Tampereen yliopisto

Annika Viitanen

SKILL MAPPING USING NATURAL LANGUAGE PROCESSING

Master's Thesis
Faculty of Engineering and Natural Sciences
Examiners: Professor Matti Vilkko
University Instructor Mikko Salmenperä
November 2022

ABSTRACT

Annika Viitanen: Skill Mapping Using Natural Language Processing
Master’s Thesis
Tampere University
Master’s Program in Automation Engineering
November 2022

Digitalization transforms the world faster than ever seen before. It is also impacting the ways of working and work itself. Companies are trying to implement modern technologies to enhance their production but in order to do so, companies need skillful employees. Due to new technologies and new ways of working, employees need new skills in working life to be able to perform in their job.

These future skills can be divided into two categories: soft and technical skills. Future soft skills include skills such as communication, critical thinking, ability to learn, teamwork skills and leadership skills. These skills are typically hard to teach. The future technical skills include modern technology skills such as artificial intelligence, machine learning, cloud computing and cyber security. However, even more important technical skills in the future are data related data literacy skills that everyone should have. Research suggest that even non-technical roles require some technical skills more and more in the future.

Because skilled personnel is important for companies, companies also need to know what kind of skills their employees currently have to develop a skill strategy for the future. Building a company-wide skill profile is a time-consuming task that needs a lot of resources since the skill data is usually in various text documents. Therefore, in this thesis it was research whether a company skill profile could be formed using modern natural language processing techniques.

Job description and job advertisement texts were analyzed and compared using artificial intelligence application that used natural language processing. It was investigated if the application was able to recognize the skill related words from the text data and build a visual skill profile from them.

Keywords: natural language processing, future skills, digitalization. machine learning

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

TIIVISTELMÄ

Annika Viitanen: Osaamiskartoitus luonnollisen kielen menetelmin
Diplomityö
Tampereen yliopisto
Automaatiotekniikan DI-ohjelma
Marraskuu 2022

Digitalisaatio muuttaa maailmaa nopeammin kuin koskaan ennen. Se vaikuttaa työskentelytapoihin ja työpaikkoihin. Yritykset yrittävät ottaa käyttöön nykyaikaisia teknologioita tehostaakseen toimintojaan, mutta siinä onnistuakseen yritykset tarvitsevan osaavia työntekijöitä. Uusien teknologioiden ja uusien työskentelytapojen myötä työntekijät tarvitsevat työelämässä uusia taitoja suoriutuakseen työssään.

Nämä tulevaisuuden taidot voidaan jakaa kahteen kategoriaan: pehmeät ja tekniset taidot. Tulevaisuuden pehmeisiin taitoihin sisältyy taitoja, kuten kommunikaatio, kriittinen ajattelu, oppimiskyky, ryhmätyötaidot ja johtamistaidot. Pehmeitä taitoja on yleensä vaikea opettaa. Tulevaisuuden teknisiin taitoihin kuuluu esimerkiksi tekoäly-, pilvilaskenta- ja kyberturvallisuusosaaminen. Näitäkin tärkeämpiä teknisiä taitoja ovat datan käsittely- ja tulkintataidot. Datataitoja tulisi olla jokaisella työntekijällä. Teknisiä taitoja tullaan vaatimaan entistä enemmän tulevaisuudessa myös ihmisiltä, jotka työskentelevät muissa kuin teknisissä rooleissa.

Koska osaava henkilöstö on yrityksille tärkeää, yritysten on myös tiedettävä, millaisia taitoja työntekijöillä on tällä hetkellä kehittääkseen tulevaisuuden osaamisstrategiaa. Erityisesti suuren yrityksen laajuinen osaamisprofiilin tekeminen on aikaa vievä tehtävä, joka vaatii paljon resursseja, koska tieto henkilökunnan osaamisesta löytyy yleensä monista eri tekstidokumenteista. Siksi tässä opinnäytetyössä tutkittiin, voitaisiinko yritysten osaamisprofiileja muodostaa nykyaikaisilla luonnollisen kielen käsittelytekniikoilla.

Tässä työssä työnkuvauksia ja työpaikkailmoitustekstejä analysoitiin ja vertailtiin keskenään luonnollisen kielen käsittelyä käyttävän tekoälysovelluksen avulla. Diplomityössä tutkittiin, pystyikö sovellus tunnistamaan osaamiseen liittyvät sanat tekstidatasta ja rakentamaan niistä visuaalisen osaamisprofiilin.

Avainsanat: luonnollisen kielen käsittely, tulevaisuuden osaaminen, digitalisaatio, koneoppiminen

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck –ohjelmalla.

PREFACE

This thesis perfectly combined my interest in technology and future skills, and I could not have thought of a more interesting topic. Some chapter just flew by and were easy to write, while some sections felt really clammy and it felt like I was stuck in them for longer that I would have wanted to. For the most part writing felt nice, but other events in life, such as moving abroad, finding a new apartment, giving up my own apartment and almost all my belongings, and starting a new job, slowed the process more than I was expecting and made the writing a bit stressful. I am happy that this is now done.

I want to thank my examiners Matti Vilkko and Mikko Salmenperä helping me with this faster than usual thesis project. Especially I want to thank Mikko Pelkonen, Mathias Höglund and Jessica Enholm from Cargotec for letting me do my thesis for Cargotec and supporting me along the way. Also, thank you Erika Palm, for hiring me to Cargotec in the first place and believing in me. Another round of special thank yous goes to Headai, especially Anu Passi-Rauste and Neslihan Dinçel for brainstorming with me and supporting me while doing the practical part of the thesis.

Biggest thank you goes for my parents for encouraging and supporting me and always believing in me. Last but not the least, huge thank you to all my friends as well. Even if I forget everything that I learned in the courses, I will always remember our best moments together.

Katwijk aan Zee, 6th of November 2022

Annika Viitanen

CONTENTS

1. INTRODUCTION 1
2. COMPETENCY NEEDS IN THE DIGITAL REVOLUTION 3
    2.1 Digitalization 3
        2.1.1 Industry 4.0 4
        2.1.2 Circular economy 4
    2.2 The effects of digitalization on work 5
    2.3 The effects of digitalization on skills and competencies 9
3. NATURAL LANGUAGE PROCESSING 15
    3.1 The evolution of NLP: Then and now 15
    3.2 Natural language processing & machine learning 16
        3.2.1 Tokenization & preprocessing 17
        3.2.2 Machine learning 19
        3.2.3 Semantic computing 20
    3.3 Corpus-based methods 21
    3.4 Neural networks & self-organizing map 22
4. IMPLEMENTATION 25
    4.1 Short company introductions 25
    4.2 Input data 25
        4.2.1 Cargotec data 26
        4.2.2 Comparison data 27
    4.3 The AI solution 27
        4.3.1 How does it work? 27
        4.3.2 How is it used? 28
5. RESULTS 31
    5.1 How to read the skill mind maps? 31
    5.2 Cargotec skill mind map 31
    5.3 Skill mind maps of the industry clusters 40
        5.3.1 Machine & metal industry's skill mind map 40
        5.3.2 Electronics & electrical industry's skill mind map 41
        5.3.3 IT industry's skill mind map 43
6. CONCLUSION 45
REFERENCES 47

ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
AR	Augmented Reality
BOW	Bag of Words
BMU	Best Matching Unit
CBOW	Continuous Bag of Words
CNN	Convolution Neural Networks
CPS	Cyber-Physical System
EEI	Electronics & Electrical Industry
HR	Human Resources
IDF	Inverse Document Frequency
ILO	International Labour Organization
IoT	Internet of Things
IT	Information Technology
ITI	Information Technology Industry
JSON	JavaScript Object Notation
LSA	Latent Semantic Analysis
ML	Machine Learning
MMI	Machine & Metal Industry
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
OECD	Organization for Economic Cooperation and Development
RNN	Recurrent Neural Networks
SOM	Self-Organizing Map
URL	Uniform Resources Locator
UX	User Experience
VR	Virtual Reality

.

1. INTRODUCTION

Digitalization is transforming the world at a rapid pace and changing the business sector by replacing physical products with services, increasing the use of digital platforms, and improving production efficiency. Companies are trying to boost and grow their business by exploiting the benefits of digitalization using modern technologies and providing new products and services to satisfy their customers' needs. To succeed in this, companies need competent employees with the right skill sets. However, with modern technologies emerging and customer needs changing continuously it has become even harder to recognize what kind of competencies are needed now, and in the future (Fonseca & Picoto 2020).

To get the skills transformation right, an organization should start by identifying their current skills and competencies to be able to recognize the possible skills gaps (De Smet et al 2021). This requires a lot of manual labor since the competencies are usually documented in an open-text format, such as a job description if documented at all. In the case of hundreds or even thousands of employees, creating an overall picture and categorizing the organization's current competencies becomes a laborious and time-consuming task, when all the competency-related texts need to be first reviewed by people.

Even though competency mapping is still strongly human-driven, using modern natural language processing (NLP) has shown markable potential in solving this problem (Garman et al. 2021). For example, Goldenstein et al. (2015) studied organizational research texts trying to categorize them using NLP. Montelisciani et al. (2014) studied team formation using NLP analyzing ideal skills and wishes to form a well-cooperating team. Garman et al. (2021) used NLP to format a competency framework using NLP and compared the framework to a human-made competency framework. Recently, Technology Industries of Finland (fin. Teknologiateollisuus ry) produced a Skills Data Playbook in 2022 where they gathered case examples of how data analytics and artificial intelligence approaches can be used to better predict the future skill needs of industrial industries. The results of these studies have varied but overall NLP has shown a potential to at least assist humans when categorizing large text sets. Technology Industries of Finland believe that the use of different data analytics and AI approaches to predict skill needs will increase in the future (Technology Industries of Finland 2022a).

The purpose of this thesis is answer to these questions:

- What are the future skills needed during the digital revolution in industrial companies? and;
- Can NLP pick skill related words from job descriptions and job advertisements, and create a skill profile from text data?

The aim is to form an in-depth understanding of the possible skills gaps between the existing skills in the Finnish manufacturing company, Silicon Valley-based companies, and the researched future skills. To achieve this, an understanding of the future skills and skills transformation is formed with a literature review, the company’s current competencies are mapped to a mind map and that mind map is compared to other companies’ competency mind maps. The mind maps are formed from job descriptions and job advertisements using NLP instead of using, for example, LinkedIn profiles due to legal restrictions.

The first part of the thesis consists of conducting a literature review of future skills, skills transformation, and what drives it. In the second part the research method, NLP, is introduced and the third part focuses on describing the implementation of the research. After the implementation chapter, the results of the research are introduced and reviewed, and in the last chapter, a conclusion is formed.

2. COMPETENCY NEEDS IN THE DIGITAL REVOLUTION

This chapter focuses on what affects the competency changes and what are the needed competencies now and in the future. To understand the change in competency needs, it is important to understand digitalization, Industry 4.0 and the circular economy. After having the background information, the next sections focus on how digitalization transforms work, what are the future skills and competencies and how competencies are mapped in companies.

