word	pto_publish
LS4	registration	publish	file	application	receive	word	apply	grant_patent	grant	patent
LS5	word	application	company	grant	publish	method	apply	grant_patent	say	file
LS6	word	abstract_release	accord	device	publish	registration	title	file	share	method
LS7	company	say	application	word	grant	file	market	title	expect	apply
LS8	pto_publish	publish_application	application	receive	device	company	abstract_release	word_dateline	accord	apply
LS9	device	title	abstract_release	accord	grant	receive	company	pto_publish	publish_application	company
LS10	say	company	grant_patent	grant	share	expect	launch	plus_grant	device	company_have
LS11	grant_patent	grant	say	accord	plus_grant	abstract_release	title	publish	word	word
LS12	share	\$	say	launch	method	expect	market	company_have	cent	company
LS13	method	share	launch	\$	device	apparatus	manufacture	application	title	company
LS14	launch	method	share	\$	market	expect	device	word	say	application
LS15	company_have	launch	use	market	share	expect	sale	application	say	abstract_release
HDP1	title	patent	registration	publish	receive	file	application	grant	word	grant_patent
HDP2	grant	title	grant_patent	patent	word	method	file	receive	plus_grant	application
HDP3	title	grant	patent	file	grant_patent	word	application	receive	abstract_release	device
HDP4	title	receive	application	apply	word	patent	grant	grant_patent	file	pto_publish
HDP5	title	grant	patent	grant_patent	file	word	receive	application	registration	publish
HDP6	title	patent	file	publish	registration	device	grant	grant_patent	word	application
HDP7	title	patent	receive	file	application	word	grant	grant_patent	apply	pto_publish
HDP8	title	patent	grant	file	grant_patent	word	application	receive	say	publish
HDP9	title	patent	file	grant	grant_patent	word	application	abstract_release	receive	method
HDP10	title	patent	file	grant	grant_patent	word	application	receive	company	say
HDP11	title	grant	grant_patent	patent	file	application	word	receive	method	registration
HDP12	file	title	patent	share	say	grant	company	word	grant_patent	stock
HDP13	title	patent	file	grant	grant_patent	application	word	receive	publish	abstract_release
HDP14	title	patent	file	grant	grant_patent	publish	registration	form	word	application
HDP15	title	grant	patent	grant_patent	file	word	application	receive	company	say
AP1		lead_player	personal_computer	global_smartphone	star	computer_giant	shoot	smartphone_sale	sale_build	aim_high
AP2	word_dateline	pto_publish	apparatus	receive	publish_application	title	fiberlike_speed	fiberlike_speed	fiber_network	fiber_network
AP3	shoot	new	support	new_mode	composite_image	free_download	resolution_composite	compatible_tv2019s	suite_call	raw_developmentbox0
AP4	application	plus_apply	word	receive	title	field_communication	fiber_network	field_communication	fiddle	phone_receive
AP5	video	protocol_ip	mobile_phone	ip_network	conduct	controller_configure	communication_unit	transmit_network	terminal_include	fiber_network
AP6	registration	patent	title	publish	field	fiberbased_fx	fiberlike_speed	fiber_optic	fiber_network	fickle
AP7	portiontofixed	drum	blada_portion	surface	provide	member_include	support_member	portion_include	remote	outside
AP8	enterprise	print_solution	tag_printer	corporate_environment	resource_deploy	build	enterprise_mobile	mobile_print	rich_directory	map
AP9	nm	wavelength	absorbance	range	vicinity	dye_base	embodiment	record_material	record_accord	record_method
AP10	manager_of's	vice_president	wireless_access	distance_away	run_service	mobile	service_like	rest	era	general_manager
AP11	state	help_boost	show	keenness	apple_production	development_project	production_section	decide	horticulture_agriculture	sector_government
AP12	node	supervise	solution	method	+	path_set	path_establish	initiate	cellular_network	determine
AP13	business_opportunity	mount_uncertainty	word_amid	brass	jail	facto_leader	company_say	vacuum_deal	blow	future
AP14	develope	build_guard	line	office_build	new_antibiotic	woman	succeed	go	new_need	lingerie
AP15	registration	patent	title	publish	field	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fickle

TF-IDF vectorizers for the N-gram tokenizer, these existed less difference in the results compared to the case of the plain tokenizer results. The HDP and LSI algorithms behaved as in every previous case, except now the results for LSI seemed to incorporate more bigrams in the created topics than for the BOW vectorization case. Affinity Propagation was again very much more repetitive and redundant for the TF-IDF vectorizer case, similarly as with the plain tokenizer.

Regardless, the same trends could be noticed as previously for the K-means algorithm: The TF-IDF vectorizer case incorporated less technology specific terms and made the division between technology and event related clusters clearer. This did not however make the clustering result any more intelligible, and there were many nonsensical clusters with only a few of the clear event or technology text clusters among them. However, an interesting observation is that K-means with N-gram tokenization and TF-IDF vectorization was the only method combination to capture a clear cluster of sports related tokens in the first sample. Otherwise, the results were similar across all samples, with sample

two including slightly more technology terms in the created topics and clusters, which is very likely just due to the sample of news containing more texts with more technology vocabulary.

The LDA algorithm came across as more intelligible for the TF-IDF vectorizer for N-gram tokenization, as opposed to the plain tokenization, in which LDA results were more coherent for BOW vectorization. The TF-IDF vectorization incorporated less technology specific terms per created topic, and seemed to be the overall best approach in terms of N-gram tokenization.

**Table 4.4.** N-gram tokenizer with TF-IDF vectorizer. Top 15 clusters and topics.

K-means1	revenue	term	expect	year	billion	contribute	say	growth	previous_forecast	method
K-means2	device_word	grant_patent	title	plus_grant	electronic_device	memory_device	display_device	semiconductor_device	device_method	represent
K-means3	registration	publish	patent	title	method	image_apparat	device	system	apparat_method	storage_device
K-means4	file	patent		field_apply	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle	fiddle
K-means5	application	file	patent	field_apply	field_apply	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle
K-means6	launch	say	smartphone	model_lose	leadership_late	smartphone_volume	volume_leadership	create	year	smartphone_market
K-means7	conduct	company_try	consistent	technology	's_code	human_right	court	legal_justification	portray	stifle_competition
K-means8	expect	say	end	year	analyst	profit	sell_unit	close	company	transaction
K-means9	device_word	apply	receive	application	title	memory_device		image_device	method	lightemitting_device
K-means10	customer	offer	company	market	purchase	say	company_have	refund	swap	region
K-means11	share_fall	\$	close	fall_cent	's_share	company_have	cent	rebound_slightly	low	trade
K-means12	form_apparatus	grant	image_form	apparatus_word	grant_patent	title	word_dateline	device	process_cartridge	form_method
K-means13	title_word	publish_patent	olympus_title	sory_title	plus_patent	line_publish	memory_title	plus_publish	publish	patent
K-means14	publish_application	pto_publish	receive	title	word_dateline	word	apparat	apparat_word	length_word	device_word
K-means15	predict	screen	loss	£	make_headway	42-inch_screen	chip	current_correction	outperform	outperform
LDA1	company_have	help	way	difficult	new_business	prove	design_patent	control_program	say	control_apparatus
LDA2	company_say	crystal_display	problem	fiscal_year	liquid_crystal	game	compute_device	scale	of's_sale	say_note
LDA3	form_apparatus	image_form	apparatus_word	image_word	hold_company	global_recall	service_provider	device	attention	sell_smartphones
LDA4	hold	datum	reserve	new_device	apparatus_image	module_word	accord	customer_demand	lose_grind	radio_frequency
LDA5	model	configure	new_technology	strategy	mobile_broadband	hardware_maker	spend_\$	mobile_ph	glance	buy_found
LDA6	report_earnings	far	feature_phone	3d	competitor_like	new_solution	eliminate	international_sale	builtin_capability	global_reach
LDA7	mobile_phone	talk	phone_market	partnership	company	segment	rate	similar	market	act
LDA8	apparat_word	agree	maker_say	consider	week	sell_share	article	notebook	user_experience	outsource
LDA9	focus	device_have	rival	struggle	capture	patent_publish	recently_launch	plus_patent	recent_year	web_site
LDA10	gain	position	manufacturer	big	panel	size	economy	thick	pioneer	new_plant
LDA11	group	light_source	people	adjust	variant	revise	report_acquisition	instance	company_include	collapse
LDA12	publish_patent	patent_application	pto_publish	title_word	word	reduce	title	place	announce_plan	cover
LDA13	sell	product	build	store	open	present_invention	say	part	protect	part
LDA14	market_share	deal	aim	cost	total	manage	leader	esk_word	section_esk	dvd_recorder
LDA15	profit	mobile_device	early	profit_fall	market_leader	switch	worth	expect	resign	choice
LS1	file	patent	application	registration	publish	title	device	receive	apply	issue
LS2	title	publish	registration	application	receive	grant	file	word	apply	grant_patent
LS3	application	publish	registration	receive	apply	word	patent	file	plus_apply	grant
LS4	grant	grant_patent	application	publish	registration	receive	apply	title	file	apparatus_word
LS5	application	pto_publish	publish_patent	publish_application	grant	word	receive	title	publish	patent_application
LS6	publish_patent	patent_application	application	pto_publish	word	grant	title	apply	publish	grant_patent
LS7	word	publish_patent	publish_application	pto_publish	receive	grant_patent	grant	plus_grant	word_dateline	apparat
LS8	patent_grant	plus_patent	word	grant_patent	publish_patent	publish_application	grant	pto_publish	plus_grant	receive
LS9	apply	pto_publish	publish_application	grant_patent	publish_patent	application	word	publish	plus_grant	receive
LS10	word	device	publish	plus_grant	device_word	grant_patent	grant	title	apply	application
LS11	device	device_word	word	plus_grant	grant_patent	method	title	grant	apply	apparat
LS12	publish	registration	device_word	device	apply	issue	plus_grant	title	apparat	plus_apply
LS13	apparat	device_word	method	plus_grant	method_word	registration	apply	plus_apply	word	publish
LS14	patent_application	share	internationally_file	company	publish_patent	say	apply	device_word	plus_apply	\$
LS15	patent_application	share	company	internationally_file	say	publish_patent	\$	plus_apply	apply	launch
HDP1	patent	file	title	grant	application	grant_patent	publish	word	receive	registration
HDP2	file	patent	application	title	grant	receive	grant_patent	apply	word	method_word
HDP3	patent	file	title	application	grant	publish	grant_patent	registration	receive	word
HDP4	patent	file	title	application	grant	publish	grant_patent	publish	registration	receive
HDP5	patent	file	title	application	grant	publish	grant_patent	word	registration	receive
HDP6	patent	file	title	application	grant	publish	grant_patent	receive	registration	word
HDP7	patent	file	title	application	publish	grant	registration	word	receive	grant_patent
HDP8	patent	file	title	application	publish	grant	registration	word	receive	receive
HDP9	patent	file	title	application	publish	grant	registration	word	receive	grant_patent
HDP10	patent	file	title	application	publish	grant_patent	publish	registration	word	receive
HDP11	patent	file	title	application	publish	grant	registration	grant_patent	receive	word
HDP12	patent	file	title	application	publish	grant	registration	grant	receive	grant_patent
HDP13	patent	file	title	application	grant	word	publish	grant_patent	registration	receive
HDP14	patent	file	title	application	grant	word	receive	publish	registration	grant_patent
HDP15	patent	file	title	application	grant	publish	receive	registration	word	grant_patent
AP1	file	patent		field_apply	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle	fiddle
AP2	file	patent		field_apply	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle	fiddle
AP3	file	patent		field_apply	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle	fiddle
AP4	registration	publish	patent	title	field	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle
AP5	file	patent		field_apply	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle	fiddle
AP6	application	file	patent	field_apply	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle	fiddle
AP7	enterprise	tag_printer	corporate_environment	resource_deploy	enterprise_mobile	rich_directory	mobile_print	print_solution	map	build
AP8	file	patent		field_apply	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle	fiddle
AP9	file	patent		field_apply	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle	fiddle
AP10	file	patent		field_apply	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle	fiddle
AP11	registration	publish	patent	title	field	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle
AP12	column	multiple_sense	sensor_value	sum_sensor	panel_system	sense_element	row	touch_panel	array	inventor
AP13	file	patent		field_apply	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle	fiddle
AP14	file	patent		field_apply	fiberbased_fix	fiberlike_speed	fiber_optic	fiber_network	fiddle	fiddle
AP15	maker	steak_sauce	matt_unite	person_familiar	near	purchase	deal	accord	word	fight_like

To summarize thus far, N-gram tokenization was difficult to read and analyze compared to the plain tokenization case for both instances of vectorization. The main reason for this being that N-gram often mixed very specific bigrams with seemingly unrelated unigrams

in the clusters and topics. This raised questions on how are the topic or cluster tokens related to each other, and these connections were more difficult to make than for plain tokenization.

### 4.3.5 Chunk Tokenizer with Bag-of-words Vectorization

For the chunk tokenizer combined with bag-of-words vectorization, the top terms per top 15 topics for each considered algorithm are presented in table 4.5. For this case of the K-means algorithm, many very specific ungeneralisable - both technology and event wise - terms and chunks were incorporated in the clusters: clusters 37, 61, and 84 incorporated tokens "method\_storage\_medium\_exposure\_apparatus\_exposure\_method", "lead\_encapsulation\_resin\_body", and "reassessment\_of's\_overall\_business\_profile".

For the LDA algorithm, the case was opposite. The topics created by LDA were much more sensible than for the N-gram methods, and did not include overly specific long chunks. "Market\_share", "patent\_application", and "new\_supply\_contract" are examples of chunks that brought more information compared to even plain tokenization in the LDA results.

**Table 4.5.** Chunk tokenizer with count vectorizer. Top 15 clusters and topics.

K-means1	come	expect	word	like	time	available	release	company	announce
K-means2	grant	title	patent	word	plus	include	device	equip	form_apparatus_word
K-means3	action_assistance_method	registration	patent	action_assistance_device	title	flat_panel_display_device	flat_plate_member	flat_screen_division	flat_screen
K-means4	patent	publish	application	issue	word	plus	title_word	word	sony
K-means5	file	patent	infringe	include	lawsuit	perform	provide	year	suit
K-means6	share	company	fall	rise	close	's	new	compare	percent
K-means7	launch	word	announce	market	set	new	request	design	unveil
K-means8	end	server	cloud	indicate	extend	process	receive	heat_dissipator	response
K-means9	company	say	chief_executive	president	word	aggressive_market_campaign	replace	part_supplier	labrate
K-means10	say	company	expect	sale	statement	include	set	plan	continue
K-means11	stock	share	disposition	involve	action	conventional_film	sell	\$	file
K-means12	grant	patent	plus	patent_patent	word	trademark	title	federal_contract	worth
K-means13	network	pocket	includes	extract	associate	identifier	identify	programmable_hardware_processor	programmable_logic
K-means14	win	trillion	operate_profit	extract	associate	sale	accord	post	report
K-means15	word_datetime	receive	application	publish	application	pto	kabuki	method	device
LDA1	business	drop	confirm	event	company	break	oversee	uncertainty	scale
LDA2	receive	application	title	apply	word	publish	method	apopt	kabuki
LDA3	recently	million	segment	interview	pixel	near	recognize	fingerprint	include
LDA4	get	network	be	keep	euro	electronic	engineer	company	say
LDA5	provide	computer	link	link	serve	relationship	clearly	service	entry
LDA6	help	drive	attempt	job_out	expectation	company	company	say	word
LDA7	require	concern	investor	asset	digital_camera	possible	handset_maker	reason	combination
LDA8	base	store	equation	beat	memory	remove	include	laurel	approach
LDA9	less	electronic_giant	beat	beat	worker	grind	heat	compensate	company
LDA10	available	generate	prevent	include	host	record_medium	current	specify	big_company
LDA11	acquire	term	apparatus	control_method	material	stock_market	user_equipment	discuss	movement
LDA12	design	module	effectively	wirelessly	wide_range	spot	error	boss	cheap
LDA13	market	say	suffer	player	company	hire	word	disrupt	product
LDA14	numb	find	task	achieve	fire	battery	rank	catch	disrupt
LDA15	produce	stake	skid	format	smartphone_sale	company	alsoipable	record_profit	say
LSI1	grant	patent	title	word	publish	application	application	registration	receive
LSI2	grant	patent	receive	application	title	say	word	patent	include
LSI3	file	grant	publish	application	registration	application	title	word	plus
LSI4	say	receive	title	application	company	file	patent	apply	grant
LSI5	\$	say	company	share	include	application	have	receive	release
LSI6	include	accord	word	related	configure	application	configure	apply	provide
LSI7	publish	file	registration	word	include	application	receive	apply	grant
LSI8	say	company	word	include	have	sale	application	analyst	expect
LSI9	word	application	receive	say	publish	grant	announce	share	registration
LSI10	company	say	announce	have	apply	abstract	launch	file	share
LSI11	include	accord	abstract	release	+	provide	file	have	title
LSI12	have	measure	control	melter	process	configure	pressure	sensor	publish
LSI13	have	measure	control	melter	include	process	pressure	configure	magnetic
LSI14	publish	pto	apply	device	registration	title	grant	+	patent
LSI15	launch	share	self	expect	company	market	word	announce	safe
HDP1	patent	grant	title	publish	application	registration	word	release	method
HDP2	patent	title	include	word	abstract	patent	release	word	provide
HDP3	say	company	\$	word	patent	launch	sell	expect	share
HDP4	title	accord	patent	abstract	grant	release	configure	grant	receive
HDP5	say	patent	title	word	company	include	share	configure	file
HDP6	title	patent	title	say	grant	word	\$	include	file
HDP7	title	patent	include	grant	accord	word	release	receive	word
HDP8	title	patent	include	patent	say	word	company	file	accord
HDP9	title	application	receive	patent	say	word	publish	grant	file
HDP10	patent	patent	say	title	grant	word	file	include	pto
HDP11	patent	title	title	word	say	file	company	publish	application
HDP12	patent	title	grant	file	word	include	publish	accord	say
HDP13	patent	patent	title	say	grant	word	company	publish	application
HDP14	title	patent	title	grant	file	word	include	accord	say
HDP15	title	patent	grant	title	include	accord	receive	word	abstract
AP1	new	support	design	call	high_quality_fotage	new_mode	file_download	dephad	dephad
AP2	configure	network_information	communicate	video_terminal	controller	mobile_phone	include	mobile_phone_receive_network_information	internet_protocol_ip_network
AP3	portionboard	include	trade_portion	respect	fix_portion	fix_portion	contact	board_portion	drum
AP4	enterprise_mobile_print_solution	tag_printer	corporate_environment	enterprise	build	resource	map	deploy	rich_directory
AP5	boost	ecommerce_boom	camera_printer_maker	aim	revenue	country	tourism	\$	eye_heanager_senior
AP6	nm	accord	organic_dye_base_record_material	have	vicinity	vicinity	comprise	define	storage
AP7	distance	of's	general_manager	application	era	world	say_vice_president	wireless_access	instead
AP8	help	boost	state	decide	horticulure_agriculture	keenness	show	support	dairy_sector_government
AP9	supervise	of's	comprise	send	node	number	abstract	trigger_message	deemina
AP10	flat_panel_tv_segment	flat_screen	production	flat_sale	flat_revenue	flat_portion_light	flat_plate_member	flat_plate_member	flat_panel_tv_segment
AP11	\$	product	printedcircuitboard_assembly_plant	production	example	progress	call	cut	half
AP12	sense	array	inverter	sensor_value	abstract	column	provide	device	touch_panel_system
AP13	grant	recall_assembly_word	title	have	patent	flat_panel_tv_segion	flat_panel_tv_segion	flat_revenue	flat_portion_light
AP14	monitor	accord	determine	record	detect	network_monitor_device	abstract	website	include
AP15	purchase	familiar	person	steak_sauce	word	deal	accord	unite	maker