2.1 Digitalization

Simply said digitalization is a process that transforms analog to digital which is sometimes described with the word digitization. However, digitalization is more than just digitization. It is a visible phenomenon in the whole society, and the changes caused by it are transferred through industries to the operations of companies as well (Ilmarinen & Koskela 2015). Parviainen et al. (2017a) define digitalization as a phenomenon that makes Big Data available to large audiences and automates work, services, and production further than before. Digitalization transforms how manufacturing companies interact with their stakeholders, changes the value chain creation, and how data is shared with customers and suppliers (Porter & Heppelmann 2014). Kohtamäki et al. (2020) and Parvainen et al. (2017) state that digitalization increases the efficiency of industrial companies and provides new business opportunities due to servitization and new technologies. Data gives companies the possibility to analyze their performance more accurately which improves their value creation (Kohtamäki et al. 2020) and makes it possible to monitor functions in real-time (Parviainen et al. 2017a).

In addition to digitalization changing the business models, it changes also work. Modern technologies can replace existing jobs and work tasks, but it also creates new ones. Digitalization changes both the ways of working and places of working and it affects each industry but on a different scale. It offers new opportunities for companies, and individuals, and changes the operations and roles of organizations. (Parviainen et al. 2017a) However, digitalization also creates challenges for companies. For example, integrating all the systems into a single entity, liability and security risks, and privacy problems related to data. Still, adapting modern technologies is seen beneficial for the companies. (Langley et al. 2020) Digitalization can also boost sustainable circular economy by streamlining production which then creates less waste, optimizes energy consumption, and helps to make more durable products (Antikainen et al. 2018).

Bouée & Schaible (2015) have categorized digital transformation into four categories that are digital data, automation, digital customer access, and connectivity. These categories hold within many technologies that are connected to digitalization such as robotics, Internet of Things (IoT), data analytics, cloud computing, wearable technologies, and platform applications. Working with data allows better prediction and makes decision-making easier and more reliable when automation combined with artificial intelligence (AI) makes machines more autonomous which lowers the error rates and operating costs. With connectivity, value chains become more agile when everything is synchronized automatically and with connectivity, the digital customer access is also possible. (Bouée & Schaible 2015) All these technologies affect work since a large part of manual work is now automated and some decisions can be made without human intervention. Still, humans are much needed in value creation, and it seems that digitalization has created a lot of new job opportunities.

2.1.1 Industry 4.0

The digitalization of the industrial industry is often referred to as industry 4.0. To portray how radical the changes caused by digitalization in the industry are the term Industry 4.0 is the fourth industrial revolution. It follows the invention of steam machines in the first industrial revolution, mass production during the second industrial revolution, and moving from analog electronics to digital electronics in the third industrial revolution. (Skender & Ali 2019)

The characteristics of Industry 4.0 is the connectivity of objects with IoT, cloud computing, and cyber-physical systems (CPS) (Marr 2016). Important technologies of Industry 4.0 are also AI and autonomous systems (Skender & Ali 2019). Industry 4.0 describes how the opportunities created by digitalization are utilized in the industrial industry's software and device development (Sanders et al., 2016) and as a phenomenon it aims to meet personalized needs as well as improve competitiveness by bringing production costs down despite the customized production (Wang et al. 2016).

When implementing Industry 4.0 it is important to pay attention to the company's solvency so that it can be slowly introduced into existing factories without disturbing ongoing production. Industry 4.0 production system should be reliable, accessible, and sustainable and extra attention should be paid to the system's cyber security due to the large amount of data. Unauthorized access to systems and data could jeopardize business continuity and cause monetary damages. Implementation of Industry 4.0 must happen so that it follows energy and emission policies and limits. (Drath & Horch 2014) With Industry 4.0 technologies, companies can increase their value production and product quality as well as create new business possibilitites. Also, with better production systems companies can achieve shorter manufacturing and delivery times. (Fonseca 2018)

2.1.2 Circular economy

Digitalization drives the circular economy and modern technologies enable the change towards circular thinking and sustainable innovations. (Stock & Seliger 2018) Modern technologies help to locate used products and estimate their condition and show the availability. (Antikainen et al. 2018) With circular economy, product lifecycle changes from linear to circular as the name suggests. This means that instead of throwing unused products to the landfill, the product lifecycle would form a loop, where products are recycled, reused, or remanufactured (Auktor 2020, p. 16). In these fields, circular economy has huge potential for value creation but this to happen needs a lot of cooperation between different industries, new policies and appropriate financial models (Auktor 2020,

p. 29). Circular economy can create more job opportunities, and new green skills are needed from the workforce (Auktor 2020, p. 30).

Instead of selling product to one customer who then uses the product *x* amount of time and then throws it away, durable products can be rented or shared. Like digitalization, also circular economy can act as a competitive advantage for businesses. (Antikainen 2018) Antikainen et al. (2018) define three different circular business models that are slowing, closing, and narrowing the loop. Slowing refers to slowing down the linear economy by extending the product lifecycle whereas closing the loop means improving material recycling, and narrowing the loop tries to use fewer resources during production. These circular business models can be all used at the same time. (Antikainen et al. 2018)

It is estimated that circular economy would create over 6 million jobs if people have the needed skills. Therefore, training and education play a huge role in circular economy and digitalization. It is also important that these skills are taught in various countries that need the skills in order to build a more sustainable economy. (Auktor 2020, p. 30)

2.2 The effects of digitalization on work

As shortly mentioned previously, digitalization transforms work at an unforeseen pace. Digitalization has caused some fears of people losing their jobs. However, according to some forecasts, it is likely to create more jobs than before, following the trend of previous industrial revolutions. (Wellener et al. 2018) Autor (2015) emphasizes that it is often overestimated how able machines are to replace human labor due to the lack of assessing the importance of automation and workforce cooperation in productivity growth. He also mentions that the role of intuition and human judgment in work tasks is underestimated (Autor 2015). It is still quite evident that some jobs will disappear, but new jobs will emerge too. According to Hirschi (2018), it is very difficult to reliably estimate what percentage of jobs will disappear since it depends on various factors.

Level of education plays a big role in employability. Highly educated people working with complex tasks will have a better chance of employment while less educated people and routine task workers will have weaker opportunities for employment as digitalization progresses. (Parvianen et al. 2017) On the other hand, it is argued that as digitalization progresses job polarization too. This means that the amount of high-skilled and low-skilled jobs would both increase but then the mid-skilled jobs would slowly disappear. This would be due to the middle-skilled jobs, such as machine operators, being easy to automate. (Hirschi 2018) Technology Industries of Finland says that in the year 2035 half of the young adults in Finland should have a university degree (Technology Industries of Finland 2022b) which means that there would be more high-educated people in Finland in 2035 than ever before. Yet, in October 2022, Finnish Broadcasting Company Yle reported that the education level of young adults in Finland has stayed in 40 percent when other Organisation for Economic Cooperation and Development (OECD) countries education level has increased by 27 percent to 48 percent (Hallamaa 2022). Hallamaa (2022) states that highly educated people have around 10 percent better employment rates than those who have completed only secondary education in OECD countries. This applies especially to women, since 86 percent of women with higher education are employed, while only 38 percent of women without a secondary degree are employed. Education level also affects salaries as people with higher education levels earn distinctly more compared to those without high level of education. (Hallamaa 2022)

Nonetheless, what is certain is that there will be major changes in the job structures (Hirschi 2018) and some of the structural changes are location dependent. It is estimated that there will be more new jobs close to large residential areas close to universities and technology centers whereas the employment possibilities are predicted to weaken outside of big cities. (Parviainen et al. 2017a) This difference can be found also between OECD countries meaning there will be more jobs in highly digitalized countries than in those that are not as digitalized (Hirschi 2018). However, since digitalization has made remote work possible, which was even boosted by the COVID-19 pandemic, location dependence is less significant. When employees can communicate with each other and modify documents in real-time using different digital platforms it is easy to stay connected no matter where employees are located physically. (Leonardi 2021) For example, augmented reality (AR) and virtual reality (VR) enable remote maintenance, production planning, quality control, displaying different functions remotely, remote guidance, and simulating processes from anywhere. This means that also other jobs than jobs traditionally done in offices can be done remotely. (Alcácer & Cruz-Machado 2019)

Work is no longer a place to go but rather a place to connect (Accenture 2017). When work can be done from anywhere and at any time, the line between leisure and work becomes blurred and can lead to employee stress and increase employee burnout. Additionally, some employers have started to monitor their workforce, creating distrust between employees and employers which contributes to a stress and anxiety factor for employees too. (Perna 2022) On the other hand, digitalization and remote work make it possible for work to become more independent and flexible. Besides work being more flexible, work becomes more productive and efficient when different technologies are used smartly alongside people. Due to the increase in the use of cloud services, employees can access devices and information completely digitally and can communicate with colleagues, customers, and partners from anywhere. However, remote work may negatively impact the relationships that develop in physical workplaces and the quality of work-life relationships deteriorates, as face-to-face encounters decrease. (Degryse 2016)

One of the significant employment trends is the growth of the gig economy which includes working through applications and crowd work. Crowd work means work that is done online doing a series of simple tasks for various organizations worldwide. Gig work via applications refers to physical work tasks that the worker acquires through the application. These tasks are such as cleaning and transportation tasks. Usually, in the gig economy, employees receive payments after each task and the relationship between the employee and the customer is short term. Even though the gig economy has become more popular, and more and more people report getting income from gig work, most of the people use the gig economy to get only additional income. Only a small percentage of gig employees do gig work out of necessity. (Hirschi 2018) Parviainen et al. (2017b, p. 15) write that several sources speculate that digitalization will increase the fragmentation of work and that more and more people would get their livelihood from various sources instead of working only for one employer in a traditional way. The most likely self-employed are the least educated, but even the most highly educated have a relatively high probability of self-employment doing either gig work or acting as light entrepreneurs (Parviainen et al. 2017b, p. 15).

In manufacturing, there are many possibilities of what the future of work will look like since the industry has soaked up modern technologies for a long time now (Wellener et al. 2018). Automation and robotization will abolish some of the routine work in factories but it is argued that automation and robotization could also possibly increase the amount of lower-skilled routine tasks. This is possible because of smart assistance technologies could make the work more ergonomic as well as more customizable to the employee's needs. (Stock et al. 2018) Still, the general understanding seems to be that manufacturing jobs will decrease as digitalization proceeds (Birgün & Ulu 2020, p. 42; Hirsch-Kreinsen 2016). However, the digitalization of manufacturing will make the remaining manufacturing jobs safer since dangerous and heavy tasks can be done using robots (Stock et al. 2018). Bzhwen et al. (2019) found that even though robots can do dangerous and heavy tasks, some safety concerns rise from the close human-machine cooperation. Especially working with cobots which are robots designed to work alongside humans in the factories (Bzhwen et al. 2019). Furthermore, it is projected that so-called blue-collar work in factories will be more techno-centric and therefore, the work itself will be simplified (Waschull et al. 2022). Even though the work might get simpler, this transformation requires new skills from the blue-collar workers as well (Birgün & Ulu 2020, p. 42).