The HDP and LSI algorithms behaved similarly as in every previous case, with LSI incorporating a little less chunks in its topics, just as it incorporated less N-grams than the

other algorithms. The behavior of Affinity Propagation was very similar to that of Affinity Propagation with N-gram tokenization and BOW vectorization: it incorporated very specific clusters with relatively many specific technology chunks with little generalisability, and often clusters made of similar chunks such as AP9 in table 4.5.

Over the three samples, all algorithms behaved similarly. K-means behaved slightly better for samples two and three than for sample one in terms of number understandable clusters. All samples had a split of clusters into clusters with very specific technology terms and more general event terms, but the latter two samples had overall more understandable clusters in terms of what types news they were related to. LDA, LSI, HDP and Affinity Propagation behaved identically over all samples, of course taking into account that the sample of texts affected the topic and cluster top terms.

### 4.3.6 Chunk Tokenizer with TF-IDF Vectorization

For the chunk tokenizer combined with TF-IDF vectorization, the top terms per top 15 topics for each considered algorithm are presented in table 4.6. The results for the K-means clustering with chunk tokenization with TF-IDF vectorization are very much similar to those of plain tokenization with TF-IDF vectorization, with a difference in that chunk tokenization incorporates a few more detailed technology chunks as tokens. These clearer technology chunks bring more clarity into the clusters compared to those of plain tokenization.

The topics created in this case by the LDA algorithm are also very comparable to those of the plain tokenizer with the exception of a few chunks here and there - but fewer than for the K-means results, which seemed to prefer incorporating chunks more than other methods in this approach.

Over all the samples, the results were very similar for the algorithms. HDP and LSI behaved as in previous cases, LSI noticeably preferring unigrams over chunks compared to the other methods, while HDP insisted on making every patent news item its own topic. Affinity Propagation repeated the same few clusters over, and got hung up on certain chunks.

As a general overview, the chunk tokenizer behaved very similarly to the plain tokenizer, with the exception of including some more specific chunks in the topics and clusters, which were mostly otherwise useful except for the K-means algorithm with BOW vectorization. Compared to N-gram tokenization, the ratio of chunks to unigrams was very small in every approach.

Before moving on to the discussions and conclusions chapter, may it be noted that the LDA algorithm systematically created fewer topics for patent news compared to the other approaches, and that Affinity Propagation was very much less repetitive in the cases of bag-of-words vectorization. This was similar behavior to the smaller samples tested with

**Table 4.6.** Chunk tokenizer with TF-IDF vectorizer. Top 15 clusters and topics.

K-means1	rank	company		accord	term	globally	world	list	share	leader
K-means2	file	patent	flat_global_growth	flat_screen	flat_sale	flat_revenue	flat_portion_light	flat_plate_like_light_emit_panel	flat_plate_member	flat_panel_tv_segment
K-means3	apparatus	registration	publish	title	patent	method	device	process	include	image
K-means4	growth	say		year	revenue	expect	company	drive	market	sale
K-means5	patent_patent	plus	grant	application	title	flat_panel_television_set	flat_screen	flat_revenue	flat_portion_light	flat_plate_like_light_emit_panel
K-means6	apparatus_word	receive	application	apply	title		publish	apparatus	image	kabuki
K-means7	developer_image_form_apparatus	process_cartridge_word	grant	application	title	flat_screen_division	flat_screen	grant	flat_revenue	flat_portion_light
K-means8	device_word	apply	receive	application	title		publish	grant	patent	method
K-means9	word	file	include	publish	share		patent_application	patent	accord	year
K-means10	operation	official	say	set	semiconductor_market_company	chalk	transfer	employee	sale	blame
K-means11	abstract	release	accord	include	title	provide		method	have	inventor
K-means12	post	loss	say	operate_profit	net_profit	year		net_loss	profit	win
K-means13	grant	patent		title	device	write		thruout	process	apparatus
K-means14	power	by's	word	announce	software	operate_system	supply	processor	generate	power_supply
K-means15	\$	share	fall	million		say	sell	company	invest	drop
LD41	pto		receive	application	publish	title	kabuki	word	apparatus_word	dateline
LD42	compete		update	event	reportedly	free	tbreak	announce	company	offer
LD43	s		way	force	award	away	enough	ensure	shed	be
LD44	billion		improve	plant	maker	upgrade	decision	system_word	type	calculate
LD45	compare		base	add	let	intend	eye	user	scanner	instance
LD46	director		settle	jury	evidence	function	withdraw	report_acquisition	threat	technology_giant
LD47	's		work	despite	have	net_profit	large	tv	plunge	will
LD48	platform		cooperation	revive	title_word	allege	core_business	replacement	flood	virtually
LD49	chip		production	enter	establish	contract	test	movie	prove	exclusively
LD410	revenue		hope	acquisition	company	presence		device_manufacture_method	similar	consolidate
LD411	firm		rank	stake	firm	shipment	shift	new_device	production_facility	correct
LD412	configure		include	need	transmit	computer	select	memory	receive	signal
LD413	issue		application	publish	patent	area	merge	world_leader	interface	direction
LD414	purchase		subject	ready	profitability	market_researcher	arise	transaction	apart	operate_loss
LD415	run		new_range	wj	ad	information_process_system	length_word_dateline	detection	pen	launch
LS11	file	patent	application	grant	title	publish	registration	word	device	receive
LS12	grant		publish	file	registration	word	application	receive	patent	apply
LS13	grant		application	publish	receive	apply	registration	pto	title	patent
LS14	application		registration	receive	apply	word	patent	device	file	plus
LS15	publish		application	pto	registration	title	patent_application	grant	device	word
LS16	application		word	registration	pto	patent_application	publish	receive	say	file
LS17			application	say	company	share	receive	apply	sell	pto
LS18	plus		pto	device	publish	patent_application	apply	method	title_word	registration
LS19	ward		share	receive	device_word	pto	apply	device	plus	patent_application
LS110	device		plus	word	share	method	include	apparatus	patent_application	accord
LS111	share		own	device	directly	\$	say	word	company	method
LS112	patent_application		include	accord	registration	plus	release	abstract	publish	word
LS113	device		plus	apparatus	method	apply	registration	receive	accord	device_word
LS114	method		include	apparatus	accord	patent_application	release	abstract	say	word
LS115	\$		own	directly	company	launch	share	expect	manufacture	sale
HDP1	patent	file	grant	title	application	publish	word	receive	registration	registration
HDP2	patent	file	grant	title	application	publish	word	registration	receive	registration
HDP3	patent	file	grant	title	application	publish	word	receive	registration	registration
HDP4	patent	file	grant	title	application	publish	registration	word	receive	registration
HDP5	patent	file	grant	title	application	publish	word	receive	registration	registration
HDP6	patent	file	grant	title	application	publish	word	registration	receive	registration
HDP7	patent	file	grant	title	application	publish	word	registration	receive	registration
HDP8	patent	file	grant	title	application	publish	word	receive	registration	registration
HDP9	patent	file	grant	title	application	publish	word	say	registration	registration
HDP10	patent	file	grant	title	application	publish	word	word	receive	registration
HDP11	patent	file	grant	title	publish	application	word	registration	receive	registration
HDP12	patent	file	grant	title	application	publish	word	registration	receive	registration
HDP13	patent	file	grant	title	application	publish	word	registration	receive	registration
HDP14	patent	file	grant	title	publish	application	word	receive	registration	registration
HDP15	patent	file	grant	title	application	publish	word	registration	receive	registration
AP1	file	patent	flat_global_growth	flat_screen	flat_sale	flat_revenue	flat_portion_light	flat_plate_like_light_emit_panel	flat_plate_member	flat_panel_tv_segment
AP2	file	patent	flat_global_growth	flat_screen	flat_sale	flat_revenue	flat_portion_light	flat_plate_like_light_emit_panel	flat_plate_member	flat_panel_tv_segment
AP3	registration	publish	title	patent		flat_screen_television	flat_screen_division	flat_screen	flat_sale	flat_revenue
AP4	enterprise_mobile_print_solution	tag_printer	corporate_environment	rich_directory	map	deploy	resource	enterprise	build	
AP5	file	patent	flat_global_growth	flat_screen	flat_sale	flat_revenue	flat_portion_light	flat_plate_like_light_emit_panel	flat_plate_member	flat_panel_tv_segment
AP6	file	patent	flat_global_growth	flat_screen	flat_sale	flat_revenue	flat_portion_light	flat_plate_like_light_emit_panel	flat_plate_member	flat_panel_tv_segment
AP7	registration	publish	title	patent		flat_screen_television	flat_screen_division	flat_screen	flat_sale	flat_revenue
AP8	sense	sensor_value	touch_panel_system	multiple_sense_element	row	column	sum	array	element	inventor
AP9	patent_application	pto	publish	flat_screen	flat_sale	flat_revenue	flat_portion_light	flat_plate_like_light_emit_panel	flat_plate_member	flat_panel_tv_segment
AP10	file	patent	flat_global_growth	flat_screen	flat_sale	flat_revenue	flat_portion_light	flat_plate_like_light_emit_panel	flat_plate_member	flat_panel_tv_segment
AP11	steak_sauce	unite	person	familiar	mat	near	maker	purchase	deal	accord
AP12	supplychain_analysis	of's_revenue	account	customer	make	have	company	accord	company	
AP13	registration	publish	title	patent		flat_screen_television	flat_screen_division	flat_screen	flat_sale	flat_revenue
AP14	file	patent	flat_global_growth	flat_screen	flat_sale	flat_revenue	flat_portion_light	flat_plate_like_light_emit_panel	flat_plate_member	flat_panel_tv_segment
AP15	registration	publish	title	patent		flat_screen_television	flat_screen_division	flat_screen	flat_sale	flat_revenue

the Mean Shift algorithm, which behaved similarly to Affinity Propagation in the bag-of-words cases by singling out text documents, favouring chunks and N-grams over uni-grams, and not being very generalisable.

### 4.3.7 Content Analysis

Content analysis, as described in the methodology chapter 3.4.6, was performed for a sample of 58 134 text documents. The sample was run for K-means and LDA algorithms with the chunk tokenizer for 150 topics and clusters, since the top terms for these approaches were the most coherent and interpretable for a human reader previously. 150 clusters and topics was chosen as an appropriate amount, since it is comparable to the other results studied previously.

While doing the content analysis and content coding, the goal was to stay true to the human interpretation of the content, which is why the coding, the names and descriptions of text groups, has more general themes as well as specifications according to the interpretations of the reader. The full coding and results of the content analysis is presented

in appendix A, along with some syntax explanations.

As described in the previous chapter, the congruence of top terms in a cluster or topic to human interpretation of the contents was assessed. The scale and tagging used for this was “in congruence”, “not in congruence”, “clear business action cluster or topic in congruence”, and “clear business event cluster or topic in congruence”. Amounts of these evaluations for each method combination are presented in table 4.7.

**Table 4.7.** *Congruence of top terms to content analysis per algorithm*

	KM - CC	KM - CT	LDA - CC	LDA - CT
Clear action in congruence	37	35	7	11
Clear event in congruence	51	48	13	25
In congruence	59	77	38	42
Not in congruence	40	25	99	83

Clearly, K-means performs the best at creating representations of clusters that are in congruence with their contents. K-means with TF-IDF vectorization locates the largest amount of clusters in congruence with their contents, but K-means with BOW vectorization locates the largest number of clear business events or actions. LDA performs poorly in comparison with either case of vectorization. The LDA results in general were more difficult to interpret, vague and the found themes were broad. Furthermore, the topics were always comprised of many documents, while K-means was happy to create clusters of single or two documents.

However, while there were plenty unintelligible and term incongruent topics, the LDA top terms did provoke thoughts and produce insights about categories that should be included. For instance, LDA with TF-IDF vectorization found a topic with terms “struggle”, “job\_cuts”, “worry”, “market”, and “hope”, which is clearly indicative of a company in distress. This could indeed be included in the categorization with reason. The reader did not pick up on this latent topic in the content analysis, and this case is indicative of a situation in which LDA and latent topic models provide serendipitous insights into categorization creation. This theme persisted - many of the topics categorized as “unintelligible” by the reader could be interpreted as potentially interesting categories, see appendix A. On the other hand, sometimes when the content of a topic was clear, the top terms completely failed to reflect it.

While K-means did perform significantly better at this task, it often created clusters based on single terms, and many clusters of the same contents. For instance, there exist many patent application and patent grant clusters that could be merged. Let it be noted that all these multiples of clear clusters and topics are still counted in the results as congruent. Table 4.8 presents the amounts of content wise overlapping created clusters and topics.

Both LDA and K-means made overlapping topics and clusters. These overlapping groups often held similar terms among them - only in a different order. The exception here was LDA with TF-IDF vectorization, which very much had different term foci in the overlapping

**Table 4.8.** Amounts of overlapping topics and clusters

Topic/cluster content	KM - CC	KM - CT	LDA - CC	LDA - CT
Dispositions/acquisitions by director	2	2		
Misc "\$"	2	2		
Misc "according"	2			
Misc "launch"	2	2		
Misc "like"	2			
Misc "plan"	2			
Misc "say"	2	3		
Patent applications	14	15	3	5
Patent grants and registrations	14	13	8	4
Poor formatting	7	7	3	15
Technology descriptions	25	19	7	6
Unintelligible	3	3	29	29
Acquisitions			2	
Announcements				2
General themes			8	2
Partnerships			2	2
Metatext				2
Market standing			2	
Business targets			2	
Misc "earn"				2
Clear general themes			9	34

topics – often the others were more technology term heavy than the others. LDA with TF-IDF vectorization also stands out from the BOW vectorization in the sense that it created clearer themes and topics content wise. In table 4.8 "general themes" represent more vague themes such as "phones" for example, and "clear general themes" more relevant an actionable themes such as "defensiveness". "Misc" indicates that a cluster or topic was mostly formed around the implied term. However, as by table 4.7, while the found general themes were clear from the content analysis, they were not always interpretable from the top terms, which were often too specific, derailing, or technology heavy.

K-means could be interpreted as more vulnerable to WSD, since it often clustered together documents with synonyms, such "own shares" and "share pictures". K-means also created very many technology specific clusters, which were not studied in detail since technology details are not the focus of this thesis. They also count towards the favouring of K-means over LDA, since K-means clearly made separate clusters of technology details, while LDA mixed them into other topics, making them too specific to technology and

hard to interpret. This was the most common incongruence-causing factor – top terms of a cluster or topic were too specific to truly represent the contents of the subset of texts. This was especially the case on the located broader themes in LDA topic contents, such as the theme of warfare in LDA with BOW vectorization, in which the reader recognized important keywords such as “battle”, “rival”, and “target”, which were not reflected in the top terms. At other times the top terms were unintelligible, which was in congruence with the cluster or topic contents. These cases appeared with all approaches.

Other general level observations were that in some cases WSD was unavoidable. In all the approaches, the word “won” – for currency, was lemmatized falsely as “win”. The word “patent” did not appear in some of the top terms of patent application clusters and topics, and interpreting these topics or clusters correctly requires some familiarity with the data. However, the other top terms in them were very similar to other patent application clusters that did contain the term “patent”. The overlap of the clusters or topics could be inferred from the similarity. Still, this was a slight challenge and not straightforward. Moreover, rather than creating distinctive topics or clusters on profits, revenue, or share values, these terms were often grouped together by all methods – possibly due to semantic similarity. If these terms would benefit from separation in the categorization creation, this phenomenon should be taken into account.

The implications of the results presented in this chapter are discussed further in the following chapter. Moreover, the presented research questions are answered and the limitations and scope of this thesis are addressed further. Following this discussion, a conclusion for this thesis is provided.

## 5 DISCUSSION AND CONCLUSIONS

May it be once again stated, that the results presented in this thesis are the result of unsupervised machine learning methods, and are as such very much dependent on subjective perceptions and understanding of the topic and data. The goal of the research was that of exploratory nature, intending to venture out to a fuzzy realm of unknowns, hoping that some insights may be gained for answering the posed research questions:

- How can unsupervised machine learning methods be exploited in the creation of an action and event categorization framework for business intelligence purposes?
- How do these exploitation possibilities differ in different, common unsupervised method approaches?

### 5.1 Reliability and Validity

It is obvious from the results, as it was insinuated by the studied literary theory, that unsupervised text clustering and topic modelling is by no means an exact science. Mostly, the top term results of the studied topic modelling and clustering approaches were confusing and nonsensical for a human reader, and did not always match the content of representative clusters or topics.

Furthermore, due to the subjective nature of assessing unsupervised learning method results and content analysis, the results would be interpreted differently by different people. If one had hoped to end up with clear cut clusters and topics representing unambiguous business related events and actions, would they have been in for a lamentable disappointment.

However, this is the reality of the practice, the state-of-the-art methods, available today for categorizing text documents without a human annotator. The situation must be dealt with as it stands. Let it be noted however, that in this thesis, no method or approach was specifically optimized or tinkered with to suit the specific data, but used in their standard form for reasons of universality.

Regardless, these methods exist, and are used as they are, and are therefore relevant to understand in such way. The set premise for this thesis has been wholly pragmatic: were one to hope for some reference or framework in creating a system of categorization of events and actions for business intelligence with machine learning methods, what might

they need to consider, what may be the gains of doing so, and by what methods can they be achieved?

The premises set for the research executed in this thesis is, as described, are very subjective, and therefore this raises questions on the reliability and validity of the done work. While the subjectivity matter cannot be avoided, it may try to be mitigated. The resources for this thesis included one person to study the data, preprocess the data, interpret the results, and draw conclusions. This limits the reliability of the done work. However, this risk of subjectivity was mitigated by taking multiple random samples of the data for assessment. This aided in avoiding situations in which the researcher would have been too familiar with the exact dataset before analyzing the results of topic modelling and clustering.

In the content analysis however, only a single random samples was taken. Reliability, in this case, was considered in such a way, that the content analysis was done unaware of the topic modelling and clustering top terms, which were not studied while doing the content analysis. This was done to avoid being biased by the results.

Moreover, the clustering and topic modelling results will vary each run and for each different dataset. This makes making the exact same observations every time impossible. However, since the research problem allows for this type of a situation, the generalisability of the observations is the focus over repeatability. Generalisability was achieved by the multiple runs, and by using easily available methods that anyone could test for themselves with relative ease.

Regarding the validity of the done studies, it can be said that the measured and gained results truly do answer the posed research questions and address the research problem. The setting was exploratory, and all types of information how to utilize unsupervised learning in categorization creation counted towards answering the questions. The main threat to validity is that not all approaches of unsupervised learning could be studied in the scope of this thesis, and therefore, a totally comprehensive answer could not be gained. This is a common type of a situation that must be dealt with in research.

## **5.2 Discussions**

Let the question of "how can unsupervised machine learning methods be exploited in the creation of an action and event categorization framework for business intelligence purposes?" be considered first. It is rather clear that not much can be gained from simply using any of the studied unsupervised methods for creating the categorization by itself - most of the categories would be very difficult to understand for a human interpreter. The main gains of the computerized approach are the thoughts and insights the results evoke in a person reading the top terms of topic modelling results, and the good starting ground to build categorization on provided by clustering methods, which give a good understanding of the general contents in the relevant data.

The analyzer must - and will - consider both what is seen in the top terms representation of categorization, and what is not. Sometimes, the results will not realize some of the events a human can clearly distinguish from data content analysis, while at other times, the machine learning approach results will make the human interpreter understand details of data in a way that brings value to the resulting categorization framework in an unanticipated manner via serendipitous understandings of categorization requirements. For instance, a human may notice that a certain event type will refuse be easily defined by NLP methods, and that some extra work, such as keyword location, may be needed to ensure that certain text types will be surely located in the future.

To clarify, at other times the computerized approach representation will create a categorization that makes the human interpreter understand their need for the categorization better. For example, in the case of the data used for this thesis, the human interpreter reading the events made a simple category of "lawsuit" for certain types of news, whereas topic modelling and clustering methods sometimes distinguished between different phases of a lawsuit process, such as an allegation and a verdict. This created a moment of realization to the human interpreter that these two are indeed very different events of different interests to different stakeholders, and should be treated as such. This realization would not have happened without the aid of computational means. Moreover, some event categories were apparent in the clustering and topic modelling results that did not even cross the human interpreters mind such as a "licensing agreement" category found in the clustering and topic modelling results.

Similar moments of realization sparked from having to decipher why certain tokens were in the same term clusters. In some cases it became clear that certain events had been separated into different clusters based on technology, such as hardware development versus software development. Depending on the decision maker, these two clusters may be of equal standing, and may be combined into the same category, but sometimes it may be beneficial to keep them separate. For example, a company may be interested in a competitor's software development, but not hardware development. The knowledge that a clustering method splits these into separate events enables the possibility for an entity to decide on how they want to deal with these types of situations - should some clusters be combined, and some manually split for different intents and purposes? Machine learning may find categories such as "music" or "sports" and the top terms associated with them. The location of these categories once again allows for a natural person to decide on how they want to deal with these categories, and whether they are of interest to them as they are, or do they require some action. All in all, clustering and topic modelling can aid a decision maker to create the best categorization framework for their specific purposes.

On the topic of enabling a decision maker to be able to make informed decisions, let the question of "how do these exploitation possibilities differ in different, common unsupervised method approaches?" be addressed. As already stated in the results, some methods and approaches were more event and action focused, and some were more technology term oriented. Some method combinations were better suited to creating top

term representation that actually reflected the contents of a cluster or topic more accurately. The type of categorization framework one wishes to create dictates whether the term congruence with the contents is the first priority, which is a very reasonable approach to a more realistic data based representation. However, in another situation the coverage of all potential events and actions may be the main goal of a categorization framework. In such cases, the previously described insights from studying the top terms of various approaches with differing results may be more important than the extent to which the categorization reflects the exact contents of a certain dataset.

Moreover, running these different types of approaches may bring insight into what is truly interesting for the categorization. For instance, a decision maker may realize that clustering with the more technology focused methods, such as K-means, along with their published dates may help create a technology development historical roadmap, which may be useful information for, for example, technology forecasting. Different approaches may bring different insights, and running the different combinations allows again for a decision maker to make the observations most useful for their business intelligence.

However, while running different types of combinations for different information is suggested as useful, it may be now useful to go over what one might expect from the different, tested, approaches. As might have been expected, clustering algorithms mostly behaved in a way that clustering algorithms are meant to behave compared to topic modelling algorithms: K-means and Affinity Propagation created many more clusters on the technology term heavy side, especially with bag-of-words vectorization, since clustering algorithms are meant to create separate clear clusters, whereas topic models are meant to locate underlying latent, abstract topics. Logically, this makes sense in the way that the texts can be more clearly categorized into separate clusters based on the defining technology terminology - TV texts into the TV clusters, and camera texts into the camera clusters - while topic modelling methods LDA and LSI tended to try to find more generic terms found across the whole corpus. However, often times these topic model latent themes contained too specific technology terms unrelated to the content, which then made the topics confusing. Moreover, clustering methods made more hard clustering based on a few terms that were easy to analyze, while topic models created more vague themes and topics, which were sometimes difficult to interpret, since they contained various different events and news topics within the created topics.

If one hopes to locate more event and action based clusters or topics, plain or chunk tokenizer with the LDA algorithm and bag-of-words vectorizer, or plain or chunk tokenizer with the K-means algorithm and TF-IDF vectorization is the way to go - if looking only at the top terms. If judging also based on the content representativeness of the top terms, K-means performs significantly better than LDA. However, LDA is faster for more topics and larger datasets. For more technology based terminology in clusters and topics, the best approach according to this thesis' methodology would be plain or chunk tokenization with the K-means algorithm based on both top term representations and content representativeness, or plain or chunk tokenizer with the LDA algorithm and TF-IDF vec-

torization. However, this only works for top term based categorization creation, and is not very representative of data contents.

The LSI algorithm with any of the tokenization and vectorization combinations is an appropriate approach for anyone truly looking to get a broad overview of the scope of topics in a corpus based on the top terms per topic. It was not the best approach for locating events or categories of events or actions, but if one had a completely unknown corpus of text data, LSI would give a very fast, good glimpse into the general topics found in it.

Affinity Propagation would be the way to go to accomplish the exact opposite of LSI. It was not very good at finding underlying topics, but with the cases of the bag-of-words vectorization, it was a good approach for overviewing the text corpus in more detail, without having to read the whole texts. This was of course possible, because Affinity Propagation made thousands of clusters, and many of them represented singular text documents, while some represented more general topics and events. This provided insights to what types of themes can be found in the corpus, but also interesting singled out documents that might be of interest for some cases. At least the natural person interpreting the clusters felt that reading over three thousand lines of ten terms is faster and more informative than skimming over thousands whole documents of text.

Affinity Propagation and HDP combined with TF-IDF vectorization, were useless, creating just repetitive patent related clusters every run. This was somewhat surprising, since the studied literature gave both methods a lot of credit. On a similar note, it was interesting that the N-gram tokenization approaches were the least legible for a human reader, when the studied literature left an impression that N-grams were a good approach compared to plain tokenization. Furthermore, TF-IDF surprised by behaving worse for more algorithms, and by not bringing noticeable advantage over bag-of-words, except being slightly faster to compute for the N-gram methods. These may very well be case sensitive results, and possibly other versions and approaches to N-grams or HDP, for instance, may yield different results, because these results are, of course, not exhaustive.

There exist different types of possible combinations of approaches that will all yield differing results. Lemmatized tokens and not lemmatized tokens will give a differing outcome. Different tokenization methods, perhaps regarding casing or punctuation, combined with different vectorizers, just Scikit-learn has many more to try, will yield yet again different results. Perhaps HDP from a different source would function better than the one from Gensim.

Other different preprocessing choices affect the results: Deleting all nouns from the texts may have prevented creating clusters by specific technology terms, but at the same time would have prevented clustering a document by, for example, the term "lawsuit", which would have probably resulted in the document being miscategorized, if it did not contain further informative verbs, such as "sue".

Other clustering and topic modelling algorithms will yield different results, and simply different runs of the same algorithm will yield yet again different results. The review done

by Xu and Tian (2015) for clustering algorithms based on time complexity and suitability for high dimensional data would suggest that better algorithms for text clustering might be CURE, DBCLASD, DBSCAN, STING, or OPTICS algorithms. These were not as easily available and comparable in terms of methodology to the algorithms studied in this thesis, and were therefore omitted, but might be studied more in the future.

Regardless, the point from earlier still stands - different approaches bring different benefits into categorization creation - and it is simply a matter of time and resources, how many different types of approaches one wants to implement for as many different insights and observations as possible, if the goal of the categorization is to be as broad as possible. Different unsupervised methods cannot definitely and objectively be said to be better than any others, they are simply different. They all contribute their own difference to the richness of information, aiding in the creation of a conclusive, exhaustive, human interpretable, and data congruent framework for event and action categorization for business intelligence.

### **5.3 Conclusions**

In conclusion, the setting that raised the relevant research questions was that of often not finding any clear reasoning for event categorizations in business intelligence literature. For example, for the event categorization software SIE-OBI, made for business analysts for locating important events, no clear, exhaustive, reference framework is presented. The user must define themselves what are the interesting categories are. (Castellanos et al. 2012) The question was raised how machine learning can positively contribute to creating these categorizations, and the first research question was formed. After this, different types of machine learning approaches were studied and considered. It became clear that the choices made would affect the results and input to categorization creation. This raised the second research question.

In the beginning, it was very unclear what types of results and answers could be found for the questions, because the setting was very general and even vague. Many different approaches were considered and trialed on a dataset of mostly digital camera related news, and in the end three different tokenization methods, two different vectorization, three clustering algorithms, and three topic modelling algorithms and their combinations were tested for numerical coherence, time requirements, human interpretability of top terms, and their correspondence to the actual contents of the relevant clusters and documents. The methods were chosen based on their easy availability and understandability to someone not necessarily too familiar with machine learning or NLP beforehand.

From all the results, it was clear that while no unsupervised machine learning algorithm tested was a straightforward way to create any ready event and action categorization frameworks by themselves, they were not devoid of value. When compared to a reference categorization made by a human reader while reading over the dataset, the clus-

tering and topic modelling results were able to bring key insights into bettering the initial subjective categorization framework further, and understanding different types of possibilities to tweak it for different needs. These types of observations are of value to anyone creating the framework, whether the observations are made by themselves based on topic models and clusters, or by an outside source who has studied different types of data processed with unsupervised learning. Moreover, some of the tested approaches were good ways to create this type of baseline, improvement requiring, categorization framework without a human having to read the data.

It was noticed that different approaches created different types of valuable results regarding the categorization. For example, certain method combinations of the K-means and Latent Dirichlet Allocation algorithms were more easily humanly interpretable based on their displayed top descriptive terms than other methods, and out of these LDA was faster than others for a larger dataset. K-means top term representations, on the other hand, were more representative of the actual contents of the created clusters, when analyzed with content analysis. Some combinations of the methods created more technology term heavy categories than others. For LDA, technology terms typically made top term representation more difficult to interpret correctly regarding topic contents. K-means created more clear cut technology themed clusters, instead of mixing technology terms to other clusters' top terms.

Latent Semantic Indexing provided a good base for a very general overview of observable topics in a dataset, while Affinity Propagation combined with bag-of-words vectorization was suitable for creating a more detailed overview of the whole dataset, but neither was well suited for locating event categories. Hierarchical Dirichlet Process implemented in this thesis behaved very poorly in all cases.

Out of all the methods, LSI was always clearly the fastest, followed by LDA, HDP, Affinity Propagation, and Mean Shift, which was so slow it was eliminated from further considerations. Out of the tokenizers, the plain tokenizer was clearly the fastest, followed by the chunk tokenizer, N-gram tokenizer being the slowest. There was no great time difference between the vectorizers. All these results are visualized in table 5.1.

Compared to the different categorizations of Wu, Tsai, Hsu et al. (2003), Wei and Lee (2004), and Zhou, Chen and He (2015), mostly the results gained from the methods that behaved well were closest to the level of detail of the categorization done by Wei and Lee (2004), which follows the logic of the content analysis employed in this thesis. The results of LSI were more comparable to the categorization detail of Zhou, Chen and He (2015). For example, LDA and K-means approaches were capable of capturing topics and clusters that would fit in the cross-section of topics as general as "business" and "law". While it may depend on the intents and purposes of the user or decision maker which result is better for them, the interesting result here is that both types of results could be obtained with unsupervised methods, which did not require manual human annotation compared to their literature obtained counterparts.