Compared to blue-collar work it is even less certain how so-called white-collar work will transform. According to Birgün & Ulu (2020, p.43), white-collar workers must know how to design, manage, and implement modern technologies to benefit the business. Depending on the white-collar job, others are easy to automate whereas others are extremely hard to automate. According to Andersson et al. (2016), simple office work such as invoicing and logical acceptance tasks are easy to automate but then expert roles that require creative problem-solving and cannot be broken down into logical rules, are the hardest jobs to automate. Waschull et al. (2022) suggest that white-collar work will become even more human-centric but what is the exact role of white-collar workers is heavily dependent on how the organization takes its' workforce into account when designing new implementations of digitalization. Where blue-collar work is predicted to become more techno-centric, white-collar work is more likely to become more socio-technical and the job opportunities for white-collar workers are predicted to enrich (Waschull et al. 2022). In other words, white-collar work is to become more challenging requiring a list of competencies. Skills complexity and skills requirements of engineers and technologists increase, while their task frequency gets lower (Maisiri et al. 2019).

In Table 1 the most impactful changes are summarized. In addition to Table 1, Kadir et al. (2019) state that problem-solving and IT skills will become a necessity for all workforces.

Table 1. *The effects of digitization on the labor market and work (Degryse 2016, modified)*

At risk to reduce	Likely to increase
Manual and repetitive tasks	Robotization & automation
Job positions	Jobs related to IT, new job titles
Mid-level skills jobs (computerization)	Efficiency
Clear division of work and free-time	Human-machine collaboration
Employees' control of their skills	Use of wearable and handheld devices
Trust between employees and employers	Virtual models
Confluence of competencies and competency needs	New forms of collaboration
Equal opportunities	Factories returning back to the origin countries
Salary development	Innovative products and services

Parviainen et al. (2017b, p. 34) say in addition to the changes mentioned in Table 1, it is expected that digitalization will increase inequality and wage differences, new forms of work such as micro-entrepreneurship and self-employment, and people having multiple simultaneous employment relationships in the future. Accenture (2017) states in their report that work will increasingly be broken into tasks or small projects that utilize an individual’s unique skills and knowledge areas since organizations outsource projects to individuals with relevant skills and knowledge. Digitalization will also modify the contents and tasks of work through sharing economy and automation of information work. Parviainen et al. (2017b, p. 34) say in their conclusions that professions that are estimated to be significantly affected by the progress of automation and digitalization are industrial work, salespeople, product presenters, and business economist. In turn, fields of "creators of digital technology" and "builders and users of digital platforms" will create new work (Parviainen et al. 2017b, p. 34).

Concerning digitalization, also sustainability and the green economy will generate new jobs. For example, International Labour Organization (ILO) (2018), estimates that twenty-four million new jobs would be created by 2030 only because of changes in energy production and use to achieve the 2°C goal while at the same time causing the loss of around 6 million jobs. In addition to green energy production, the United Nations Industrial Development Organization estimates in its report that a circular economy would create nearly six million jobs through recycling, reuse, and remanufacturing. To gain these estimated jobs, both OECD and non-OECD countries need re-skilling. (Auktor 2022, p. 16)

2.3 The effects of digitalization on skills and competencies

The challenge for many countries is to match the education of their people with current economic requirements (Wilson 2013, 102). Digitization is possible because of knowledge and competencies, but at the same time, new expertise needs arise because of digitalization. Since digitalization is changing the world so rapidly, learning and skill development has become a necessity in work-life. (Nousiainen 2020) According to Almeida et al. (2012), acquiring knowledge and skills increases human capital and is central to economic development. Educated employees have better employment opportunities, earn more, and have more stable and rewarding jobs. Some companies even complain that the lack of relevant skills limits their business growth. (Almeida et al. 2012, 14)

The situation becomes even more complex due to the diversity of skills needed at work. Moreover, empirical studies show that both core work and soft skills, and cognitive and non-cognitive skills determine an individual's employability and affect an individual’s earnings later in life. Also, technical skills are important for an individual’s employability. The emerging pattern shows that when economies evolve and expand, the demand for higher-level cognitive skills increases compared to the demand for physical work-related skills. (Almeida et al. 2012, 14)

Only a high-quality education system alone can no longer meet all training needs created by digitalization. Therefore, employees need skill development throughout life. In addition to traditional degrees, continuous learning is needed, which usually happens at work without a precise plan. (Nousiainen, 2020) While a lot of learning happens on the job, formal mid-career training programs are still needed (Wingard & Farrugia 2021, p. 27). Also, Parviainen et al. (2017b, p.11) highlight the importance of education in tackling skill needs. Training needs must be identified, and after identification, necessary training must be available for employees in various forms of training. A good basic education is emphasized and the importance of natural sciences such as mathematics and physics contribute to in-service training. This is because they create a good basis for learning new skills. (Parviainen et al., 2017b, p. 11)

ILO (2019, p. 19) defines skills as "*the knowledge, competence and experience needed to perform a specific task or job*". In a survey conducted by the World Economic Forum (WEF), in 2018, the majority of responding employers estimated that the skills related to job performance have changed significantly, and although the necessary skills vary from industry to industry, basic skills were thought to change significantly in the future as well. (WEF, 2018) Various experts have made empirical analyzes of employability using both direct and indirect methods. In 2017, Suleman summarized the researched employability skills to detect the exact employability skills and whether these skills can be specified in a skills catalog (Suleman 2017). Table 2 presents a modified version of Suleman's skills summary.

Table 2. *Summarized employability skills (Suleman 2017, modified)*

Skills found using direct methods	Skills found using indirect methods
Ability to learn	Self-management
Problem-solving	Communication
Using information	Numeracy
Communication, written and spoken	Teamwork
Teamwork	Analytical skills
Technical skills	Language skills
Cost and time management	Leadership skills
Critical thinking	General writing skills
Interpersonal skills	Socio-emotional skills
Real-life skills	Organizational skills
Application of knowledge	Professional skills
English language skills	Learning skills
Analytical skills	Presentation skills
IT skills	Ability to perform under pressure

In Table 2, also so-called foundation skills are represented. Foundation skills include skills such as numeracy, written and spoken communication, ability to build knowledge and skills together with the ability to solve and analyze problems. Foundation skills are the minimum level of competence that every graduate should have. (Oliver et al. 2014) Andersson et al. (2016, p. 73) say that the most important future work-life skills are:

- Cooperation and communication;
- Self-management, prioritization, and concentration;
- Experimentation, quick learning, and defining the learning goal;
- Entrepreneurial spirit as an attitude, and skill to strive for something new;
- Creativity, looking for alternatives, and lateral thinking.

According to Marr (2022), the most needed skills for the next 10 years are digital literacy, data literacy, critical thinking, creativity, collaboration, flexibility, leadership skills, time management, and curious and continuous learning. He says that the most important skill is to know how to navigate in a digital world at work and in everyday life and everyone should know how to use digital devices, applications, and software (Marr 2022). According to Zavyiboroda (2022) also professionals who are not in charge of software development directly still need to embrace programming and overall digital literacy, have hands-on data analysis and statistics skills within their area of responsibility and stretch towards computational and algorithmic thinking. Since data is now one of the most important tools to companies, individuals need to draw meaning from data. In addition to data interpretation, it is important to be able to evaluate the integrity and validity of data. (Marr 2022)

Collaboration and cooperation are done either with machines, artificial intelligence, and robots or with colleagues, customers, and suppliers Accenture, 2017, p. 9) and in various working modes such as hybrid work and full remote work (Marr 2022). In this diverse environment, it is crucial to effectively communicate and collaborate (Marr 2022). With good self-control, individuals can understand, control and adapt their emotions and behaviors in a team environment (Accenture, 2017, p. 9-12) and part of that is emotional intelligence that enables everyone to understand how their emotions affect others (Marr 2022). Also, self-control, self-motivation, and understanding of own strengths are important in today's work-life (Zavyiboroda 2022).

Self-management is also related to time management. The ability to manage one's time effectively is essential for your workplace performance, and it is about working smarter not necessarily harder. Good time management skills support employees' mental health too and it enables employees to create better work-life balance. (Marr 2022) Other valued interpersonal skills are mental flexibility, the ability to build relationships, and teamwork effectiveness are all required for effective collaboration, building relationships with others, and moving towards common goals (Zavyiboroda 2022).

In addition to self-management and self-leadership, leading others and embracing leadership are on-demand future skills. Leadership skills bring the best out of other people and enable them to thrive. Leadership skills are important to everyone because of concepts like distributed teams, increasing diversity, the gig economy, and more fluid organizational structures becoming more common. (Marr 2022) Leadership skills include skills such as role modeling, creating an inspiring vision, organizational awareness, ownership and decisiveness, grit and persistence, and especially the ability to cope with uncertainty (Zavyiboroda 2022).

Critical thinking emerges in various sources as an important skill for working life. Critical thinking includes the ability to understand structured problems, search for relevant information, logical reasoning, and master agile thinking (Zavyiboroda 2022) and it means analyzing issues and situations based on evidence rather than hearsay, personal opinions, or biases and questioning the validity of evidence (Marr 2022). Critical thinking creates the basis for understanding structured problems (Zavyiboroda 2022). It allows employees to use logic and reasoning to pinpoint the strengths and weaknesses of alternative solutions, approaches to problems, and conclusions that are all important for problem-solving (Islam 2022).

In addition to soft skills, technical skills are no longer solely the purview of experts, and digital and emerging technology skills will remain or become critically important to everyone (Accenture 2017). Maisiri et al. (2019) list technology skills, programming skills, and digital abilities as technical skills. According to Islam (2022), programming skills are the technical skills that allow employees to create and develop software. Other technology skills include, for example, the ability to collaborate with robots, IoT, and other Industry 4.0 technologies such as connectivity and cloud computing. Technical skills include also clever design skills such as simulation. Simulation can also be part of programming skills, which include calculation and coding. Digital abilities emphasize data expertise, cyber security, and cloud service expertise. (Maisiri et al. 2019) Islam (2022) also lists data visualization and data interpretation skills as important technical skills. Skills related to Industry 4.0 technologies such as understanding cloud services and knowledge of IoT are directly related to employability (Islam 2022).

Besides cloud computing and cyber security, other important technology skills include AI and machine learning (ML), Big Data analytics, VR and AR, blockchain technology, and data security, 3D printing and user experience (UX) (Zavyiboroda 2022; Heideman Lassen & Waehrens 2021). Especially cloud computing skills are on-demand due to companies changing their server infrastructure to cloud solutions and many AI and ML services are provided through cloud platforms (Zavyiboroda 2022). AI and ML technologies and tools help organizations deliver more relevant, personalized, and innovative products and services and therefore, these skills are on-demand (Zavyiboroda 2022). Overall, the need for technology, engineering, and math skills is significantly growing (Wingard & Farrugia 2021, p. 19) and more and more people need to have fluent advanced technology skills to create, implement and maintain modern technologies in the workplace (Wingard & Farrugia 2021, p. 22).