**Table 5.1.** Relative time requirements and potential suitable use cases of the different clustering and topic modelling approaches - a conclusion

Vectorizer	Tokenizer	Algorithm	Suitable Uses	Time Requirements
Bag-of-words	Plain	LDA	Locating event and action based business events and more general technology topics	Low to moderate
		LSI	Overviewing the general topics found in a corpus	Very low
		HDP	Unuseful	Moderate
		K-Means	Locating clusters of detailed and related technology terms	Moderate
		Affinity Propagation	"Speed reading the corpus" - finding detailed singular news and some more general cluster topics found in a corpus	High
		Mean Shift	"Speed reading the corpus" - finding detailed singular news and some more general cluster topics found in a corpus	Very high
	N-gram	LDA	Locating various topics of specific technology bigrams	Low to moderate
		LSI	Overviewing the general topics found in a corpus	Very low
		HDP	Unuseful	Moderate
		K-Means	Locating various topics of specific technology bigrams	Moderate
		Affinity Propagation	"Speed reading the corpus" - finding detailed singular news and some more general cluster topics found in a corpus	High
		Mean Shift	"Speed reading the corpus" - finding detailed singular news and some more general cluster topics found in a corpus	Disqualifyingly high
	Chunk	LDA	Locating event and action based business events and general technology topics with more specific technology and business noun and verb chunks	Low to moderate
		LSI	Overviewing the general topics found in a corpus	Very low
		HDP	Unuseful	Moderate
		K-Means	Locating various clusters of specific technology bigrams with mainly more specific technology noun chunks	Moderate
		Affinity Propagation	"Speed reading the corpus" - finding detailed singular news and some more general cluster topics found in a corpus	High
		Mean Shift	"Speed reading the corpus" - finding detailed singular news and some more general cluster topics found in a corpus	Disqualifyingly high
TF-IDF	Plain	LDA	Locating topics of detailed and related technology terms	Low to moderate
		LSI	Overviewing the general topics found in a corpus	Very low
		HDP	Unuseful	Moderate
		K-Means	Locating event and action based business events and general technology clusters	Moderate
		Affinity Propagation	Unuseful	High
		Mean Shift	Unuseful	Very high
	N-gram	LDA	Locating various topics of specific technology bigrams	Low to moderate
		LSI	Overviewing the general topics found in a corpus	Very low
		HDP	Unuseful	Moderate
		K-Means	Locating various clusters of specific technology bigrams	Moderate
		Affinity Propagation	Unuseful	High
		Mean Shift	Unuseful	Disqualifyingly high
	Chunk	LDA	Locating topics of detailed related technology terms with more detail mainly due to noun chunks	Low to moderate
		LSI	Overviewing the general topics found in a corpus	Very low
		HDP	Unuseful	Moderate
		K-Means	Locating event and action based business events and general technology clusters with more specificity due to noun and verb chunks	Moderate
		Affinity Propagation	Unuseful	High
		Mean Shift	Unuseful	Disqualifyingly high

An u-mass coherence measure meant for assessing topic modelling results numerically was tested for all clustering and topic modelling algorithms, and in the general picture gave the best values for K-means and HDP, the second best for LSI, and typically the worst for LDA or Affinity Propagation. These results did not hold well with the human judgment of the coherence of the topics, and no conclusions were drawn from this test, except that further studies are needed.

While it may seem obvious now that different approaches yield differing result, and that applying as many as possible brings more possible insights to take into account in the categorization, may it be noted, that it was unsure in itself in the first place, whether any insights or usefulness may have been possible to gain at all, or that potentially the whole categorization could have been possible to do with the unsupervised methods in this thesis. Neither was the case. Furthermore, the results of this thesis provide some guidelines for the used method combinations on what to expect and what to use for acquiring certain types of results, instead of having to shoot at the dark with different method combinations and simply hope to get the type of result one is looking for.

The results and conclusions of this thesis hold well within the grand picture of business intelligence and its role in decision making. Recalling the literature on data quality by Wang et al. (1996) and Detlor et al. (2013) from chapter 2.2.2: Data quality is the quality of raw facts representing an entity or an event. This quality is the result of data correctness, objectivity, completeness, timeliness, relevance, understandability, interpretability, presentability, obtainability, and actionability.

The insights from event categorization with the aid of unsupervised learning approaches help provide an objective reference to a simple, human decided framework. The most critical offering of this thesis is the pointing out of a lack of standard and objective frameworks for event and action categorization, and giving some input on how to start to fix this issue. Data categorization becomes more complete when unsupervised methods enrich a natural person's understanding of different possible, both present and missing, categories. This enrichment further aids in making decisions on what types of event categories are relevant and interpretable, and should thus be included in the framework. Thus, the information gained from unsupervised methods is actionable in itself, and therefore valuable in the context of business intelligence.

## **5.4 Limitations and Future Research**

Even the presented literature and theory on unsupervised learning and natural language processing is simply a scratch on the surface of the field. The methods of unsupervised learning are a popular field of study, and more methods are developed constantly. All could not possibly be covered in a single thesis. This required limiting the research, and setting an appropriate scope for this thesis. The scope was set, with result generalisability in mind, to easily available an implementable methods.

While all the possible combinations of varying approaches to clustering and topic modelling text data cannot realistically be covered, a more realistic, potentially fruitful endeavor that might be pursued in the future, is the study of how the number of topics or clusters created affects the interpretability and content congruence of located categories.

This idea for further study on different topic and cluster numbers is reinforced by the notice that the coherence measures acquired in the results for this thesis were a very poor indicator of the behavior of the clustering or topic modelling methods. For instance, HDP received very good scores despite being very much mostly useless, while Affinity Propagation received very high coherence scores in the TF-IDF cases while looking very alike to HDP results. Moreover, there was no coherence difference between LDA and K-means between the bag-of-words and TF-IDF vectorization cases, while their top term representations appeared very much different to a human reader. For this reason, the coherence value per topic number may not be a very trustworthy indication of how many clusters or topics to create. However, may it be noted that the content analysis results were similar for both vectorizations for K-means and LDA algorithms. In the future, different types of coherence measures, such as UCI coherence might be tested to see whether they happened to be noticeably more suited to this task as a proxy of human understanding of the categories.

To better the generalisability of these observations of unsupervised methods for text categorization in the future, different types of datasets and different sizes of samples of documents from these datasets should be tested - for multiple human analyzers instead of one if possible. This may bring also further insights into the question of what value these methods might bring to a person already somewhat familiar with the dataset compared to a person, who has never seen the original data. In the best case, the whole datasets would be run for all the methods multiple times in order to confirm that the observations are in line with each other for every run. An interesting question that may be answered by doing this, is whether HDP may behave better for text sets that are do not have a disproportionate number of a certain type of data - patent news in this case.

## REFERENCES

- Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B. and Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.
- Alpar, P. and Schulz, M. (2016). Self-service business intelligence. *Business & Information Systems Engineering* 58.2, 151–155.
- Altaweel, M., Bone, C. and Abrams, J. (2019). Documents as data: A content analysis and topic modeling approach for analyzing responses to ecological disturbances. *Ecological Informatics* 51, 82–95.
- Aryal, S., Ting, K. M., Washio, T. and Haffari, G. (2019). A new simple and effective measure for bag-of-word inter-document similarity measurement. *arXiv preprint arXiv:1902.03402*.
- Atriwal, L., Nagar, P., Tayal, S. and Gupta, V. (2016). Business Intelligence Tools for Big Data. *Journal of Basic and Applied Engineering Research* 3.6, 505–509.
- Benedetti, F., Beneventano, D., Bergamaschi, S. and Simonini, G. (2019). Computing inter-document similarity with context semantic analysis. *Information Systems* 80, 136–147.
- Berger, A., Caruana, R., Cohn, D., Freitag, D. and Mittal, V. (2000). Bridging the lexical chasm: statistical approaches to answer-finding. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 192–199.
- Blei, Carin and Dunson (2010). Probabilistic Topic Models: A focus on graphical model design and applications to document and image analysis. *IEEE signal processing magazine* 27.6, 55.
- Blei, Ng and Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research* 3.Jan, 993–1022.
- Castellanos, M., Gupta, C., Wang, S., Dayal, U. and Durazo, M. (2012). A platform for situational awareness in operational BI. *Decision Support Systems* 52.4, 869–883.
- Chen, Chiang and Storey (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly* 36.4.
- Chew, C. and Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PloS one* 5.11, e14118.
- Coelho, L. P., Peng, T. and Murphy, R. F. (2010). Quantifying the distribution of probes between subcellular locations using unsupervised pattern unmixing. *Bioinformatics* 26.12, i7–i12.
- Coffey, A. and Atkinson, P. (1996). *Making sense of qualitative data: complementary research strategies*. Sage Publications, Inc.

- Colas, Finck, Buvat, Nambiar and Singh (2014). *Cracking the data conundrum: How successful companies make big data operational*.
- Cole, F. L. (1988). Content analysis: process and application. *Clinical Nurse Specialist* 2.1, 53–57.
- Cordeiro, M. and Gama, J. (2016). Online social networks event detection: a survey. *Solving Large Scale Learning Tasks. Challenges and Algorithms*. Springer, 1–41.
- Cvitaš, A. (2010). Information extraction in business intelligence systems. *The 33rd International Convention MIPRO*. IEEE, 1278–1282.
- Dave, Lawrence and Pennock (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web*. ACM, 519–528.
- Davenport, T. H. et al. (2006). Competing on analytics. *harvard business review* 84.1, 98.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science* 41.6, 391–407.
- Denny, M. J. and Spirling, A. (2018). Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Analysis* 26.2, 168–189.
- Detlor, Hupfer, Ruhi and Zhao (2013). Information quality and community municipal portal use. *Government Information Quarterly* 30.1, 23–32.
- Downe-Wamboldt, B. (1992). Content analysis: method, applications, and issues. *Health care for women international* 13.3, 313–321.
- Du, M., Pivovarova, L. and Yangarber, R. (2016). PULS: natural language processing for business intelligence. *Proceedings of the 2016 Workshop on Human Language Technology*. Go to Print Publisher, 1–8.
- Duncan, D. F. (1989). Content analysis in health education research: An introduction to purposes and methods. *Health Education* 20.7, 27–31.
- Elo, S. and Kyngäs, H. (2008). The qualitative content analysis process. *Journal of advanced nursing* 62.1, 107–115.
- Ertemel (2015). Consumer insight as competitive advantage using big data and analytics. *International Journal of Commerce and Finance* 1.1, 45–51.
- Friedl, J. E. (2002). *Mastering regular expressions*. " O'Reilly Media, Inc."
- Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.
- Ghasemaghahi, Ebrahimi and Hassanein (2018). *Data analytics competency for improving firm decision making performance*. ID: 271674. DOI: // doi . org / 10 . 1016 / j . jsis . 2017 . 10 . 001. URL: <http://www.sciencedirect.com/science/article/pii/S0963868717300768>.
- Gibson, M., Arnott, D., Jagielska, I. and Melbourne, A. (2004). Evaluating the intangible benefits of business intelligence: Review & research agenda. *Proceedings of the 2004 IFIP International Conference on Decision Support Systems (DSS2004): Decision Support in an Uncertain and Complex World*. Citeseer, 295–305.
- Grishman (1986). *Computational linguistics: an introduction*. Cambridge University Press.

- Hardeniya, N. (2016). *Natural Language Processing: Python and NLTK*. Packt Publishing. ISBN: 9781787285101. URL: <http://search.ebscohost.com/login.aspx?direct=true&AuthType=cookie,ip,uid&db=nlebk&AN=1426890&site=ehost-live&scope=site&authtype=sso&custid=s4778523>.
- Harwood, T. G. and Garry, T. (2003). An overview of content analysis. *The marketing review* 3.4, 479–498.
- Hazen, Boone, Ezell and Jones-Farmer (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics* 154, 72–80.
- Herrero, Á., Corchado, E. and Jiménez, A. (2011). Unsupervised neural models for country and political risk analysis. *Expert Systems with Applications* 38.11, 13641–13661.
- Hsieh, H.-F. and Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research* 15.9, 1277–1288.
- Hu, X. and Liu, H. (2012). Text analytics in social media. *Mining text data*. Springer, 385–414.
- Indurkha and Damerau (2010). *Handbook of natural language processing*. Vol. 2. CRC Press.
- Izenman, A. J. (2008). Modern multivariate statistical techniques. *Regression, classification and manifold learning*.
- Jourdan, Z., Rainer, R. K. and Marshall, T. E. (2008). Business intelligence: An analysis of the literature. *Information Systems Management* 25.2, 121–131.
- Kämäräinen, J. (2018). *Introduction to Pattern Recognition and Machine Learning: Introduction [PowerPoint slides]*. URL: [https://moodle2.tut.fi/pluginfile.php/548453/mod\\_resource/content/1/intro.pdf](https://moodle2.tut.fi/pluginfile.php/548453/mod_resource/content/1/intro.pdf).
- Kantardzic (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.
- Kelly (2007). *Data warehousing in action*. John Wiley & Sons.
- Lansley, G. and Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems* 58, 85–96.
- LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *nature* 521.7553, 436.
- Lienou, M., Maitre, H. and Datcu, M. (2009). Semantic annotation of satellite images using latent Dirichlet allocation. *IEEE Geoscience and Remote Sensing Letters* 7.1, 28–32.
- Lim, Chen and Buntine (2016). Twitter-network topic model: A full Bayesian treatment for social network and text modeling. *arXiv preprint arXiv:1609.06791*.
- Lincoln, Y. S. (1985). Naturalistic inquiry. *The Blackwell Encyclopedia of Sociology*.
- Litovu, L. (2019). *Research Methodology Course at Tampere University Spring 2019 Assignment 2 - Event Based Research*. [https://moodle2.tut.fi/pluginfile.php/604187/mod\\_resource/content/1/Tapahtumapohjainen\\_harkka-kev%C3%A4t2019.pdf](https://moodle2.tut.fi/pluginfile.php/604187/mod_resource/content/1/Tapahtumapohjainen_harkka-kev%C3%A4t2019.pdf). Accessed 15 April 2019.

- Liu, X., Li, Q., Nourbakhsh, A., Fang, R., Thomas, M., Anderson, K., Kociuba, R., Vedder, M., Pomerville, S., Wudali, R. et al. (2016). Reuters tracer: A large scale system of detecting & verifying real-time news events from twitter. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 207–216.
- Liu, Liu, Tsykin, Goodall, Green, Zhu, Kim and Li (2010). Identifying functional miRNA–mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics* 26.24, 3105–3111.
- Manning (1997). Authenticity in constructivist inquiry: Methodological considerations without prescription. *Qualitative inquiry* 3.1, 93–115.
- Manning, C., Raghavan, P. and Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering* 16.1, 100–103.
- Maulik, U. and Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.12, 1650–1654.
- Mimno, D., Wallach, H. M., Talley, E., Leenders, M. and McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 262–272.
- Mobasher, Jain, Han and Srivastava (1996). *Web mining: Pattern discovery from world wide web transactions*. Tech. rep. Technical Report TR96-050, Department of Computer Science, University of . . .
- Mohr, J. W. and Bogdanov, P. (2013). *Introduction—Topic models: What they are and why they matter*.
- Montani, I. and Honnibal, M. (2018). Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence* to appear. eprint: toappear.
- Morgan (1993). Qualitative content analysis: a guide to paths not taken. *Qualitative health research* 3.1, 112–121.
- Morgan and Hunt (1999). Relationship-based competitive advantage: the role of relationship marketing in marketing strategy. *Journal of Business Research* 46.3, 281–290.
- Mulunda, C. K., Wagacha, P. W. and Muchemi, L. (2019). Review of Trends in Topic Modeling Techniques, Tools, Inference Algorithms and Applications. *2018 5th International Conference on Soft Computing & Machine Intelligence (ISCFMI)*. IEEE, 28–37.
- Nikolenko, S. I., Koltcov, S. and Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science* 43.1, 88–102.
- Orriols-Puig, A., Martínez-López, F. J., Casillas, J. and Lee, N. (2013). Unsupervised KDD to creatively support managers' decision making with fuzzy association rules: A distribution channel application. *Industrial Marketing Management* 42.4, 532–543.
- Patton, M. Q. (1990). *Qualitative evaluation and research methods*. SAGE Publications, inc.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau,

- D., Brucher, M., Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.
- Pourvali, M., Orlando, S. and Omidvarborna, H. (2019). Topic Models and Fusion Methods: a Union to Improve Text Clustering and Cluster Labeling. *INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE* 5.4, 28–34.
- Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning*. Vol. 242, 133–142.
- Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. English. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. <http://is.muni.cz/publication/884893/en>. Valletta, Malta: ELRA, 45–50.
- Resnik, Garron and Resnik (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1348–1353.
- Röder, M., Both, A. and Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the eighth ACM international conference on Web search and data mining*. ACM, 399–408.
- Russom (2011). Big data analytics. *TDWI best practices report, fourth quarter* 19.4, 1–34.
- Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management* 24.5, 513–523.
- Salton, G., Wong, A. and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM* 18.11, 613–620.
- Sapir (2004). *Language: An introduction to the study of speech*. Courier Corporation.
- Särkiö, I. et al. (2019). Topic modelling of Finnish Internet discussion forums as a tool for trend identification and marketing applications.
- Sassi, D. B., Frini, A., Abdessalem, W. B. and Kraiem, N. (2015). Competitive intelligence: History, importance, objectives, process and issues. *2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 486–491.
- Saunders, M., Lewis, P. and Thornhill, A. (2009). Research methods for business students. 2007. *England: Pearson Education Limited*.
- Scott (2017). *Social network analysis*. Sage.
- SKLEARN (n.d.[a]). *Scikit-learn documentation - Bag-of-words representation*. [https://scikit-learn.org/stable/modules/feature\\_extraction.html#the-bag-of-words-representation](https://scikit-learn.org/stable/modules/feature_extraction.html#the-bag-of-words-representation). Accessed: 2019-02-18.
- (n.d.[b]). *Scikit-learn documentation - Clustering*. <https://scikit-learn.org/stable/modules/clustering.html>. Accessed: 2019-02-27.
- (n.d.[c]). *Scikit-learn documentation - Precision-Recall*. [https://scikit-learn.org/stable/auto\\_examples/model\\_selection/plot\\_precision\\_recall.html](https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html). Accessed: 2019-02-21.