Auktor (2020, p. 15) lists four groups of work tasks that are relevant for green occupations in the future. The first group, engineering and technical skills, has hard skills such as design, construction and assessment of technology that are important for renewable energy design and energy-saving R&D projects. Engineers and technicians usually master these skills. The second group, science skills, includes knowledge on for example physics and biology that are crucial for innovation activities. These skills are necessary for environmental scientists and materials scientists, for instance. The third group, operation management skills, has skills related to organizational change, lean production, life-cycle management, and cooperation with external stakeholders to support sustainable activities. These skills are important for example for sales engineers, climate change analysts and sustainability specialists. The fourth and final group, monitoring skills, are skills related to technical and legal aspects and are different from engineering and science skills. This group has skills such as observance of technical criteria and legal standards and are usually relevant for environmental compliance inspectors and officers. (Auktor 2020, p. 15)

Closely related to green skills are electrification skills. Especially, governments planning to ban diesel and petrol cars has pressured car industry to rapidly develop electrical cars. Electrification is however not only effecting the car industry but many other industries are trying to electrify their products or operations. Therefore, up-skilling and re-skilling is in dire need. (Constantinou 2021)

In a report published about electrification skills need in the UK the most important skills revolved around supply chain and logistics, circular economy and sustainability, and technology development and deployment. There is a big need for skills in battery manufacturing, and power electronics, motors and drive. Specialist such as design engineers, electronics technicians, and system engineers are really in demand. (Harper et al. 2021)

Overall, there is a clear need for both soft and technical skills in the future. The most relevant soft skills seem to be ability to learn, critical thinking, teamwork skills, and leadership skills. On the technical skill side data is clearly the most mentioned skills. It includes data visualization, data analytics, data interpretation, and the ability to make decisions based on data. Other technical skills revolved around industry 4.0 technologies. There is a need for machine learning, cloud computing and cyber security specialist to mention a few examples. What is certain is that re-skilling is needed since new technologies are emerging frequently and the ways of working are changing.

3. NATURAL LANGUAGE PROCESSING

This chapter introduces the used research method - natural language processing. It has become exceedingly popular since with NLP computers can represent and analyze human language computationally and it can be used in various fields. NLP draws from different theories and techniques to communicate using natural language with computers. (Khurana et al. 2022)

3.1 The evolution of NLP: Then and now

NLP roots back to the 1940s, when formal language theory started to evolve (Beysolow 2018) during the germination period until 1956. During the germination period, Alan Turning produced a concept called "Turning Machine". It was the theoretical origin of the modern computer, and it provided the background for machine translation and later NLP. Due to the need for machine translation, the basic research of NLP was done during the germination period. In the early 1950s, Kleene studied finite automata and regular expressions, and in 1956, Chomsky applied context-free grammar to NLP. This led directly to the creation of two rule-based and probabilistic NLP techniques. In 1956 the invention of AI boosted the development of NLP since researchers started to gradually combine these two techniques together expanding the social application of NLP. This led to the new development period of NLP; the rapid development period from 1957 to 1970. (Kang et al. 2020)

At the beginning of the rapid development period, signature work influenced by AI began, and research on rule-based methods and probabilistic methods progressed quickly. Probabilistic methods adopted statistical research methods and made great progress, but AI research funding was focused on logic. However, in 1967 a concept of cognitive psychology was introduced which led to linking NLP with human cognition. (Kang et al. 2020) The low-speed development period lasted from 1971 to 1993 (Kang et al. 2020) though in the 1980s computational grammar theory became a highly active research area. It linked to logic of meaning and knowledge and the ability to deal with user beliefs and intentions as well as functions such as emphases and themes. (Khurana et al. 2022)

In the 1990s, during the recovery period, practical resources, grammar, and parsers tools started to become available (Khunara et al. 2022). Also, the increase in memory and speed of computers and the commercialization of the Internet boosted the development of speech and language processing since it made information retrieval facile. (Kang et al. 2020) NLP was especially used for information extraction from data and automatic summarizing in the late 1990s. From the 2000s onwards development of NLP took huge steps forward. (Khunara et al. 2022) In the early 2000s the first neural language model was introduced (Kang et al. 2020). The feed-forward neural network determined the probability of the next word using *n* amount of previous words. Also, a multi-task learning model was introduced that used two convolutional models with max pooling to mark part-of-speech and named entity recognition. A word embedding process was proposed that used dense vector representation of text. After the progress of word embedding, the field of NLP introduced neural networks where variable length input is taken for further processing. The use of neural networks has played a particularly important role in NLP. It paved the way for the use of various neural networks in NLP such as convolutional neural networks (CNN) and recurrent neural networks (RNN). (Khunara et al. 2022)

Nowadays NLP can be described as a multidisciplinary effort to process, understand, or produce natural language, such as English or Finnish, mechanically (Deng & Liu 2018, p. 1). Oxford University Press's dictionary and encyclopedia defines NLP as follows:

> *"Natural language processing: abbr.: NLP; the computational analysis and interpretation of human language. NLP is used in software that provides automatic translations of text from one language to another, in robotic systems that use human-language-type commands, and in text-mining tools (e.g. to provide summaries or abstracts of large volumes of text)."* (Cammack et al., 2006)

NLP combines computer science, AI, and cognitive psychology, and it is concerned as an interaction between natural language and computers (Deng & Liu 2018). NLP can be divided into two research directions: natural language understanding (NLU) and natural language generation (NLG). NLU allows machines to understand natural language extracting concepts, emotions, keywords, and meaning whereas NLG allows machines to produce understandable natural language from text, audio, video, or other types of data sources. (Kang et al. 2020) Some think of NLP as a research area of information technology and machine linguistics, where applications that utilize natural language structures such as a word, a sentence and a document are built (Cohen & Demner-Fushman 2014 p.1-2) while others think of NLP as a subcategory of machine learning (Deng & Liu 2018).

Some applications of NLP are machine translation, text categorization, spam filtering, information extraction, summarization, and dialogue system (Khurana et al. 2022). NLP has many commercial and everyday applications such as Google Translate, Gmail spam filtering, and Apple's Siri voice assistant (Khurana et al. 2022) and it is widely used in chatbots, information collection, and machine summaries to support decision-making. NLP research is influenced by a wide range of different research fields, like machine learning, cognitive sciences, linguistics, and information technology (Deng & Liu 2018 p.1).

3.2 Natural language processing & machine learning

This chapter explains the basic steps of machine learning used with NLP. A rough presentation of the process is shown in Figure 1.

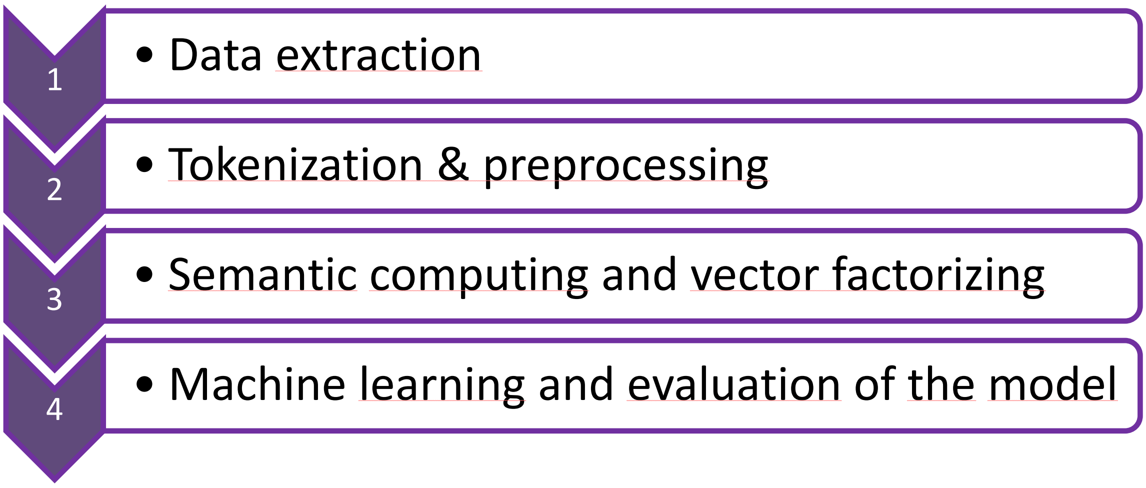

Figure 1. *The machine learning process of natural language processing*

This chapter presents all the topics mentioned in Figure 1.

3.2.1 Tokenization & preprocessing

Programming languages use tokenization since their syntax is clearly defined and characters cannot have several different meanings unlike natural language. Natural language is classified as unstructured data even though it is based on some rules and internal structures. In order to make natural language understandable to computers, it needs to be tokenized. How difficult tokenization is depends on many natural language factors. For example, languages such as English that delimiters words with space, tokenization is easier than for languages that do not use space to delimiter words. (Indurkhya et al. 2010, p. 15)

Tokenization essentially means text segmentation (Uysal & Gunal 2014). Corpus consists of independent entities called documents that can be broken down into sentences. The sentences are formed from a set of words (or tokens) that can be further broken down into single letters or numbers. (Martinez 2010) This hierarchy is presented in Figure 2. The challenge of tokenization many times comes from the use of punctuation, like periods and commas. They can serve many purposes in a sentence and in a piece of text. Periods, for example, can be used in abbreviations, numbers or to end a sentence. (Indurkhya et al. 2010, p. 16) It is also good to note, that one word appearing many times will create many corresponding tokens. Due to all these details, tokenization often requires a lot of domain knowledge of the topic since punctuation can have many meanings depending on the context. (Aggarwal 2018 pp. 6)