- SKLEARN (n.d.[d]). *Scikit-learn documentation - TF-IDF representation*. [https://scikit-learn.org/stable/modules/feature\\_extraction.html#tfidf-term-weighting](https://scikit-learn.org/stable/modules/feature_extraction.html#tfidf-term-weighting). Accessed: 2019-02-18.
- Stavrianou, A., Andritsos, P. and Nicoloyannis, N. (2007). Overview and semantic issues of text mining. *ACM Sigmod Record* 36.3, 23–34.
- Stevens, K., Kegelmeyer, P., Andrzejewski, D. and Buttler, D. (2012). Exploring topic coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 952–961.
- Tesch, R. (2013). *Qualitative research: Analysis types and software*. Routledge.
- Trieu (2017). Getting value from Business Intelligence systems: A review and research agenda. *Decision Support Systems* 93, 111–124.
- Tsujii (2011). Computational linguistics and natural language processing. *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 52–67.
- Vedder, Vanecek, M. T., Guynes, C. S. and Cappel, J. J. (1999). CEO and CIO perspectives on competitive intelligence. *Communications of the ACM* 42.8, 108–116.
- Wang and Strong (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems* 12.4, 5–33.
- Wani, M. A. and Jabin, S. (2018). Big Data: Issues, Challenges, and Techniques in Business Intelligence. *Big Data Analytics*. Springer, 613–628.
- Watson and Haley (1998). Managerial considerations: organizational resistance on many fronts can detail the most promising systems, even those designed to address a specific organizational pain. *Communications of the ACM* 41.9, 32–38.
- Watson, Wixom, B. H., Hoffer, J. A., Anderson-Lehman, R. and Reynolds, A. M. (2006). Real-time business intelligence: Best practices at Continental Airlines. *Information Systems Management* 23.1, 7.
- Wei, C.-P. and Lee, Y.-H. (2004). Event detection from online news documents for supporting environmental scanning. *Decision Support Systems* 36.4, 385–401.
- Wu, S.-H., Tsai, T.-H., Hsu, W.-L. et al. (2003). Domain Event Extraction and Representation with Domain Ontology. *IJWeb*. Citeseer, 33–38.
- Xu, Harzallah and Guillet (2019). Comparing of Term Clustering Frameworks for Modular Ontology Learning. *arXiv preprint arXiv:1901.09037*.
- Xu and Tian (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science* 2.2, 165–193.
- Zhang, Jin and Zhou (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics* 1.1-4, 43–52.
- Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y. and Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC bioinformatics*. Vol. 16. 13. BioMed Central, S8.

- Zheng, Z., Fader, P. and Padmanabhan, B. (2012). From business intelligence to competitive intelligence: Inferring competitive measures using augmented site-centric data. *Information Systems Research* 23.3-part-1, 698–720.
- Zhou, Chen and He (2015). An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Zhou, Cheng and Zhang (2019). An end-to-end Neural Network Framework for Text Clustering. *arXiv preprint arXiv:1903.09424*.
- Ziegler, C.-N. (2012). *Mining for strategic competitive intelligence*. Springer.
- Zobel, J. and Moffat, A. (1998). Exploring the similarity space. *Acm Sigir Forum*. Vol. 32. 1. ACM, 18–34.

## **A TOP TERM AND CONTENT ANALYSIS CONGRUENCE EVALUATIONS**

This appendix presents the tables for the done content analysis for whether the top terms provided by the studied methods are in congruence to the content analysis derived category for each topic or cluster. The scale used in the tables is as follows: "++" stands for a clear action in congruence, "?+" stand for a clear event in congruence, "+" stands for situations the terms are in congruence with the contents, but may be too vague to be of use, and "-" stands for situations in which the top terms are not in congruence to the contents of the cluster or topic.

Some remarks on the notation: "Misc" followed by a word indicates a cluster or topic that often was focused around the word in question, and often included documents that had been clustered into the cluster or topic due to WSD. Metatext refers to general extra notation in the news that was not important to the contents. General themes were vague but clearly within a theme. Clear general themes were less vague in comparison. Unintelligible refers to contents that did not seem to make sense to the reader when put together. For example, if the top terms would imply a topic of cluster that was coherence, there would exist and incongruence.

**Table A.1.** K-means content analysis and clustering top terms congruence with bag-of-words vectorization

Analysis on business situation (1)	-	directly	produce	like	increase	competitive	incorporate	supply	way	nd ii	import
Dispositions/acquisitions by director	-	beneficial interest	hold	disposition	change	form	note	file	director word	word	director
Dispositions/acquisitions by director	-	change	beneficial interest	hold	form	note	file	director word	report acquisition	acquisition	director
Misc "based on" - technology - large very misc cluster	-	base		include	file	determine	obtain	technology	associate	set	receive
Misc "come"	-	come		company	launch	announce	user	word	device	phone	include
Misc "company"	-	company	say	word	launch	announce	expect	offer	help	share	have
Misc "image" - lots of technology descriptions	-	image	grant	title	patent	method	obtain	accord	device method	manufacture	abstract
Misc "include" - technology descriptions - lawsuits	-	include		company	provide	file	offer	generate	device	display	allow
Misc "launch"	-	launch	announce	word		say	company	come	available	plan	phone
Misc "launch" - products - campaigns - facilities	-	launch	word		recently	smartphone	offer	device	company	smartphones	price
Misc "like"	-	like		company	look	word	user	feature	compete	build	time
Misc "like" - mostly comparing times and technology	-	like		go	need	increase	s	company	help	come	government
Misc "new" - very much product launches	-	new		have	say	word	announce	offer	drive	launch	mobility
Misc "replace"	-	replace	dn	printer3310	nos	5s	control application	remove	chip fill toner powder word	significant item	refill
Misc "say"	-	say		company	word	market	launch	share	offer	include	help
Misc "say" - officials and analyst statements loads	-	say	word	analyst	help	customer	include	launch	offer	firm	provide
Misc "tell" - product details - high scale strategy	-	tell		company	analyst	have	word	reporter	customer	phone	report
Misc "work" - "in" - "on" - "together" - "towards"	-	work		say	company	word	help	closely	user	device	use
Poor format	-	£		euro	price		worth	start	confirm	tv	cut
Poor format	-	£	unlimited	month	plus	ay	mb	day	o	cost	
Poor format - "s"	-	's		have	company	say	come	include	sale	share	market

Table A.1 continued from previous page

Poor formatting	-			thin	laptops	bk	foc	lens	ep dir nikn d500 crlls frza mtrsptr actin	requirement	mst innva- tive product amsung t
Product descriptions - shares - scandals - a very big cluster	-	word	offer	user	provide	use	file	smartphones	introduce	help	set
R&D announcement (1)	-		focus	have	company	go	desktop laptop computer	vice presi- dent	hope	entirely	little
Responsibility description (1)	-		involve	actively	responsible	driving's leadership	product	service	business	prior	flagship smart- phone lineup
Sales and profits in "won" currency	-	win	trillion	billion		operate profit	say	sale	word	post	company
Services descriptions (1)	-	power coolin- goptions security measure	blade server customer	bladed en- vironment	process improve- ment	expertise	insight	announce	offer	word	service
Technology description (1)	-	identifier	abstract	send	nonauthentic	rairie	memory	classify	determine	consumable product match	consumable product
Technology description (1)	-	need	+	focus	work	include	sell	printer	pay	outside	professional
Technology description (1)	-	correspond	include	form	dispose	display re- gion	improve	opening	accord	arrange	input
Technology description (1)	-	associate	acquire	indicate	satisfy	install	communication device identi- fication informa- tion	correspond	determine	outputcontrol	proposal condition
Technology description with "on to titled"	-	abstract	accord	release	title	include	provide	method		have	comprise
Technology descriptions	-	determine	base	include	abstract	accord	release	detect	title	method	receive
Technology descriptions with "+"	-	+	abstract	title	accord	publish	website	number	release	include	provide
Technology descriptions with numbers and serials	-	+		accord	entpr	title	number	abstract	publish	website	wo201323726
Technology descriptions with numbers and serials	-	+	include	abstract	title	website	publish	accord	number	configure	

Table A.1 continued from previous page

Technology descriptions with poor format	-	+		accord	abstract	title	publish	website	number	provide	method
Unintelligible	-		atop	prove	situation	summit	company	stand	forge	race	ultimately
Unintelligible	-	have		say	word	company	world	's	like	not	launch
Unintelligible - large cluster	-		company	word	announce	market	not	take	user	add	buy
Misc "according" - mostly market analysis	?+	accord	expect	percent		grow	research firm	analyst	reach	unit	share
Misc "cost" - costs - "at the cost of"	?+	cost	company		say	reduce	cut	help	word	save	estimate
Misc "expect" - many financial results - scandal responses - nominations	?+	expect		say	word	company	analyst	grow	market	launch	revenue
Misc "hold" - shares - seminars - equities	?+	hold		word	company	share	stake	inform	recently	announce	meet
Misc "loss" - mostly financial	?+	loss	company	post	word	expect	warn	job	announce	\$	compare
Misc "percent" - revenues - sales - shares - changes	?+	percent	say	share	fall	sale	rise	revenue	company	account	increase
Misc "plan" - general future plans	?+	plan		word	announce	launch		job	invest	say	build
Misc "plan" - product/project releases - job cuts	?+	plan	company		say	announce	sell	launch	include	unit	year
Misc "report" - earnings and profits mostly	?+	report	word	company		accord	profit	sale	fall	share	revenue
Misc "rise" - shares - incomes - shipments	?+	rise	share	say	\$	cent	revenue	trade	company	profit	fall
Misc "sell"	?+	sell	company	unit	smartphones	phone	word	launch	accord	\$	handset
Misc "sell"	?+	sell	say		company	unit	word	expect	market	plan	start
Misc "sell" - sold products - business units/operations	?+	sell		company	word	announce	share	market	unit	phone	have
Misc "share" - prices - "shared among" - "share pictures"	?+	share	own	directly	stock	fall	involve	action	cent	close	\$
Filed patent (1)	+	address		erasable	programmable	grant	read	title	patent	file	memory word
General "consumer electronics"	+	phone	flip	cover	protective	mobile	wearable	digital	portable	memory	computer
Misc "according" - shares - employment - inheritance	+	accord		company	datum	research firm	share	word	market share	market	unit
Misc "announce" - job cuts - software	+	announce	word		available	company		issue	offer	recently	design

Table A.1 continued from previous page

Misc "forecast" - profits - revenue - market analysis	+	forecast	analyst		sale	company	year	revenue	earnings	word	share
Misc "issue" - patents - metatext	+	issue	application	patent	publish	obtain	notification	inventor	include	infringe	subscription
Misc "market" - marketing and markets	+	market	say	company	launch	word	smartphones	lead	share	hit	product
Misc "product"	+	product		company	include	sell	say	launch	sale	use	market
Misc "publish" - patents + metatext from news	+	publish	patent application	pto	word	develope	title	method	website	manage	apparatus
Misc "release" - products - information - reports	+	release		company	word	include	device	accord	expect	plan	user
Misc "sales"	+	sale		word	company	include	expect	fall	increase	cent	smartphones
Misc "sales" - changes - reports - consequences	+	say	sale		company	word	increase	expect	market	analyst	fall
Misc "year"	+	year	end	expect	company	ago	period	sale	win	cent	early
Poor format	+	>	<	lose	author></author	congress	choose	visit	/categories	< category	linear
Poor format (1)	+		3	1	0		11	iness	n	ion	
Poor format (1)	+	\$		8hudson	8stephan \$ 8sea- mus power	0dominic	8	0nate	roberto	ollie	877blayne
Technology description (1)	+	measure	meter	control	pressure	sensor	process	magnetic	prism	projection	liquid
Technology description (1)	+	test	base	form	master im- age datum	read	image test method	print im- age	optically	depend	transparent image da- tum
Technology description (1)	+	configure	absorb	provide	fix	include	joint	form	light guide section	ferrule	actuator
Technology description (1)	+	heat	cook	facial	cooker	gas	bathtub	dispenser	shower	supply	stove
Technology description (1)	+	take	capture	look	catch		lonely house	near	game reserve	eyecatching picture	tree
Technology description (1)	+	display	base	receive	successively		control unit	switch	audio determi- nation unit	store	title
Technology description (1)	+	substrate	+	ia in	flexible substrate electrical connector pad	abstract	arrange	connection	accord	connect	method

Table A.1 continued from previous page

Technology description (1)	+		flexibility	standardsbased adsl product	central office	feature	offer	multiline architecture	's	comprehensive adsl solution	development
Technology description (1)	+	assume	presidency pcs	sale subsidiary	in's	plant	camcorder	company headquarter	ese firm portable video taperecorder	household use	formula player
Technology description (1)	+	online	andor	relate	portable	host	include	provide	global computer network	22aug	webbased application
Technology description (1)	+	include	receive	store	transmit	accord	match	abstract	light system	light device	radio remote controller
Technology description (1)	+	lead	optically	comprise	couple	optical	have	convert	provide	modify lead light	wavelength
Technology description (1)	+	carry	transmit	include	able	control	apparatus	network node control	provide	initiate	accord
Technology description (1)	+	measurement	component	blood	measure	apply	electrode	base	time	device	biological information measurement mode
Technology description (1)	+	control		irradiate	set	masaki	+		present invention	farred light	plant grow system
Technology description (1)	+	transmit	available	determine	use	+	example	embodiment	accordance	accord	website
Technology description (1)	+	include	execute	interlock	provide	accordance	relate	audio	service	realize	dispatch
Technology descriptions	+	connect	include	abstract	release	accord	have	electrically	title	provide	configure
Technology descriptions	+	connect	include	abstract	generate	file	allow	cause	release	potential	set
Technology descriptions	+	receive	transmit	include	accord	abstract	release	correspond	store	title	send
Technology descriptions	+	provide	accord	abstract	release	include	title		base	store	form
Technology descriptions	+	include	abstract	release	accord	title	provide	have	form	plurality	receive
Technology descriptions	+	display	include	accord	abstract	release	execute	control	title	set	content
Technology descriptions	+	have	include	abstract	release	accord	title	positive refractive power	expose	order	object
Technology descriptions	+	configure	include	abstract	accord	release	title	base	supply	provide	generate
Technology descriptions	+	relate	accord	abstract	release	title	base		embodiment	disclose	memory
Technology descriptions	+	print	accord	abstract	method	head	provide	reverse	control	title	include

Table A.1 continued from previous page

Technology descriptions	+	generate	include	accord	abstract	base	release	pixel	control	image	correspond
Technology descriptions	+	set	configure	include	abstract	title	unique	shift	video	accord	release
Technology descriptions	+	process	abstract	release	title	accord	include	provide	configure	acquire	generate
Technology descriptions	+	form	include	abstract	accord	release	title	have	connect		plurality
Technology descriptions	+	connect	include	configure	node	datum line	output	power supply voltage	sense control line	source	apparatus
Technology descriptions	+	drive	configure	accord	generate	abstract	release	transfer	comprise	receive	include
Technology descriptions	+	record	generate	reproduce	read	determine	lens	title	include	device	accord
Technology descriptions	+	enable	set	incorrect		repair	base	logic circuit	provide	initialize replacement value	register
Technology descriptions	+	record	medium	disklike	configure	include	convey	therebetween	outside	position	disk cartridge
Technology descriptions	+	configure	generate	sense	include	determine	similar	medium	accord	prepare	perform
Technology descriptions (2)	+	couple	extend	adapt	include	have	scan line	arrange	pixel	release	datum line
Technology descriptions (2)	+	describe	maintain	raise	develop	emerge	comply	relevant market	key executive manager	claim	liability
Technology descriptions (2)	+	acquire	configure	control	store	av	immediately	compare	encapsulate	abstract	+
Technology descriptions with "receive"	+	receive	title	pto application	word	include	method	configure	file	transmit	response
Technology descriptions with numbers and serials (1)	+	image	read	emit	generate	control	+	illumination light	worth	necessary	arbitrarily
Unveilings - mostly product launches at events	+	unveil	word		launch	company	device	product	expect	phone	smartphones
Donation (1)	++	\$	million	donate	etch photo	follow news release	action	miss	wound	continue upkeep	scene
Filed patents, trademarks, lawsuits...	++	file	patent	application	include	infringe		lawsuit	form	dispose	allege
Misc "\$" - funding - prices - profits	++	\$	word	revenue	worth	billion	fall	pay		company	say
Misc "\$" - shares - prices - general numbers	++	\$	share	close	rise		cent	report	say	drop	sell
Misc "cut" - costs - jobs - forecasts	++	cut	job	word		say	cost	company	plan	announce	slash
Misc "deal" - "deals with" - business details	++	deal		company	word	announce	close	buy	sign	expect	worth