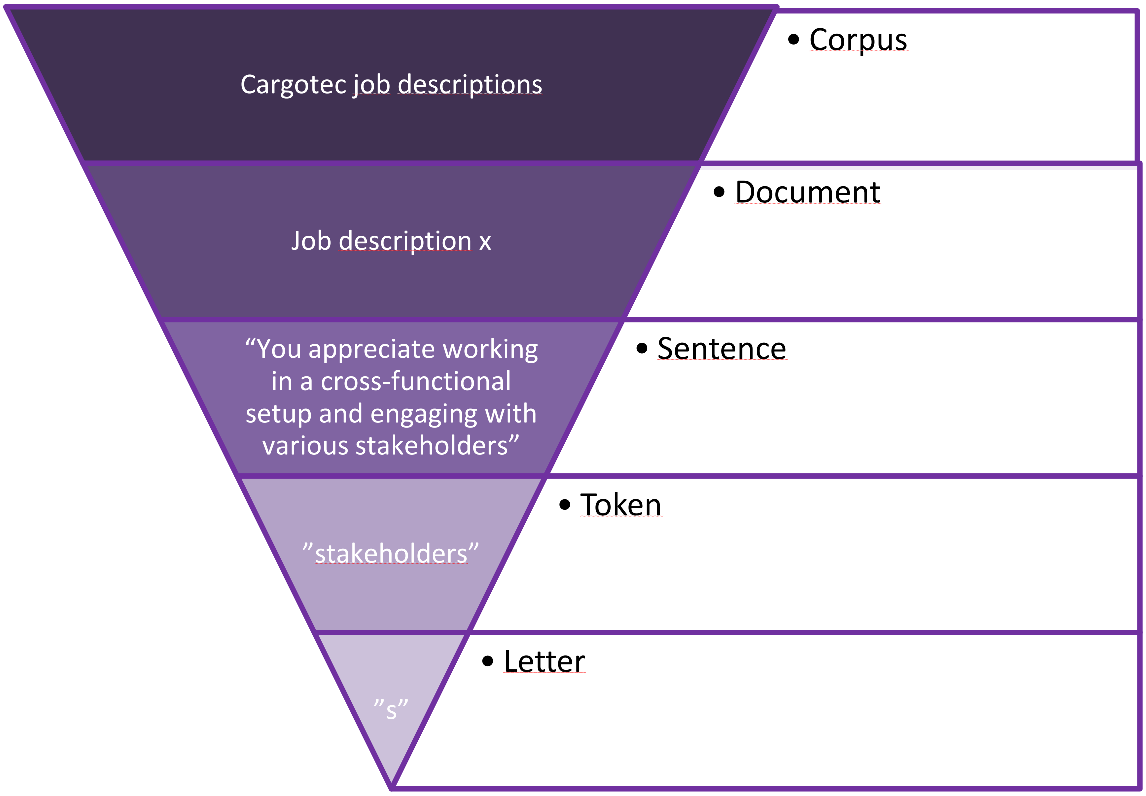

Figure 2. Hierarchy of natural language

Text preprocessing aims to clean the raw data file, remove all the irrelevant parts and keep only the meaningful word units. These irrelevant parts might be special characters, filler words, or stopwords (i.e. *the*). Preprocessing is a crucial step in natural language processing since the preprocessed data is used in every single step after preprocessing. (Indurkhya et al. 2010, p. 9) The text preprocessing has four phases:

1. Text extraction
2. Stop-word removal
3. Stemming, case-folding, and punctuation
4. Frequency-based normalization.

In the first phase of preprocessing, the text is extracted, for example, from the Internet. In the second phase words that have little to no semantic meaning, stop-words, are removed. These words can be pronouns, articles, and prepositions. In the third phase, stemming, words are changed into their stem form and for example, words *swim* and *swimming* are given only one representative, *swim*. In this stage, capital letters might be removed as well depending on the context, as well as punctuation. (Aggarwal 2018 pp. 6) Another tactic is lemmatization, where words are changed into their base form. Lemmatization is more complex than stemming. In lemmatization, the goal is to find the basic form of the word. (Hu & Liu 2012 s.389) In the fourth and last step of preprocessing, less frequently occurring words are given more weight than frequently occurring words. This is done due to less frequently occurring words are usually more distinct and important for the text and function as feature separators. This technique is also known as inverse document frequency (IDF). (Aggarwal 2018 pp. 6) All these preprocessing steps assure that the used natural language is transformed into a summarized computable form that machines can process. How effective these methods are, is dependent on how infections the used language is as in how much the words are separated with spaces (Aggarwal 2018, p. 24).

3.2.2 Machine learning

Once data is collected and preprocessed it is time for machine learning. Machine learning is a subcategory of AI, an algorithm that learns from data and has roots in statistics. ML makes AI adaptive and able to react to changes. (Pietikäinen & Silvén 2019, p. 7, 68) Machine learning can learn, for example from text, features, and properties that allow it to predict, classify or create new values and it can improve its performance or knowledge during its task (Flach 2012, p. 3). As the name suggests, machine learning simulates human learning mechanically with a continuous iterative process.

Machine learning can be divided into three categories: unsupervised learning, supervised learning, and reinforcement learning. Supervised learning can be further divided into classification and regression, unsupervised learning is used for clustering and reinforcement learning is used for decision-making. (Pietikäinen & Silvén 2019, p. 69, 71, 91, 95) This categorization is presented in Figure 3.

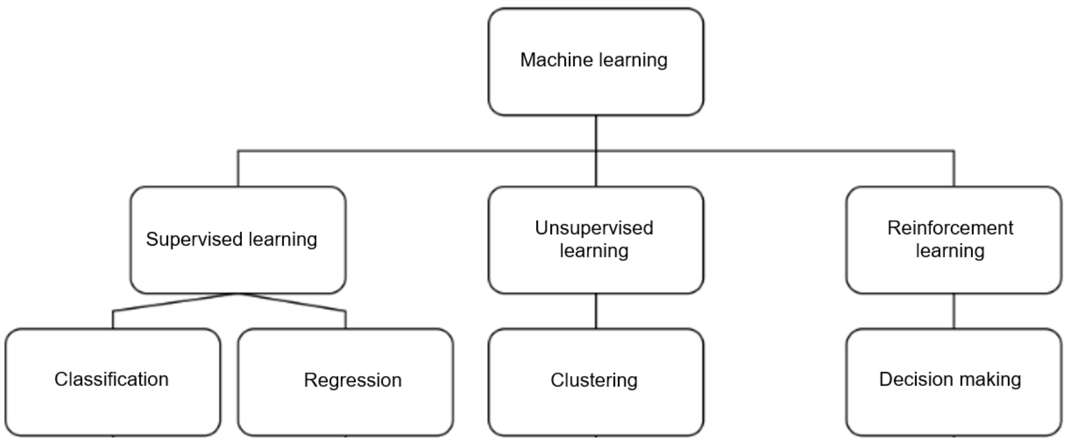

Figure 3. *Categorization of machine learning*

Supervised learning aims to teach the machine learning model with labeled data points that are divided into classes or under labels. Both the classes/labels and the inputs are known in the learning phase, but after learning the aim is to classify new inputs under the known classes/labels. Training data can be thought of as instructions for the machine on how to deal with specific types of features. In unsupervised learning, the algorithm does not get labeled data, but by looking for similarities from the input it tries to build a description of the data structures, and cluster similar inputs close to each other. This method is called clustering and it does not need a teacher who provides labeled data for it. In reinforcement learning, the algorithm examines its operating environment, makes decisions, and gets feedback from those decisions. It can receive positive and negative feedback, but the aim is to find a solution that results in as much positive feedback as possible. With reinforcement learning, it is possible to implement forgetting which is not possible with other learning techniques. Reinforcement learning also allows to continue searching for even better solutions than the first solution reached. Reinforcement learning usually learns the hypothesis without instructions or labeled data. (Pietikäinen & Silvén 2019, p. 69) In addition to these three categories, there is also semi-supervised learning that combines supervised and unsupervised learning (Pietikäinen & Silvén 2019, p. 70). Semi-supervised learning uses both labeled and unlabeled data when forming the model and is a popular method when there is a limited amount of labeled data available.

Validation is an important part of developing a machine learning application. In addition to training data and testing data, there needs to be validation data. Validation data ensures that the learned model can do predictions with data that was not part of the training or testing phase. It ensures the ability of the model to make reliable predictions. This validation process is called generalization. (Pietikäinen & Silvén 2019, p. 70-71) It is important to find a model that does not follow the characteristics of the training data precisely but the useful structures from the data are recognized. If the model follows the training data too precisely, it forms a very complex model that causes overfitting. The opposite of overfitting is an overly general model that underfits data points. An underfitting model is generalizable but does not learn useful features from the data. (Müller & Guido 2017) The goal of a machine learning model is to learn the features of data but not to copy them directly.

3.2.3 Semantic computing

Natural language has rules and structures, known as grammar, even though it is classified as unstructured data. These structures and rules compose the syntax of the language through which the meaning of the language is created. Semantics examine the meaning of words, sentences, and thoughts. (Martinez 2010) Some sentences might have the same meaning but have different words in them. For example, sentences like *“I like dogs”* and *“I love puppies”* have similar meanings, but only one common word, which is likely to be removed in preprocessing. How can computers understand that these vastly different-looking sentences hold the same meaning? It is unequivocal, and the complexity of natural languages creates challenges.

Semantic computing is used to classify the text based on semantic factors. The semantic study can be divided into two categories: lexical semantics which focuses on the meaning of words, and supralexical semantics which focuses on the meaning of sentences and phrases. However, these two study areas are not separated from each other but rather interact with each other. Also, the semantic similarity is divided into two categories: homonymy and polysemy. Homonymy means words that have different meanings but look like or sound like the same. For example, the word *bark* can mean either dog barking or the bark of a tree. Polysemy means words that have various related meanings, i.e., *play* can be used as playing with your dog as well as playing sports. Both polysemy and homonymy create challenges for NLP. Also, an individual's background, cultural context, and level of education affect the interpretation. (Indurkhya et al. 2010, p. 94)

Methods that measure the semantic likeness of words are divided into corpus-based methods and knowledge-based methods. Knowledge-based methods usually identify the semantic similarity of two words using knowledge sources such as ontologies whereas corpus-based methods need a large amount of data, and those rely on the idea that similar words occur in similar circumstances. (Chandrasekaran & Mago 2021)

Because corpus-based methods can draw from several large corpora, they tend to cover better the vocabulary than knowledge-based approaches, which can only cover the concepts contained in the given data graph (Zhu & Iglesias 2017). Since corpus-based methods are based on statistics, words need to be changed to vector format. Breaking down the corpora into smaller sections makes it easier to separate features into vectors.

Vectors can be formed using the IDF technique that creates bag-of-words (BOW) vectors. In bag-of-words vectors, the word order does not matter and they can be in different order in the vectors, than they are in the text. (Chandrasekaran & Mago 2021) BOW saves the words with their term frequency number to a vector. For example, a sentence "Dogs are cuter than cats" after preprocessing would be presented in a vector form as {'dog':1, 'cute':1, 'than':1. 'cat':1}. This way the raw text data is changed into numeric data, which is understandable for computers. Using vectors makes it easier to compare documents and determine their similarity. Similarity can be calculated using cosine similarity. (Aggarwal 2018, p. 27)

However, this TF-IDF format is insufficient since it relies so much on using the same words in different documents to determine the similarity. Even though the BOW method reduces the information from the original raw data and the data cannot be restored to its original form after vectorizing, the BOW method can still be used in simple binary classification problems. Corpus-based methods are not dependent on specific language and do not need an external information system unlike knowledge-based methods. (Aggarwal 2018, p. 305-306) Corpus-based methods work well when there is a large amount of text available that can form a large enough corpus.

In addition to knowledge-based methods and corpus-based methods, deep learning methods have become popular as well. Deep learning methods are based on multi-layered neural networks. With deep learning methods features are recognized in the layers of the neural network, and the data is transformed. After this the transformed representation is fed hierarchically to the next layer of the neural network, which then transmits the semantic information. Using neural networks for preprocessing can use the continuous bag of words (CBOW) method. CBOW separates the words into a vector by evaluating each word based on its surrounding context. (Altınel & Ganiz 2018)

3.3 Corpus-based methods

This chapter focuses on corpus-based semantic-similarity methods and introduces a few popular word embeddings techniques and semantic-similarity methods. Word embeddings are used to measure the semantic similarity of texts in different languages by mapping the word embeddings of one language into another vector space.

These word embeddings use approaches like neural networks and co-occurrence matrix. One of the most popular ones is word2vec. It was originally developed from Google News dataset. It is a neural network model used to generate a distributed vector representation of words based on the underlying corpus. (Chandrasekaran & Mago 2021) It is popular because it is relatively easy to use and access (Church, 2017). It can use either CBOW or Skip-gram method. Opposite to CBOW, Skip-gram predicts the words surroundings. The architecture of word2vec has an input layer, one hidden layer and an output layer. Word2vec is good at retaining the contextual similarity between words in a vector format, but it only uses word occurrences that are found from the corpus. (Chandrasekaran & Mago 2021)

FastText is able to use internal information of a word. It is an open-source library for scalable solutions for text vector representations and classification. It was originally launched by Facebook AI Research. It is a fast model to train and independent on the language. (Bojanowski et al. 2016) FastText uses Skip-gram methods and breaks the word into parts to examine the structure of the word (Chandrasekaran & Mago 2021).

Semantic similarity methods are easy to use with a lot of different languages, but they do not take account the meaning of words. One of the methods is called Latent Semantic Analysis (LSA). It is a method to measure semantic similarity of words and it is the most popular method. LSA forms a co-occurrence matrix where words are presented on rows, paragraphs on columns and the cell store the word counts. This makes it easier to estimate the semantic similarity since it also uses information on the whole paragraphs. Dimensional reduction is done with Singular Value Decomposition which gives the result based on three matrices. It retains the similarity structure of words by reducing only the number of columns. (Chandrasekaran & Mago 2021)

Another method is dependency-based models. Dependency-based models find the meaning of a given word or phrase by using the word's neighbors in each window. They initially parse the corpus based on its distribution using inductive dependency parsing and forms a vector by adding each window across the position whose root word is the word to be considered and when the vector is done, the semantic similarity is calculated using the cosine similarity between the vectors. (Chandrasekaran & Mago 2021)

3.4 Neural networks & self-organizing map

Neural networks are part of the corpus-based methods such as CBOW (Chandrasekaran & Mago 2021) but because of the attention that they have received neural networks deserve their own chapter. Artificial neural networks are based on data driven connectionist information processing and they have simple teachable computing elements that are connected to each other. Simple one-layer structure is presented in Figure 4. Neural networks are widely used to automatically recognize for example speech and make predictions. Their teaching requires a lot of data in the teaching phase and yet there are not able to justify their solutions. (Pietikäinen & Silvén 2019, p. 6) The first layer of artificial neural network is input layer followed by *n* number of hidden layers and lastly an output layer (Nelson & Illingworth 1991, p.46).

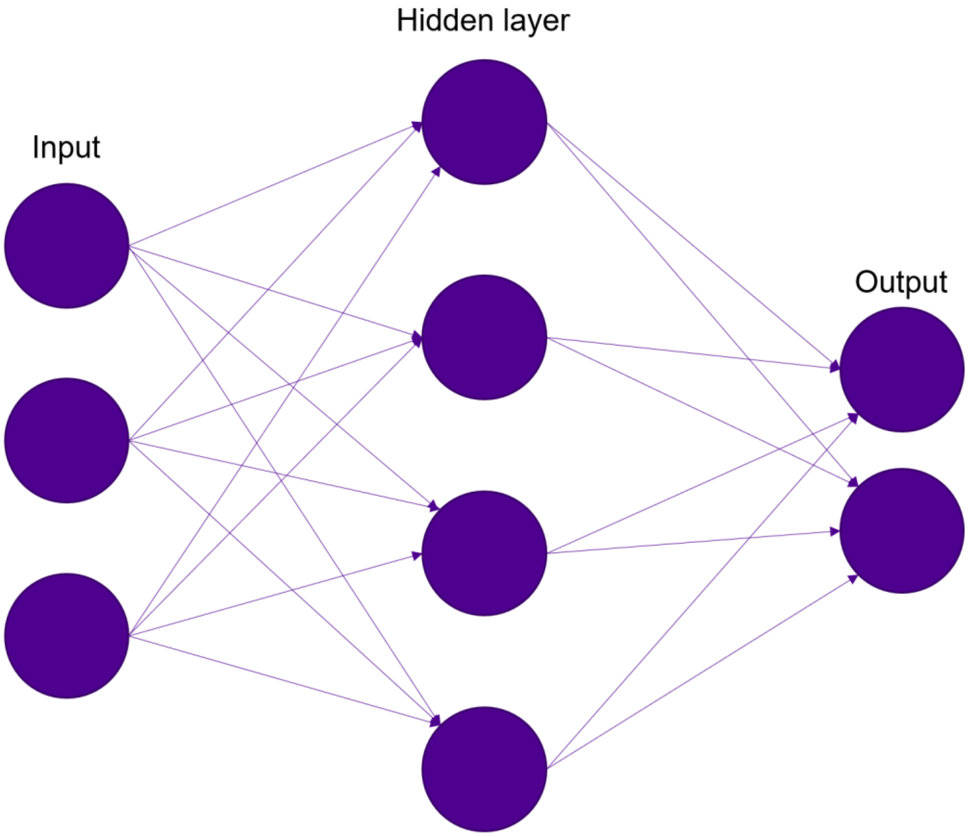

Figure 4. *Structure of neural network*

Each neuron of the neural network has its own adjustable weight. When a signal reaches a neuron, it moves to the next neuron and if the sum of different signals coming to the next neuron is high enough, also the next neuron is activated. Decisions are then made based on these weighted sums. (Nelson & Illingworth 1991, p.36-39) Like other machine learning techniques, different neural networks can be build using either supervised or unsupervised learning.

Like other machine learning methods, also neural networks can learn either supervised or unsupervised. One of the unsupervised learning methods is called self-organizing map (SOM). Like other neural networks, SOM needs a large amount of data to be trained but during the internet-era, data is relatively easy to get. Another benefit of SOM is that it provides simultaneous visualization and clustering based on topographic map formation. SOM is an excellent option for example processing scientific articles, since they are hard to transform into traditional Euclidean feature vectors without losing valuable information. (Laaksonen & Honkela 2011, p. 1-2)

SOM learning can be explained as follows. Each node map contains a model vector, and the node map learns by processing one input at a time iteratively. The best matching unit (BMU) is calculated for the input sample and the model vector of the BMU is adjusted towards the input vector. The model vectors of the neighboring nodes of the BMU are also adjusted per input sample, but slightly less than the model vector of the BMU. After iterative rounds of training, each region of the map or cluster is specialized to represent some features of the input data. The training is usually done in two phases where in the first phase, neighborhood is kept big, and the learning effect is high. The second phase focuses on fine tuning the results with smaller learning effect and neighborhood. When the learning phase is done, SOM can place samples to the map that it created by calculating again the BMU for the input and the place of the BMU is then also the place of the input sample. At first, the map is created during the learning process and then after that the clustering is added. (Saarikoski 2014) In the Figure 5, an example of SOM clustering is presented.

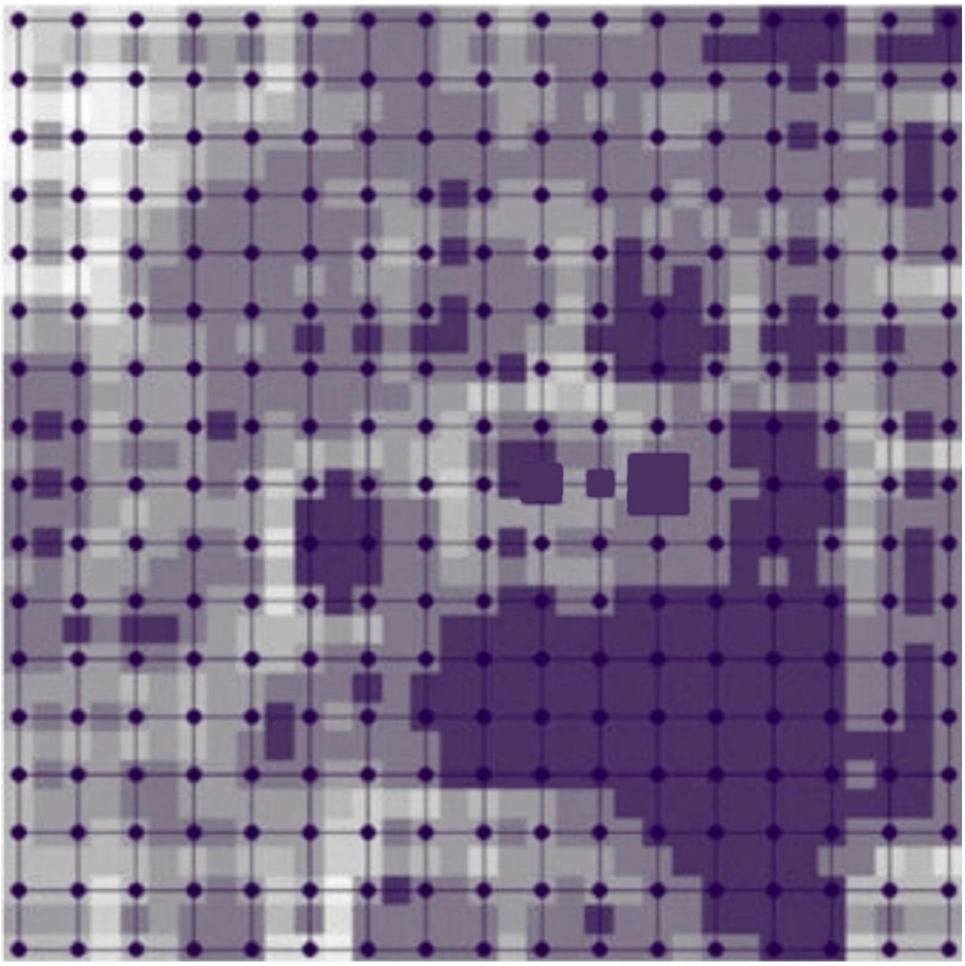

Figure 5. *Example of SOM clusters (Saarikoski 2014, modified)*

4. IMPLEMENTATION

The practical part of this theses was done together with a Finnish AI company Headai and a Finnish cargo handling portfolio company Cargotec. The aim of the practical part was to analyze Cargotec's job descriptions and form a skill mind map of Cargotec's current skills using Headai's AI solution. In addition to forming a skill mind map from Cargotec's data, three other skill mind maps were done from different industries and then compared to Cargotec's skill mind map. Since Headai is commercialized business and their solution is fully developed by them, this thesis will not go into details explaining the exact functions of Headai's application. However, some basic functionalities and principles will be discussed.