Table A.1 continued from previous page

Misc "sue" - many patent infringement lawsuits	++	sue	word		company	claim	patent	allege	infringe	stop	file
Patent application	++	receive	title	device word	application	apply	plus		device	method	image
Patent application	++	receive	publish	word	pto	title	application	dateline	kabhiki	apparat	method
Patent application on "communication"	++	application	communication device	apply	title	receive	communication method word	plus	transmission channel word	beamforming	uned channel word
Patent application on "image"	++	kabhiki	apply	application	receive	title	apparat word	apparat	information process	word date-line	method
Patent applications	++	file	word	publish	patent application	device	method	have	provide	application	program
Patent applications	++	receive	device	title	application	apply	word	method	include	plus	thereof
Patent applications	++	receive	publish	pto	application	title	kabhiki	apparat word	device word	apparat	program word
Patent applications	++	receive	application	apply	title	plus	apparat word	apparat	method	title word	word date-line
Patent applications	++	receive	application	word	title	apply	plus	apparat	method		thereof
Patent applications	++	file	method	publish	word	patent application	apparatus	patent	provide	application	system
Patent applications	++	application	publish	word date-line	receive	title	pto	kabhiki	apparat	device	method
Patent applications	++	receive	apparat	title	application	method	pto	publish	word	word date-line	kabhiki
Patent applications - international	++	file	patent application	internationally	pto	company	title	flagship industrial group	flagship insurance partner	flagship image capability	flagship have camera
Patent applications + inventors	++	publish	application	patent	invent	issue	journal	anes	ohtsuka	flagship electronic business	flagship console
Patent grants	++	grant	title	patent	method	apparatus	control	word	plus	kabhiki	communication
Patent grants	++	grant	process	title	patent	program word	plus	word	device	method	storage
Patent grants	++	grant	plus	patent	title	word	apparatus	kabhiki	method	device	device word
Patent grants	++	grant	read	title	patent	apparatus	kabhiki	image	have	method	plus
Patent grants	++	grant	patent	title	method	apparatus	kabhiki	device word	apparatus word	device	system
Patent grants	++	grant	thereof	patent	title	word	method	control method	apparatus	plus	device

Table A.1 continued from previous page

Patent grants	++	grant	title	patent	word	include	device	display	plus	image	image form apparatus
Patent grants	++	grant	have	title	patent	word	device	apparatus	plus	kabhiki	display
Patent grants	++	grant	method	title	patent	word	manufacture	device	fabricate	plus	drive
Patent grants + trademarks + contracts	++	grant	patent	title	plus	patent patent	apparatus	device word	trademark	method	apparatus word
Patent published	++	patent	publish	title	word	plus		title word	sony	olympus	program word
Patent published	++	publish patent	word	title word	title	sony		long evolution	plus	system	title device host device
Patent published	++	registration	patent	title	publish	method	apparatus	system	include	medium	have
Patent registrations	++	title	apparatus	publish	patent	registration	method	image	kabhiki	process	device
Patent registrations	++	device	title	patent	registration	publish	method	include	have	thereof	system
Trademark grants + few patents	++	grant	word	title	patent	trademark	trade mark	apparatus		tramarck	tra

**Table A.2.** K-means content analysis and clustering top terms congruence with TF-IDF vectorization

Advertisement (1)	-	use	s	mobile		fit	chance	need	erratic electricity supply	web browser	phone
Building a nuclear reactor (1)	-		say	depend	build pwr	senior an- alyst	overseas project	able	primarily	begin	curb
Exploding phones	-	catch	switch		urge	halt	handset	follow	report	owner	issue
Licensing agreements	-	source	image	technology	talk	license agree- ment lodde code	rp	w	transmit	mpany bstracts	report
Misc "account" - sales, user accounts - "hack"	-	account		percent	accord	sale	say	revenue	sell	company	smartphones
Misc "buy" - customers, companies, outsourcing	-	buy		say	word	share	company	\$	customer	announce	have
Misc "control" - technology and business	-	control	accord	abstract	release	include	title	method	apparatus	comprise	receive
Misc "hold" - competitions, events, shares, equities	-	hold		company	word	stake	accord	announce	market	patent	make
Misc "include" - job cuts - tech - large cluster	-	include	company	provide	device	announce	have	accord	file	word	smartphones
Misc "print" - printer technology	-	print	accord	abstract	release	scan	control	comprise	include	correspond	determine
Misc "provide" - technology - mergers	-	provide	accord	abstract	release	include	title	base		feature	connect
Misc "report" - strategy, financial results, comparing business	-	report	word		company	say	accord	announce	profit	offer	recall
Misc "say" - plans and evaluations	-	say		company	word	analyst	have	sell	provide	vote	report
Misc "stand" - "opposite standings"	-	stand		say	compare	phone	man	come	burn oil platform	brand	company"poured oil
Misc "win" - also "won" currency	-	win	say	operate profit	trillion	billion	company	post	sale	percent	word
Misc "work" - Unintelligible	-	word		announce		company	say	issue	have	plan	set

Table A.2 continued from previous page

Misc "year" - targets - reports - trends - comparisons	-	year	company		end	period	expect	ago	quarter	compare	early
Patent applications	-	device	receive	application	title	apply	word	method	plus	include	thereof
Poor formatting	-	+	abstract	title	accord	publish	website	number		provide	include
Profits - sales - revenues in won currency	-	win		billion	trillion	project	total investment budget	figure	operate profit	reinstate	chairman
Technology description (1)	-	enable	increase	function	compare	ensure	equivalent	effective	adopt	reduce	quick
Technology descriptions	-	drive	abstract	include	accord	release	title	output	display panel	method	control
Unintelligible	-	announce	's	come	user	accord	offer	provide	use	like	smartphones
Unintelligible	-		word	reserve		title	vendor	summary	datum	abstract	accord
Unintelligible - sites and facilities	-		have	announce	come	launch	use	include	like	user	make
Misc "\$" - stock, revenues, price	?+	\$	share	company		cent	say	worth	sell	revenue	billion
Misc "\$" - stock, revenues, price	?+	\$	share		close	say	fall	cent	rise	company	revenue
Misc "available" - results - products - times	?+	available		include	launch	device	price	start	announce	offer	user
Misc "close" - facilities - many stock "closed at"	?+	close	share	fall	\$	cent	's	trade	rise	low	
Misc "fall" - stocks, profits, also "autumn"	?+	fall	say	sale	percent	cent	profit		report	's	revenue
Misc "introduce" - products - plans - schemes	?+	introduce		word	company	plan	have	launch	market	user	include
Misc "involve" - shares and corporate scandals	?+	involve	share	stock	action	disposition	acquisition	merger	s	spy	employee
Misc "price" - price cuts and price statements	?+	\$	cut	price	cost	reduce	rival	console	instant camera	follow	twoyear agreement spokesman
Misc "sale" - units, money, forecasts, go on sale	?+	sale		company	word	rise	report	increase	expect	profit	product

Table A.2 continued from previous page

Misc "sell" - products and business	?+	unit	sell	accord	grow		company	ship	say	launch	worldwide
Misc "share" - changes, share pictures"	?+	share	company		say	cent	fall	rise	's	accord	percent
Misc "stock" - price changes mostly	?+	stock	share	fall	company	rise	's	value	trade	gain	investor
Share ownings	?+	share	own	directly	indirectly		dion	company	e	say	option
Metatext	+	publish	title	abstract	website	accord	include	comprise	word	determine	
Misc "add"	+	add		company	device	customer	support	user	provide	expand	increase
Misc "company" - tech focus	+	company		say	sell	plan	word	statement	have	add	include
Misc "company" - vague strategy	+	company	say		announce	have	plan	offer	like	come	's
Misc "expect" - business moves, products - analysts	+	expect		company	say	analyst	launch	accord	increase	grow	word
Misc "filed" - patents, lawsuits, bankruptcy	+	file	patent	lawsuit	infringe	suit	include		seek	provide	accord
Misc "grow" - business growth	+	company	business	grow		say	have	market	cent	expand	's
Misc "help" - possibilities and opportunities - business and for customers		help		say	customer	word	user	market	design	increase	business
Misc "market" - marketing and market as noun	+	market		company	like	come	announce	product	sell	have	hit
Misc "pay" - businesses or customers	+	pay		company	\$	order	word	agree	damage	say	settle
Misc "phone" - unveilings and features - sold units	+	phone		sell	user	use	have	launch	run	's	feature
Misc "say" - management and analyst statements	+		say	word	expect	sell	plan	market	have	share	sale
Misc "say" - management and analyst statements	+	say	analyst	market	include	statement	's	deal	plan	sale	come
Misc "sell" - products and business	+	sell		company	word	say	product	plan	smartphones	accord	handset

Table A.2 continued from previous page

Patent infringement lawsuit proceedings (1)	+	open statement	violate"s patent	use	smartphones	lawyer	feature	infringe	sell	tablet	tell juror
Poor formatting	+	>	<	< format = linear	nearly	see	visit	sale	provision	results</	fall
Poor formatting	+	>	/td></tr	8</td><td	date	long	r	< table	android</td><td><td><td>	> < >	charge
Poor formatting	+	5		2	9	88	4		55	<	hold
Poor formatting	+	cable	speed	world	bring	world http- nerksnoki- acom- newsev- entspress- room- press- releas- esnokia- belllabs achievesworlds- gbpsym- metrical- dataspeedover- tradition- alca technical back- ground informa- tion	+	test	deliver	be	proof+of+concept+by+%40nokianetw
Poor formatting	+	plus	grant	patent patent		title	£	reserve	cost	trademark	grant trade mark word
Poor formatting (1)	+	atomic%	q	+	2261281[jp	cell wall phase	r	release	ep2979279	accord	rpfeqrucusco100pqrs
Technology description	+	form	include	abstract	release	accord	title	plurality		supply	have
Technology description (1)	+	configure	delete	processor information process method	reproduction information	release	program inventor	accord	update	delete portion	share

Table A.2 continued from previous page

Technology description (1)	+	spectrally	emit	dispose	include	diffract	block	visible light wavelength band	optical axis	white light observation	subject
Technology description (1)	+	apparatus	digital	projector	set	telephone	perform	allow	generate	ecommerce	contain
Technology description (1)	+	datum	store	receive	process	include	identify	refer	mail datum	correspondence	apparatus
Technology description (1)	+	configure	include	sample		optical path	analog signal	release	tomographic image obtain unit	converter	accord
Technology description (1)	+	supply	configure	control	utilize	case	20140303accord	necessary	stop	detect	open state
Technology description (1)	+	cool	liquid	conduct	calculate	flow	detect	refrigerant circuit	heat	evaporator	engine cool unit
Technology description (1)	+	manage	authenticate	external	acquire	perform		information process apparatus authentication control method	registration	release	external computer
Technology description (1)	+		high performance from"s processor	development	slim design	feature	business	say	manager	interest choice	new pcs
Technology description (1)	+	form	store	define	know	heat	use	release	accord	process apparatus	positional relationship
Technology description (1)	+	bear	drive	nip	transfer	sheet detector	release	image form apparatus	abstract	detect	transfer portion
Technology description (1)	+	image	pass	include	enter	because	p2	cause	release	optical property	allow
Technology description (1)	+	receive	associate	disable	sign	prevent execution	apparatus	private key	proceed	public key infrastructure	secure memory area

Table A.2 continued from previous page

Technology description (1)	+	include	substantially	turn		plurality	circular polarization	circular polarization pattern	glass	release	right lens
Technology description (1)	+	output	specify	request	display	obtain	photograph	base	computer-readable record medium	data transitory	release
Technology description (1)	+	obtain	include	base	find	negotiation	use	search	and/or	configure	title
Technology description (1)	+	provide	application	monitor	research	process	migration	peripheral	troubleshoot	warehouse	feature
Technology description (2)	+	read	transmit	configure	store	abstract	receive	include	+	supply	connect
Technology description (2)	+	pixel	accord	belong	generate	drive	extend	integrate	acquire	release	arrange
Technology description (2)	+	add	configure	comprise	reporter word	new capability	personal audio	image datum	pixel region information	pixel portion	apparatus
Technology descriptions	+	device	title	patent	registration	publish	method	include	have	image	apparatus
Technology descriptions	+	include	abstract	release	accord	dispose	title	configure	provide	have	correspond
Technology descriptions	+	configure	generate	correspond	base	include	output	release	transmit	abstract	accord
Technology descriptions	+	execute	include	image	obtain	release	abstract	accord	base	determine	plurality
Technology descriptions	+	mean	configure	title	detect	abstract	release	receive	accord	comprise	reference
Technology descriptions	+	generate	accord	include	abstract	base	release	title	perform	determine	comprise
Technology descriptions	+	transmit	include	accord	abstract	release	title	perpendicular	direction	reflect	obliquely
Technology descriptions	+	monitor	cloud	provide	comprise	create	analyze	store	drive	array	sound
Technology descriptions	+	provide	access	text	search	create	address	enable	process	store	make
Technology descriptions	+	apparatus	printer	display	digital	camera	light	memory	plasma	consist	terminal

Table A.2 continued from previous page

Technology descriptions	+	form	reflect	transmit	comprise	have	abstract	cathode	include	consist	therethrough
Technology descriptions	+	substrate	dispose	seal	include	spacer	provide	thereof	display	extend	display panel
Technology descriptions	+	save	configure	say	require	update index information	association	release	manage	authenticate	accord
Technology descriptions	+	detect	release	abstract	accord	include	generate	detector	title	output	base
Technology descriptions	+	form	have	release	light	abstract	include	accord		title	transmit
Technology descriptions	+	include	accord	abstract	release	title	perform	apparatus		supply	determine
Technology descriptions	+	specify	access	datum	cellular	point	include	determine	connect	base	order
Technology descriptions	+	display	abstract	base	determine	comprise	accord	signal	release	title	detect
Technology descriptions	+	configure	accord	abstract	release	include	base	title	receive	comprise	generate
Technology descriptions	+	light	provide	transport	generate	laser diode	intersect	curvature	comprise	exit pupil	additionally
Technology descriptions	+	couple	include	abstract	release	protrude	accord	node	have	electrically	transistor
Technology descriptions	+	record	read	include	information	accord	medium	determine	abstract	method	generate
Technology descriptions	+	release	abstract	accord	title	include	provide		have	inventor	ccording
Technology descriptions	+	base	determine	include	abstract	accord	release	associate		receive	method
Technology descriptions	+	operable	store	couple	tag	comprise	network	datum	convert	response	energy harvest unit
Technology descriptions	+	display	light	provide	device	include	backlight	shutter control portion	different image	drive	blink
Technology descriptions - cameras	+	image	base	video	title	apparatus	set	erase	+	lack	abstract

Table A.2 continued from previous page

Technology descriptions with "configure"	+	configure	include	abstract	release	accord	file	dispose	plurality	base	receive
Technology descriptions with "connected"	+	connect	include	abstract	release	accord	test	extend	title	configure	supply
Technology descriptions with "lens"	+	have	include	abstract	release	accord	emit	object	title	order	
Technology descriptions with "on to titled"	+	accord	abstract	release	title	method	provide	comprise		apparatus	include
Technology descriptions with "output"	+	output	configure	obtain	accord	signal	title	abstract	comprise	form	differentiate
Technology descriptions with "processing"	+	process	title	grant	patent	release	abstract	include	accord	configure	store
Technology descriptions with "receive"	+	receive	include	abstract	accord	release	perform	transmit	title	base	store
Technology descriptions with poor formatting	+	mean	sheet	convey	feed		+	surface	website	follow	accord
Appointments	++	appoint	word	upcoming	manage director	position	phone	maker	broadcaster	kid entertainment space	kid channel word
Dispositions/acquisitions by director	++	change	form	note	beneficial interest	file	hold	director word	disposition	report acquisition	acquisition
Dispositions/acquisitions by director	++	change	form	hold	note	file	beneficial interest	director	word	disposition	acquisition
Misc "deal" - mostly business deals	++	deal		word	announce	company	sign	buy	have	sell	not
Misc "launch"	++	launch	word		new	say	market	report	device	+	plan
Misc "launch" - new products	++	launch	word	company	announce	say	market	plan	device	set	recently
Misc "series" - product launches, layoffs, hackers	++	series		company	include	launch	smartphones	come	say	design	unveil
Misc "unveil" - models, titles, products	++	unveil	word		company	plan	launch	new	set	smartphone	
Patent applications	++	receive	application	apply	title	word	plus	apparatus word		include	title word
Patent applications	++	file	word	publish	patent application	method	patent	apparatus	application	device	thereof