4.1 Short company introductions

Cargotec is a Finnish company specialized in freight handling solutions. As a portfolio company, Cargotec operates through three subsidiary companies (Kalmar, Hiab, and MacGregor) that focus on automatic terminal and cargo solutions, sea transport and offshore operations. Kalmar offers cargo handling machinery for ports and terminals and offers maintenance services for its customers. Hiab is focused on providing products and maintenance services for ground transportation and MacGregor operates at sea suppling cargo handling products for ships and maintenance services for its customers. Cargotec operates worldwide in over 100 countries, and it has around 12 000 employees. Recent years, Cargotec has put efforts on developing an eco-portfolio to lower its customers carbon footprint in their operations. Cargotec's goal is to become a global leader in sustainable cargo flow. Cargotec aims to achieve this goal utilizing digitalization and its modern technologies and data analytics.

Headai is a Finnish AI-enabler company founded in 2015 that focuses on analyzing cognitive texts automatically. It uses machine learning algorithms and natural language processing technology to do an analysis from text data. Headai's solution is fully self-developed. After a long research and development period, Headai has moved to production phase of their solution. Headai helps other organizations to analyze their text data and to ease their decision-making by providing them fact-based information. Headai's solution can be used to discover the skill trends on the current job market, make future market predictions and many other skill related features as well. Headai's goal is to be the largest cognitive text analytics platform in the whole world by 2029.

4.2 Input data

This chapter gives an overview of the data that was used in the implementation phase and how the data was collected. There were two types of data given to the AI solution. First, Cargotec's internal data was posted to the server to be processed and then data from various sources from the Internet was processed on Headai's platform.

4.2.1 Cargotec data

From Cargotec over 1500 actual job descriptions documents formed the input corpus. Job description is a document that contains the job title, a description of the role and tasks, and a list of necessary competencies and skills that are needed to perform in the role. It is comparable with a job advertisement but often it includes more insights on the skill requirements of the role. The job descriptions are usually done by human resources (HR) business partners in cooperation with the hiring managers. Job advertisements are made based on the job descriptions and recruiters can use job descriptions as reference sheets when looking for qualified candidates for job roles.

There were two limitations for the research material:

- Job description had to be from 2017 or newer;
- Only job descriptions written in English were used.

The dataset included job descriptions both of white- and blue-collar jobs but due to Cargotec's employee structure, there were more white-collar job descriptions than blue-collar job descriptions. All kinds of roles were included from engineering to admin roles but due to Cargotec being an industrial company, it naturally has more engineering roles than, for example, finance related roles. Even though each of the subsidiaries has their own products there is a quite significant similarity between skill needs of each of these business areas. There were around 450 job descriptions from Kalmar and Hiab, 500 from MacGregor and a bit over 130 job descriptions from Cargotec Group as seen in the Figure 6.

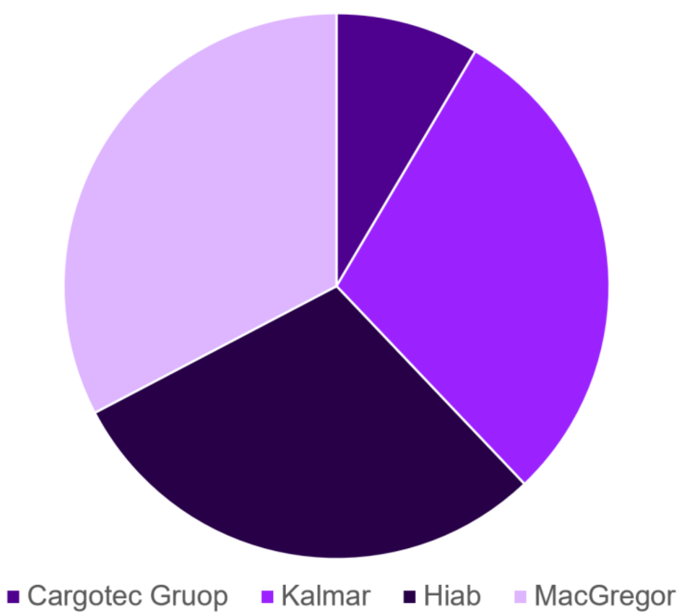

Figure 6. *Distribution of Cargotec's data between the business areas*

The research material was collected from several sources, and it required quite a lot of manual work. Due to Cargotec's organizational structure, job descriptions were not stored in the same place or in similar ways. The data was requested from each of the subsidiary's HR professionals and then collected into a same place. Some preprocessing was done even before entering the corpus to the AI solution. All line breaks and special characters except periods were removed.

Original idea was to web scrape Cargotec's employee profiles from LinkedIn. LinkedIn profiles usually contain more information on the individual's skill set whereas a job description describes only the necessary and nice-to-have skills for a specific role. However, LinkedIn prohibits data scraping from its domain and therefore this was not possible to do. Employees' profile information would have provided a different angle to the problem than job descriptions but the results from job description were still very useful.

4.2.2 Comparison data

The comparison data that describes the skill needs of the job market currently, was gathered from several job search sites. This part is done with Headai's software that automatically searches the web for content and the solution has read over 25 million job advertisements from selected locations all over the world. In addition to job advertisements, Headai's solution has also read hundreds of thousands university curriculums, millions of news and blog articles, and millions of science articles. The analysis was done during August in 2022 and only 2022 data was used.

Instead of specifying which data sources or locations were used, specific industrial clusters were used instead. These clusters contain job advertisements published from these sectors. Headai has five industrial cluster made, but for this research only three of them were used: machine & metal industry, electronics & electrification industry, and information technology (IT) industry. These clusters were chosen due to having similarities compared to Cargotec's operations. Cargotec could be categorized as part of the machine & metal industry but due to its electrification and digitalization ambitions, electronics & electrical industry and IT industry were included for the comparison as well.

4.3 The AI solution

This chapter is an introduction to the Headai's AI solution. At first, the basic principles of the solution are presented and then and how the solution was used.

4.3.1 How does it work?

Headai's AI solution uses machine learning and natural language processing technology that allows Headai to do an automated qualitative analysis of any text data. They combine cognitive psychology, semantic computing, reinforcement learning, and self-organized learning. The algorithm can automatically detect new concepts and add them to the already existing ontology. The ontology is compatible with other major ontologies, and it works as a dynamic language model for words, semantics, and meanings.

The dynamic ontology is based on self-organizing maps that has been taught unsupervised. It mimics the way humans learn. First, it learns the work context through common unstructured content and human teaching. The content given to the algorithm can be for example text documents or graphs. The learning process can be divided into two steps. In step one, it learns the basic semantics of the work context relationships unsupervised. In step two, the process applies reinforcement learning where the user gives feedback to it based on its decisions and performance. Because the learning process is done with various contents, the algorithm can do many different tasks related to skills and keeps the algorithm unbiased. With self-organized maps the words and semantics are learned unsupervised which allows the algorithm to be dynamic. and up to date. It has been learning using millions of job advertisements, scientific articles, reports, and curriculums which enables the algorithm to draw conclusions and supplement missing information.

4.3.2 How is it used?

The solution can be reached using Headai's application programming interface (API) that uses token-based authentication and allows users to protect their Uniform Resource Locator (URL) requests. Headai's API provides a clear understanding of source datasets by calculating and presenting concept mapping, concept relevance, and their relationships. This information can give insights into real-time skill predictions, find talent matches, or detect skill gaps. In this research two different APIs were used: BuildKnowledgeGraph and TextToMindMap. The API calls were made using Postman which is an API platform for building and using APIs.

When used, the BuildKnowledgeGraph reads a large amount of text according to the given parameters. For this API mandatory parameters are token key, ontology and output, and optimal parameters include language, dataset, cluster name, country, city, word type, noise list and search year, month, and day. In Figure 7 there is an example API call and all the parameters that were used in this research when using this API.

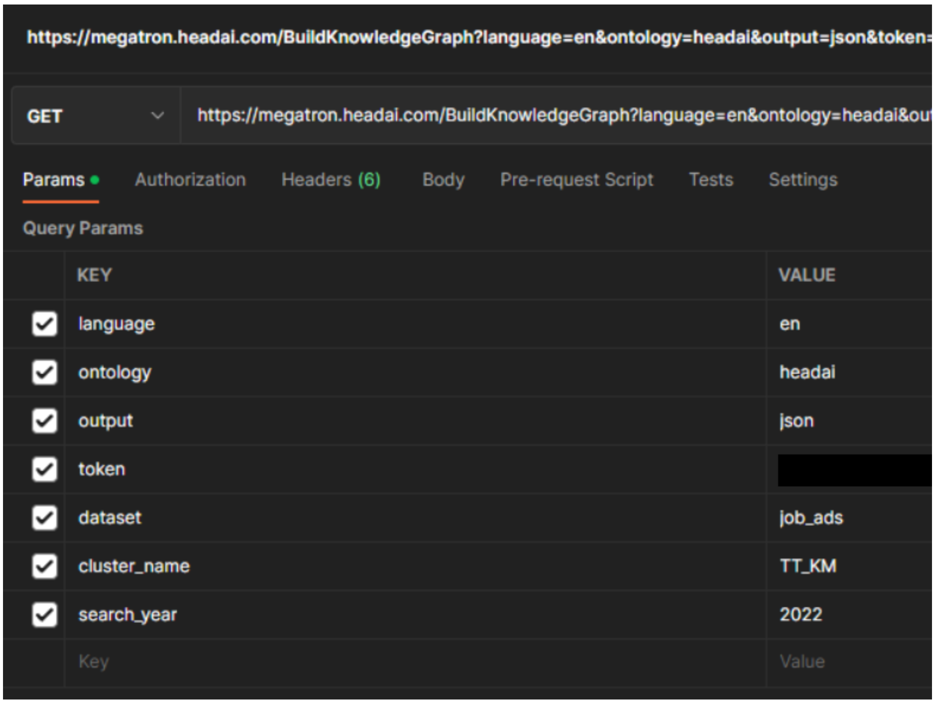

Figure 7. *An example of BuildKnowledgeGraph API call*

For this research, the only used dataset was job advertisements. These job advertisements have been gathered into the database using web scraping. After collecting the job advertisements, the texts are preprocessed, and the skill-related words are taken out from the scraped data with NLP. AI must find the semantic meaning of these skills-related words to compare the words with each other and combine words with similar meaning. Then each skill-related word is given value and weight. Value is the number how many times the word appeared in the dataset, and weight is a number from 1 to 5 given based on how descriptive the word is. The more descriptive the word is the higher the weight is as well. The relevancy of a word is calculated by multiplying the value with the weight. The word, its value and weight are given in the output link in a JSON format. JSON stands for JavaScript Object Notation and it is a lightweight data-interchange format that both humans and machines can process easily. The JSON link can then be used in Headai's data visualization tools for easier data interpretation.

The TextToMindMap API works using the same principles but there are few differences. This API works with users' own data instead of retrieving the data from different sources. Since the input is users' own text data, the API call has to be made using POST method instead of GET method. With POST, the input data is sent to the server for processing and analyzing. Figure 8 shows an example of TextToMindMap API call.