Table A.2 continued from previous page

Patent applications	++	publish	patent	application	issue	invent	journal	output	sense request	memory cell array	datum
Patent applications	++	receive	publish	application	pto	title	kabhiki	apparat	word dateline	method	apparat word
Patent applications	++	receive	apparat	apply	application	title	method	word	apparat word	plus	transmit
Patent applications	++	receive	title	pto application	word	kabhiki	method	apparat	image	dateline	word dateline
Patent applications	++	receive	device word	application	title	apply	device	plus	method		image
Patent applications	++	publish	patent application	pto	olymp	information process system information process method	gyr	coltd	nontransitory computer	readable information record medium	file word
Patent applications	++	method	receive	title	application	apply	word	plus	device	drive	thereof
Patent applications	++	publish	receive	word	application	pto	title	apparat	dateline	method	kabhiki
Patent applications	++	receive	application	title	word dateline	apply	method	have	include	kabhiki	device
Patent applications	++	receive	application	word	title	apply	apparat	thereof	plus	include	have
Patent applications - filed	++	patent application	file	internationally	grow	graphene ball technology	explore	continuously	develop	submit	commit
Patent applications + application as in "apps"	++	file	application	patent	publish	word	method	apparatus	title	delete	partially
Patent grant	++	grant	word	title	patent	include	device	trademark	trade mark	have	apparatus
Patent grant	++	grant	title	word	plus	patent	method	device	apparatus	thereof	control
Patent grant and trademark grant	++	grant	title	word	patent	trademark	method	include	manufacture	pto grant trade mark	device

Table A.2 continued from previous page

Patent grant and trademark grant	++	grant	patent	title	apparatus	process	device word	device	image	have	apparatus word
Patent grants	++	grant	workflow event	deliver	title	patent	method	document	participant	task word	fluid anesthetic
Patent grants	++	grant	thereof	patent	title	word	method	control method	device	apparatus	manufacture method
Patent grants	++	grant	patent	plus	title	method	kabhiki	apparatus	device word	device	process
Patent grants	++	grant	title	method	patent	word	control	manufacture	device	apparatus	fabricate
Patent grants	++	grant	method word	title	patent	apparatus	plus	device	process apparatus	control	decode
Patent published	++	publish	patent	plus	title word		olympus	title word	sony	title device word	obtain
Patent published	++	file	patent application	publish	word	method	include	apparatus	decode	display	system
Patent published	++	patent	publish	title	word	plus	sony		olympus	method	device
Patent registration	++	patent	registration	title	publish	apparat	method	apparatus	system	thereof	include

**Table A.3. LDA content analysis and clustering top terms congruence with bag-of-words vectorization**

Acquisitions and growth estimates	-	year	estimate	acquisition	order	term	leak	court	be	company	lawsuit
Acquisitions of shares, assets and companies	-	acquire	process	ensure	article	apparatus word	query	reserve	contact	fabricate	encryption
Analyst and manager statements - often investments	-	start	analyst	say	boost	invest		company	pc market	shrink	market value
Announcements - appointments - deals - products	-	announce	word		bring	chief executive	company	security	link	electronic giant	engage
Appointments	-	claim	recently	program word	schedule	founder	fund	appoint	word	pto application	information process device information process method
Availability and distribution channels - batteries and technology related	-	datum	battery	month	stop	interface	confidence	notify	compound		avail
Components and combining	-	meet	measure	advantage	component	volume	strength	revive	executive vice president	counter	body
Earnings - phones	-	earnings	bank	bear	game console	x	say	dub	high level	protection	company
Filed patents - reports - bankruptcies - complaints	-	file	patent	application	initiative	faulty battery	controversial	say	nonetheless	channel estimation	word
General "future"	-	issue	likely	operate	app	future		audience	easy	convert	word
General "help" - "improve"	-	help	decision	say		company	official	capacity	satisfy	2	word
General "high management" topic	-	president	different	process apparatus	integration	abandon	calculate	gaming	campaign	spread	keyboard
General possibilities and opportunities theme	-	allow	addition		total	new company	draw	accelerate	slide	fill	say
General smartphone market theme	-	cost	improve	strategy	film	smartphone market	say	ready		type	compliance
Generic manufacturing tech theme	-	generate	handset	worldwide	couple	mobile device	facility	input	image form apparatus	approve	say

Table A.3 continued from previous page

International and globality theme	-	tech giant	maintain	choose	experience	government	suffer	jolt	core	margin	overseas
Market descriptions + corporate lead changes	-	market	cent	company		chief	remain	research	name	socalled	earn
Metatext	-	model	tap	check	fly	section	map	mix	refresh	moment	refund
Misc "\$" - revenues, loans, shares	-	\$	result	shift	promote	allege	words"s	company	finally	word	land
Misc "add"	-	add	group		week	standard	live	concentrate	rush	suspect	accompany
Misc "buy" - companies - stakes in - buy back	-	buy	extend	wait		million	ago	slump	learn	word	say
Misc "come" - often as "comes with"	-	come	phone		integrate	vendor	talk	will	say	fan	responsible
Misc "company"	-	thereof	operation	equip	suggest	arrange	reportedly	familiar	mat	competitive	guide
Misc "display" - showcase and technology component	-	display	away	project	taxis	prompt	evidence	contact	low price	simplify	overall
Misc "doubled"	-	software	double	decode	roll	ing	encode	quality	property	liquid crystal display	say
Misc "expect" - finance and tech	-	expect	use	view	belong	company	n	keep	arrive	propose	reputation
Misc "face" - countries and internationality and global themes	-	face	country	install	region	method word	commit		oppose	small business	display device word
Misc "firm" - announcements and statements	-	firm	say	affect	general manager	slightly	new generation	pc business	new smart-phones	specific	settop box
Misc "get" - business opportunities theme	-	get	enjoy	invite	location	cartridge	embed	percent share	axe		cds
Misc "ink" - sign a deal, new products	-	partnership	float	especially	new version	module	ink	major	tout	lcd	contend
Misc "lower" and "plunge" - product demand	-	line	low	back	shipment	euro	respect	email	say	settlement	assembly
Misc "make" - comebacks - goods - business descriptions	-	make		slash	category	<	match	insist	expensive	proceed	undergo
Misc "pay" and "pay off"	-	pay	period	machine	describe	company	clear	wide range	eventually	say	b

Table A.3 continued from previous page

Misc "plunge" - mostly shares	-	base	plunge	extract	negotiation	analyze		manner	globe	market ex- pectation	study
Misc "printing"	-	print	reject	save	trillion	web	code	firstquarter	urge	tech	
Misc "profit" - mostly statements	-	product	profit	and"s	represent	know	say	level	cite		word
Misc "rise" - sales - revenue - shares	-	rise	gain	point	light	slip	client	computer maker	assemble	offset	
Misc "shareholders and management" theme	-	worry	trademark	dividend	strong	trade mark	defend	pledge	bonus	valuable	stabilize
Misc "show"	-	show	player	transformation	retain	invent	news word	safe	evolution	telecom	total sale
Misc "store" - as in retail outlets and storage - availability	-	store	source	range	output	beat	challenge	hand	lowend		company
Misc "take" - "steps" - "lead" - "leaf"	-	take	capable	dateline	word		asset	company	war	personal computer business	tender
Mostly operating profits	-	operate profit	detail	delay	format	new phone	load	specification	connectivity	say	user equipment
New products - profits - values	-	need	individual	locate	prevent	say	officially		side	company	vary
Outsourcing - take over - closing sites - joint ventures	-	platform	popularity	console	peripheral	shut	outsource	spokeswoman	apparat apparat	apparat system	say
Patent applications	-	patent ap- plication	publish	file	event	associate	pto	word	internationally	initially	controller
Patent applications + poor formatting	-	+	website	publish	electronic device	fast	debut	invention	attach	title	method
Patent published	-	apparatus	system	method	eject	trade	remove	low level	song	title	cell sys- tem
Patent published	-	obtain	sony	index	ban	spin	vs	word	amount	sure	shed
Patent published	-	control	comprise	method	word date- line	device man- ufacture method	embodiment	video record	accordance	exit	relaunch

Table A.3 continued from previous page

Patent published	-	device	method	display apparatus	global recall	mobile phone business	word	ength word	fingerprint	network word	title
Patent registrations	-	like	well	solution		facilitate	split	company	little	fit	4 quarter
PCs and general tech	-	mean	shake	area	suit	combination	entry		widely	explore	credit
Poor formatting	-	£	industry	tv	configuration	adopt	train	snap	speculate	embark	turnaround
Poor formatting	-	plus	semiconductor r&d memory device word		olymp	word	grow market	market researcher	telecom giant	software developer	grant
Poor formatting + job cuts	-	plan		word	job	s		say	reserve	understand	import
Rankings	-	subsidiary	history	warn	ride	goal	reflect		rank	feel	mainly
Results and revenues	-	support	revenue	post	game	apparatus word	basis	wrong	say	temporarily	
Sales, revenues, and profits in won currency	-	sale	win	billion	say		executive	return	word	follow news release	spark
Standing in the market	-	ahead	content	chance	survey	card	wireless communication system word	correct	base station	error	unlimited
Standing in the market - global	-	head	numb	reach	account	list	say	contribute		export	workstation
Sueing - patent infringement topic	-	grow	globally	factory	say		television	sue	company	introduction	staff
Targets and expectations lowered	-	lose	positive		tablet computer	workforce	fiscal year	declare	lower	distribution	tv business
Technology descriptions	-	accord	release	abstract	title		provide		number	plurality	serve
Theme of blaming - "deal" "resign" "net income"	-	smartphones	deal	say	net income	company	2 quarter	resign	blame	effective	exec

Table A.3 continued from previous page

Theme of finality - "Leave" - "stop" - "end"	-	end	place	brand	achieve	leave	mark	company	significant	tra	tramark
Theme of games and music - "Play" - "Perform"	-	select	play	mobile phone	block	word	affiliate	availability	venture	requirement	limit
Theme of met expectations - "Raises" "tops"	-	raise	wholly own subsidiary	small	surface	break	cost cut	network equipment	miss		pyright
Theme of warfare - "battle" - "rival" - "target"	-	investor	word	nearly	new product	battle	fear	highlight	company	barrier	impact
Top leadership news - "restructure"	-	time	information		home	replace	agreement	right	dvds	restructure	fact
Unintelligible	-	appear	internet	parent company	segment	promise	protect		effect	demonstrate	net
Unintelligible	-	indicate	fix	owner	assume	approval	coil	discussion	organic	modify	assignee
Unintelligible	-	reduce	variety	board	member	announcement	contract manufacturer	share price	put	exhibit	freeze
Unintelligible	-	customer	able	move	loss		late	video	present	watch	photo
Unintelligible	-	network	fail	quickly	new model		discuss	mpany	manufacture facility	company	visit
Unintelligible	-	determine	supply	drop	rumour	example	helm	home appliance		length	say
Unintelligible	-	send	deliver	response	to"s	space	respectively	decrease	unlike	resource	instance
Unintelligible	-	detect	wide	television set	operable	approach	climb	bundle	doubt	peak	please
Unintelligible	-	not		case	compatible	instead	ask	extension	slay	exploit	cater
Unintelligible	-	give		turn	of"s	plant	tv's	pcs	competitor	version	global leader
Unintelligible	-	set	produce	go	benefit		company	say	word	attract	patent infringement
Unintelligible	-	continue	power	say	statement	complete	by"s	site	showcase	trend	company
Unintelligible	-	track	book	cell	manufacture method	course	reinforce	bond	reliability	pickup apparatus	pattern
Unintelligible	-	participate	in"s	because	easily	alongside	therefor word	squeeze	new line	rebind	quote
Unintelligible	-	address	development	intensify	violate	conduct	net loss	cash	inform	say	company
Unintelligible	-	include	relate	ccording		document	deploy	electrically	stack	wire	abstract

Table A.3 continued from previous page

Unintelligible	-	utilization	crore	infrastructure	storage device	from"s	datum center	new offering	new job	new operate system	new software
Unintelligible	-	incorporate	carry	reason	mainland	hold company	implementation	notice	boom	smart tv	angle
Unintelligible	-	day	update	outside	establish	personal computer	network device	smartwatch	speculation	on"s	tool
Unintelligible - "based on" - "fined"	-	switch	fine	settle	significantly	energy	convince	negotiate	preload	organisation	regulator
Unintelligible - "edge" and "stake"	-	find	way	apparat	stake	stand	edge	sector	percentage point	word	product
Unintelligible - few events	-	people	host	state	seal	handle	download	3d	say	organize	
Unintelligible - few profit/revenue forecasts	-	forecast	agree	request	thing	largely	family	short	word	tumble	
Unintelligible - general product theme	-	want		long	spend	dominate	compute	company	cloud	far	say
Unintelligible - general tablet theme	-	tablet	tell	predict	length word	count		global market	act	mount	putt
Unintelligible - often "jump on bandwagon"	-	jump	memory device	storage	fiscal	apparatus method	possibility	web browser	resignation	yearoveryear	ownership
Unintelligible - product lineups	-	perform	screen	net profit	effort	enhance	overtake	for"s	typically	task	embrace
Unintelligible - sports - technology	-	connect	identify	disclose	execute	compute device	inside	adjust		printhead	red
Unintelligible, but general camera topic	-	hit	production	maker	digital camera	division	electronic	indirectly	word	patent patent	
Value statements	-	value	decide	intend	information process	write	apparently	apparat information process method		annual	recover
Misc "decline" + Note 7 news	?+	decline	monitor	catch	fire	attempt	question	image apparat word	favour	say	corporation
Misc "spokesman"	?+	spokesman	review	say	fuel	device method	answer	dismiss	wrap	woman	company
Mostly stock shares acquisitions	?+	stock	involve	portfolio	copy	action	share	near	disposition	size	argue

Table A.3 continued from previous page

Partnerships and employment	?+	push	employee	partner	hope	investment		employ	slow	say	new range
Product availability	?+	available	begin	early	sign	cover	capture		enter	ability	amid
Share ownings	?+	share	online	own	directly		speed	picture	say	immediately	lineup
Cameras and lenses	+	have	image	camera	reveal	g	object	company	today	say	compose
Comparisons of company size or products	+	follow	rival	manufacturer	big		distribute	soar	poise	actually	scale
General "design"	+	design	concern	engineer	substantially	optimize	stress	accessory		specifically	dip
General music topic	+	music	stream	merge		succeed	wave	tie	suite	successfully	listen
Joint ventures - partnerships - consortiums	+	lead	joint venture	package	consortium		use"s	restore	word	consumer product	occupy
Market competitiveness	+	consider	growth	competition	grind	race	appeal	say		table	compatibility
Misc "build" - plants & manufacturing - infra	+	build		force	step	large	contract	say	heat	company	and/or
Misc "cut" - costs - jobs - forecasts	+	business	cut	company		confirm	say	supplier	exclusively	hire	responsibility
Misc "headquarters"	+	bid	headquarter	prepare	developer	emerge	carrier	form apparatus	organization	word	span
Misc "high" + metatext	+	high	read	memory device word	pass	comment	subject	fully	management	feed	
Misc "introduce" - usually services but sometimes products	+	service	introduce	enable	consumer	despite		soon	say	word	compensation
Misc "look"	+	look	develop	dispose		strike	opposite	print system	say	forward	aggressively
Misc "manufacture" + poor formatting	+	manufacture	server	laptop	award	similar	>	headset	letter	startup	clothe
Misc "new" - technologies - infra	+	new	aim	join	open		occur	say	strengthen	word	recognize
Misc "provide"	+	provide	semiconductor device	utilize	semiconductor	memory chip	dollar	outlook		activate	cellphone

Table A.3 continued from previous page

Misc "purchase"	+	purchase	computer	the's		transfer	touch screen	example implemen- tation	refer		mode
Misc "sell" - products and business	+	sell		seek	company	say	figure	delist	pressure	generation	trigger
Misc "try"	+	try	leader	apart	specify	collaboration		regard	reporter	cuttingedge technol- ogy	personnel
Misc "unit"	+	drive	unit	signal	pc	lightemitting	key role	expectation	current	profit mar- gin	happen
Misc "work"	+	work	cause		interest	say	company	implement	complaint	senior ex- ecutive	closely
Patent applications	+	receive	application	title	apply	word	method	device word	dateline	rate	unable
Patent published - job cuts - market position	+	position	communicate	innovation	consult service	job cut	olympus	slate	recovery	revamp	service provider
Printers and patent applications	+	kabhiki	printer	message	wireless network	apparar method	eye	pto	touch	medium word	accept
Recalls and problems with products	+	compete	good	problem	previously		recall	struggle	found	thin	company
Sales and shipment number comparisons	+	increase	quarter	think	ship	say	global smart- phone market	research firm	success	company	pick
Shares values reports + some sales	+	's	fall	close	percent	compare	market share	cent	office	verify	share
Smartphone theme - tech and finances	+	call	smartphone	charge	rule			exceed	public	say	domestic market
Targets and result reports	+	report	demand	target	company	chairman	ceo		market leader	word	global
Technology descriptions	+	transmit	believe	correspond	plurality	direction	trial	array	halt	failure	electrode
Technology descriptions	+	title word	manage	panel	substrate	composition	saving	useful		computer monitor	word

Table A.3 continued from previous page

Technology descriptions	+	offer	world	see	power supply	company	word	portion	equipment	attention	this"s
Technology descriptions	+	configure	medium	sense	inventor	program	include	control unit	abstract	perform	past year
Technology descriptions - "form", "structure"	+	form	lot	emit		include	half	incident	image sensor	consist	
Technology descriptions - new technology	+	feature	technology	news	memory	admit	flagship		spokesperson	say	company
Theme of "hardware manufacturing" - no clear actions	+	hardware	particular	impossible	operator	oversee	web site	supply chain	unit word	strongly	
Theme of general R&D	+	expand	develope	worth	currently	say		possible	pound	researcher	premium
Theme of opportunities and benefits for customers and users	+	user	create	access	combine	license	let	test	free	say	new device
Theme of resource requirements - "Secure" - deals and funding	+	require	secure	processor	render	function	new technology	lift	centre	gap	translate
Dispositions/acquisitions by director	++	change	hold	note	file	director	form	beneficial interest	director word	word	disposition
Partnerships and mergers	++	merger	flat	team	successor	separate	cheap	person	vice	say	downturn
Patent grants	++	grant	title	patent	word	method	plus	manager	device word	probe	alter
Patent infringements and few mergers	++	damage	shareholder	infringe	frame	display device	jury	smartphone maker	"	word	unclear
Patent registrations	++	publish	patent	title	registration	pto	apparat	method	device	market campaign	word
Product launches	++	launch	run	price		word	record	company	chip	say	operate system
Unveilings - mostly product launches at events	++	unveil	focus	word	series	upgrade	giant		say	page	new camera

**Table A.4.** LDA content analysis and clustering top terms congruence with TF-IDF vectorization

Announcements and product launches	-	make	world	smartphone	manage	brand		printer	attract	say	word
Buying parts of businesses mostly	-	buy	and"s	company have headquarter	site	hard	message	initiative	word	flow	
Dispositions/acquisitions by director	-	hold	note	period	worldwide	place	change	form	file	beneficial interest	disposition
General company descriptions - market, products	-	look	unit	revenue	pay	company		people	say	ship	flagship
General market descriptions	-	hope	struggle	amid	get	market research firm	company	worry	promise	job cut	pc market
General market descriptions -phones	-	image	smartphone market	date	suppose	supremacy	premium device	pay cut	enterprise business	patent infringement case	form apparatus
General rankings	-	boost	rotate	respectively	watch	read	let	inspire		forward	rank
General theme of future plans	-	plan	title word	handheld	say	matt	familiar		company	finish	route
Geography - location - market standing and manufacturing	-	spend	electronic	house	review	hear	grab	condition	lcd	manufacture plant	video game
Launches mostly	-	time	service	plurality	output	function		generation	collect	assign	computer system
Misc "\$"	-	\$	deliver	intend	lab		describe	company	word	trouble	plenty
Misc "%"	-	cent	estimate	reach	competition	surge	fail	storage	company	say	quote
Misc "country"	-	firm	country	s	giant	consider	rule	copy	court		invite
Misc "earn"	-	supply	state	utilize	earn	discharge	root	water	word	electric power	click
Misc "officially"	-	drive	group	divide	represent	clear	image form apparatus	urge	officially	halt	feed
Misc "PC"	-	store	pcs	endoscope	walk	communicate	past	follow news release	word	useful	tech company
Misc "post"	-	give	post	purchase	think	leak		say	company	thank	lightemitting
Misc "reporter"	-	large	stand	reporter	locate	distribution	poise	fee	constantly	company	word
Misc "reports"	-	report	model	believe	word	decrease	electronic giant		adapt	code	company
Misc "takeover"	-	deal		develop	stake	talk	say	word	carry	company	sound

Table A.4 continued from previous page

Misc "tenders" and "global"	-	home	new product	smartphone sale	researcher	modify	global market share	tender	consistent	recent	word
New mobile phones - new jobs	-	phone	request	edge	oration number	variety	gadget	transmission	assess	ride	incur
Patent applications and poor formatting	-	control	obtain	word date-line	establish	affiliate	ing	record medium	negotiate	reproduce	plurality
Patent grants	-	recover	system word	form apparatus	percentage point	pound	propel	spur	apparatus control method	up	chassis
Patent grants + poor formatting	-	device	apparatus word	heavily	material	image datum	land	tip	slowly	computer printer maker	original
Poor formatting	-	cover	perform	to"s	frame	global	parameter		bounce	confident	earthquake
Poor formatting	-	's	division	keep	unlike	bolster	closely	trim	pace	company	overhaul
Poor formatting	-	£	showcase	film	climb	apparently	will		word	stick	weak demand
Poor formatting	-	plus	package	top	payment	word	image sensor	builtin	globe	electronic maker	nd
Poor formatting	-	process	device word	attempt	memory device	compensate	successfully	world have large maker	scrap	reader	stall
Poor formatting	-	work	know	test	apparatus word	manufacturer	series		device length word	b	highdefinition
Poor formatting	-		word	government	sport	plummet	install	option	rest	new version	little
Poor formatting	-	add	transmit	operate	capable		search	detail	communication	responsible	market value
Poor formatting and metatext	-	change	head	target	lack		surprise	section	say	word	effect
Poor formatting and metatext	-	call	method word	r	discount	company	vibrate	holiday	cent stake	retail store	image apparatus word
Poor formatting and metatext	-	maker	reputation	euro	fight	effectively	driver	desktop	word	utility	cash reserve
Poor formatting mostly	-	likely	feel	ahead	slip	pinch	strong	shoot	disappoint	wrong	revamp
Product availability	-	available	opportunity	client	operate loss	use"s	new generation	resource		say	international file date
Profits - sales - revenues	-	hit	numb	project	ensure	draw	present invention	security	vehicle	originally	company
Revenue and profits in won currency	-	win	system	merger	choose	net profit	said"s	company	succeed	encourage	word

Table A.4 continued from previous page

Sales, shipments and revenue	-	year	late	ago	innovation		ceo	words"s	new de- vice	score	say
Share values - sales numbers - profits	-	software	statement	say	increasingly	million		company	easy	lineup	wrap
Tenders, patents and poor formatting	-	issue	promote	present	stay	word	founder	market re- searcher	senior ex- ecutive		apparatus infor- mation process method
Theme of "restructuring" and "taking over"	-	take	save	total	step		restructure	say	company	learn	connection
Theme of apps, phones and downloading	-	offer	access	of"s		download	track	protect	company	tbreak	say
Theme of business overtakes and surpassing	-	determine	reveal	corporate	surpass	overall	profit mar- gin	activate	balance sheet	transistor	communication apparatus
Theme of comparing business sizes	-	mobile phone	view	attribute	ram	chief execu- tive offi- cer	efficient	execution	powerful	production facility	
Theme of expectations - new product features mostly	-	expect	acquisition	develope		say	word	director	position	dateine	company
Theme of flagship phones	-	discuss	disposal	asset	wait	abroad	attention	fear	studio	stress	patent bat- tle
Theme of future goals - mostly sales	-	achieve	predict	strategy	management	goal	widen	partly		antenna	expertise
Theme of general fixing	-	see	region	raise	settle	app	say	company	recovery		new tech- nology
Theme of general increases and decreases	-	end	decline	company	say	version		primarily	shortly	vs	word
Theme of lawsuits	-	open	inside	suspect	measure	allege	jury	venture	conclude	mode	amount
Theme of lawsuits	-	ask	judge	incorporate	rapidly	mention	round	heart	victory	essentially	weaken
Theme of market domination	-	dominate	far	stop	contribute	charge	week		by"s	finally	machine
Theme of outsourcing	-	shift	different	storage device	outsource	terminal	member	display ap- paratus	manufacture method	acid	behalf
Theme of shares dynamics	-	fall	low	double	break	projection	tv	say	near	regain	outlook
Unintelligible	-	display panel	link	popular	family	tumble	regard	legal bat- tle	costcutting	word	digital mu- sic player
Unintelligible	-	help	run	improve	computer		equip	confirm	photo	company	word
Unintelligible	-	build	demand	require		battery	previously	slow	company	say	prove
Unintelligible	-	result	high	reward	argue	slightly	prompt	arm	percent share	pc maker	display de- vice word
Unintelligible	-	agreement	expert	quickly	contain	easily	pressure	concern	advance	picture	slide
Unintelligible	-	couple	remain	program word	license	tie	risk	possibility	lg	successful	
Unintelligible	-	enable	small	response	correspond	money	substrate	great	deploy	campaign	pledge

Table A.4 continued from previous page

Unintelligible	-	feature	addition	big	employ	news	input	chance	company	business user	new business
Unintelligible	-	increase	lead	follow	try	go		month	sign	platform	say
Unintelligible	-	bring	design	order	mainly	roll	0	bear	word	global smart-phone market	headline
Unintelligible	-		begin	not	tell	line	currently	early	replace	television	beat
Unintelligible	-	leader	capability	general manager	study	maximum	lot	category	check	render	infrastructure
Unintelligible	-	face	subsidiary	fast	exceed	fix	depend	rumour	match	company	
Unintelligible	-	provide	relate	free	hire	block	transform	indicate		suit	basis
Unintelligible	-	percent	seek	well	say	company	engage	worker		relationship	current
Unintelligible	-	investor	technology giant	hardware	company	interact	laser	optic	race	overheat	evidence
Unintelligible	-	use	expand	decide	say	rival		operate system	team	company	complete
Unintelligible	-	industry	point	extend	rate	supplier	difference		typically	possibly	say
Unintelligible	-	associate	exchange	history	understand	sensor	trade mark	adoption	format	favour	
Unintelligible	-	core business	sale growth	eventually	correct	consolidate	erode	cope	battery word	but's	liquid
Unintelligible - general theme of recent events - mergers - patents	-	recently	move	reserve	solution	collaborate	console	merge	explain	word	indirectly
Unintelligible - possibly new products	-	push	away	cause	event	propose	provision	word	q2	motion	indication
Unintelligible - printing	-	print	excite	host	secure	impact	light incident surface	email	leadership	n	improvement
Unintelligible - tech and corporate scandals	-	focus	combine	chip	dispose	disclose	reflect		vote	portfolio	company
Unintelligible - TVs and PCs	-	with's	sector	device method	lock	image apparatus	pc business	industry leader	that's	cement	information process
Unveillings - introductions - launches - "comes with"	-	come	compete	laptop	tablet		investment	say	director word	company	interview
Market analyst statements	?+	analyst	consumer	say	catch	instead		software company	interest	announcement	company
Market share fluctuations and changes	?+	technology	market share	lose		company	say	miss	for's	grind	patent infringement
Misc "sell" - products and business	?+	sell	customer		say	camera	partner	mean	steal	be	fact

Table A.4 continued from previous page

Partnerships	?+	information	record	turn	abandon	integration	register	shelve	strategic partnership	cease	
Partnerships - mostly flipphones	?+	partnership	retailer	therefor word	shed	consumer electronic	revive	rely		ebooks	convergence
Patent infringements and lawsuits	?+	smartphones	cost	fire		infringe	tablet	accuse	compress	say	company
Sales and operating profits	?+	sale	drop	operate profit	network	g	say	list	word	delay	blame
Share ownings	?+	share	grow	own	directly	value	company		say	enjoy	movie
Stock trades and prices	?+	involve	stock	action	share	highend	effective	2	difficult	disposition	archival
Theme of fiscal year developments	?+	account	trade	quarter	figure	globally	mobile	slash	reportedly	fiscal	
Theme of general establishing and joint ventures	?+	found	soon	leverage	spot	joint venture	wj	module	company	new share	
Theme of general manufacturing	?+	method	manufacture	aim	processor	spread	word	instruction	fan	innovative product	fetch
Theme of production	?+	form	produce	production	discontinue	light		bank	carrier	company	celebrate
Theme of shipments	?+	have		shipment	question	object	overseas	automatically	irradiate	satisfy	saving
Acquisitions of assets and bids	+	acquire	highlight	bid	word	trust	emerge	plunge	combination		arrive
Announcements and contracts	+	announce	worth	word	claim	level	crore	operate revenue	share	patent	sale
General theme of threats	+	prepare	direction	portion	article	shape	instance	compatible	defend	premium	war
Handsets, phones and internet	+	create	agree	lose	handset		competitor	internet	transfer	suggest	play
Home electronics strategy	+	tv	game	chairman	say	part	course	export		vie	restore
Launches and trademarks	+	launch	gain	word	enter	recall	way	mark	say	ramp	new phone
Metatext	+	accord	release	abstract	include	title	connect	datum	provide	apparatus	
Metatext	+	publish	pto	title	receive	application	kabhiki	apparat	word	identify	method
Misc "overtake" and "headset"	+	meet	write	warn	overtake	panel	headset	dollar		recommend	contend
Misc "earn"	+	area	earnings	term	share price	admit	vendor		success	sure	largely
Misc "join"	+	join	personal computer	the's	strengthen	force	role	tech giant	expectation	initially	process apparatus
Misc "operate"	+	operation	arrange	especially	outside	day	chief		cite	partition wall	space
Misc "transaction"	+	business	commit	inventor	mobile device	online	convert	say	text	transaction	completion
Misc "unveil"	+	unveil	effort	name	fully		word	company	penetrate	cash	say

Table A.4 continued from previous page

Patent applications and technology descriptions	+	include	emit	example	advantage	light source	embed	pass	accordance	calculate	stage
Poor formatting	+	+	website	development	publish	abstract	board	accord	title	award	number
Poor formatting	+	compute	>	location	memory device word	process method	realize	device manufacture method	dependent	research company	release date
Poor formatting and metatext	+	electronic device	dateline	memory	update	pto application	exit	party	ring	subcontractor	length word dateline
Poor formatting and patents published	+	word	thereof	publish	patent application	sony	implement	scan	file	apparatus	5
Profit forecasts	+	long	forecast	president	say	notebook		net loss	company	split	thing
Technology descriptions	+	support	medium	select	enhance	occupy	park	envision	disappear	power transmission device	include
Technology descriptions	+	decode	seal	hacker	edit	manipulate	encode	extract	remove	and/or	significantly
Technology descriptions	+	base	configure	detect	include	adjust	type	transmit	perform	adjustment	slay
Technology descriptions - phones and cameras	+	3d	software maker	pixel	surface	display device	cell	lens	explore	browse	smartphone user
Technology descriptions - printers	+	comprise	send	respect	recognition	separate	exist	analysis	cartridge	express	late version
Technology descriptions with "display"	+	display	channel	photograph	right	control method	patent patent	information process method	liquid crystal display	accessory	computer maker
Theme of general newness and future visions	+	start	need	new	challenge	despite	say	size	possible	offset	contact
Theme of internal HR instability	+	set	employee	leave	word		return	lawsuit	sue	company	invention
Theme of local markets and trials	+	capture	exclusively	ccording	subject		trial	refer	restrict	sheet	home market
Theme of manufacturing facilities	+	plant	say	factory	optimize		sense	array	peak	company	construction
Theme of personalisation	+	personalize	personalization	personalize photo cover	personalize photo album greet card	personalize feature	personalize favorite	personalize expressive content	personalize experience	personalize digital camera	personalize user interface
Theme of phones and managers	+	manager	finance	debt	cellphone	streamline	responsibility	database	spinoff	mobile phone business	say
Theme of power generation	+	generate	power	controller	desktop computer	new function	performance	equipment	wide range	clean	refrigerator

Table A.4 continued from previous page

Theme of pricing	+	price	prevent	company	expansion	core	thin		detector	glass	necessary
Theme of profits	+	compare	profit	hand	smartphone maker	apparatus method	patent dis- pute	rush	strong sale	company	record word
Theme of shares and scandals	+	rise	distribute	signal	jump	video	because	scandal	resign	back	say
Theme of stability - staying the same	+	continue	show	range	retain	maintain	delivery	speed	in"s	substantially	as"s
Theme of user experience	+	user	allow	reduce	case	program	content		load	document	good
Unintelligible	+		billion	switch	senior vice president	probably	respond	web	digital	on"s	port
Unintelligible - "bribes" and "global market"	+	executive	appear	global market	conduct	say		student	bribe	mind	company
Unintelligible - Second quarter of fiscal year	+	like	find	growth		want	say	company	2 quarter	decision	word
Unintelligible - shareholders and forecasts	+	shareholder	screen	damage	speak	computer giant	previous year	world have maker	mix	multitask	unleash
Appointments - businesses or personnel	++	contract	appoint	press	address	integrate	embodiment		dub	widely	approval
Investments and benefits - for businesses and consumers	++	able	invest	company	say	benefit		pull	upcoming launch	collaboration	spokesman
Patent applications	++	application	patent	file	suffer	handset division		fault	say	company	world have cellphone maker
Patent applications	++	receive	application	title	apply	word	battle	device word	technology company	ing	developer
Patent applications - international	++	patent ap- plication	server	internationally	pyright	mpany	complaint	file	roughly	throw	translate
Patent grants	++	grant	patent	title	word	plus	apparatus	adopt	accept	home ap- pliance	firstquarter
Patent registration	++	patent	publish	title	registration	apparatus	apparatus	audio	software giant	technique	communication device
Patents filed	++	file	patent	investigate	olymp	live	bar	common	broadcast	dim	linear
Product introduced to market	++	market	product	introduce		say	problem	company	office	particularly	book
Theme of layoffs	++	close	cut	loss	job	chief exec- utive	company	avoid	shut	happen	electronic product
Trademark grant	++	source	trademark	length word	fabricate	word	include	title	grant	method apparatus computer program product	plus grant trade mark