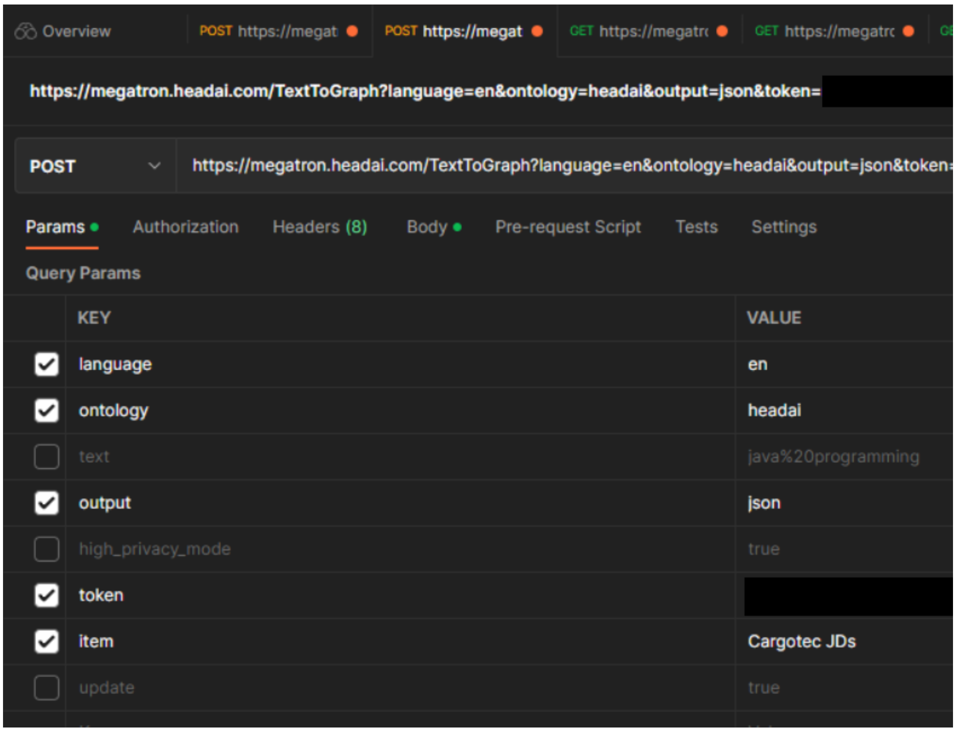

Figure 8. *An example of TextToMindMap API call*

From all the API presented above users get output in a JSON file format. These JSON files can be used in Headai visualizers to build the mind maps that represent the skills and their relations.

5. RESULTS

In this chapter the results of the thesis are presented. Skill mind maps were made for Cargotec, manufacturing & metal industry, electronics & electrical industry, and IT industry. Cargotec's skill mind map was made with the *TextToMindMap* API, and all the other skill mind map were made using the *BuildKnowledgeGraph* API. The first subchapter explains how the visualizations are read, and then each of the mind maps is presented.

5.1 How to read the skill mind maps?

The skill mind maps were done using Headai's visualization tool. The visualizations can be saved as pictures but viewing them in a browser gives the opportunity to utilize the various features of the visualization tool. When presenting the results, some pictures are screen shots from the browser version and hence, it is important for the reader to understands these features described below. The JSON links contain all the needed information to make the visualizations.

The map shows the keywords and the words that are connected to them close together. In the middle of the map are the concepts that were the most common and closer to the edges of the mind map are the least common concepts found from the input data. The skill words vary in color intensity as well. The color is related to relevancy of the word and the darker the color the more relevant the word is. Neighboring words are connected to each other expect then when there is a separator line. The separator line means that even though the words are next to each other there is no connection between them.

In the browser version, user can move on the mind map using their mouse. With scrolling, user can zoom in or out of the map and by dragging the map moves. One of the most important features is hovering the mouse over words, which then highlights all the other word that are connected to word that is pointed with the user's mouse. With this feature, even the words that are not able to be near each other but still are related to same concepts, can be recognized. By clicking a specific skill, the skill moves to the middle and the map is reorganized. Then the user is able see in the middle the most important other skills related to the skill that was clicked.

5.2 Cargotec skill mind map

In Figure 9 the entire skill map of Cargotec in shown. Due to the mind map being extremely large it must be analyzed in smaller pieces so that the reader is able to see the words on the mind map.

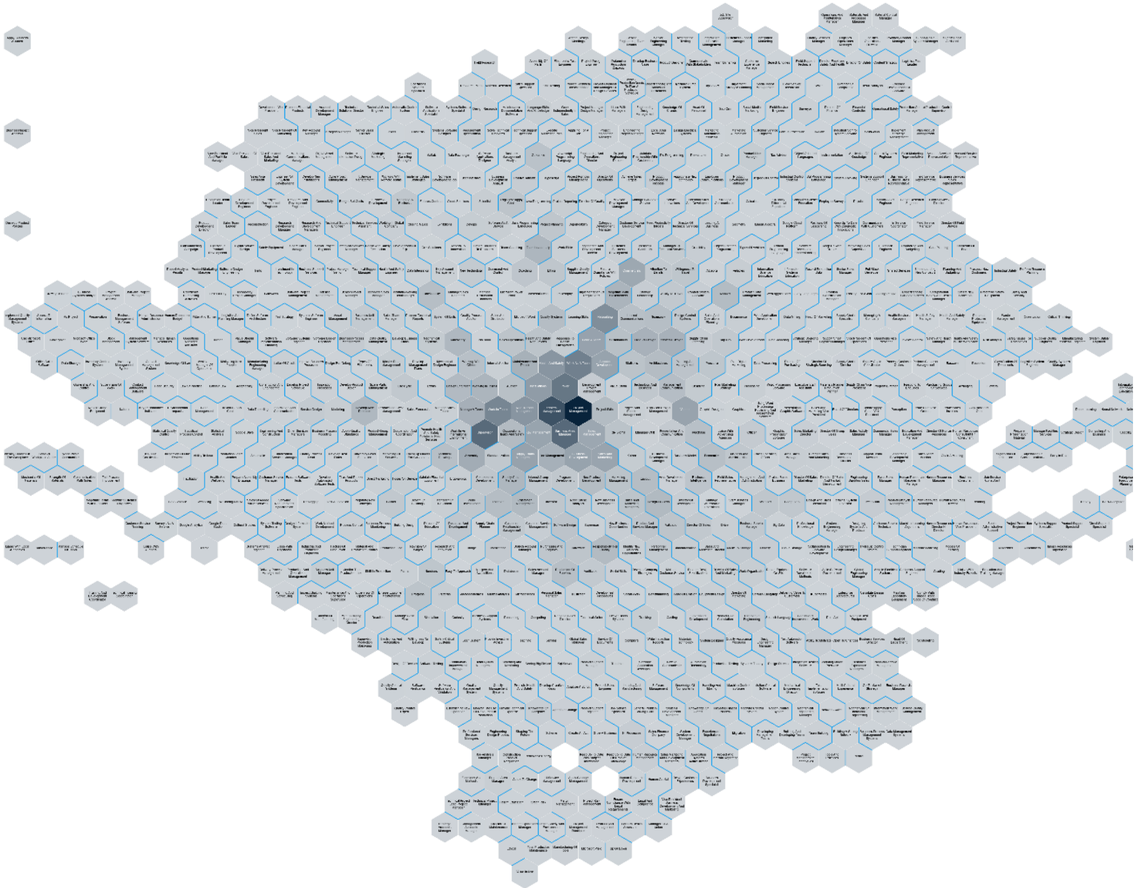

Figure 9. The entire skill mind map of Cargotec

First, the middle of the map is presented in Figure 10. The most relevant word is project management and with system management they were both as commonly found concepts from the whole corpus.

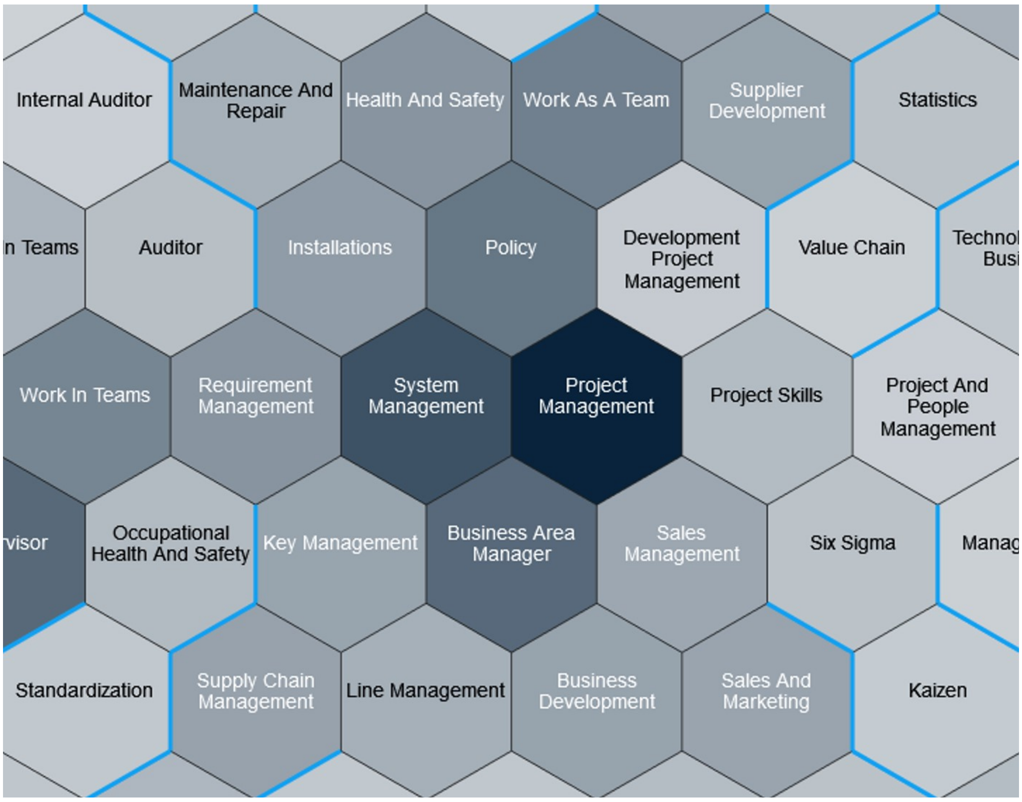

Figure 10. *The center of Cargotec's skill mind map*

Close to the center there are a lot of skills related to management, projects, and business but also teamwork is found three times quite close to the center and with a darker tone. If the AI solution would have combined these three different wording into one skill, for example, *teamwork* it probably would have been closer to the center and with darker color. From the Figure 10, the different "skill branches" start to form already. For example, from the cell *"System management"* up and a bit to the left, is a clear branch of technician skills. This technical branch is separated from auditing related skills with separator lines. Overall, the skills in the middle seem sensible, since they are skills that are needed in different professions. For example, a system engineer might need project management skills as much as a business specialist, due to work being done in projects.

Since sustainability is so high on Cargotec's agenda, it is worth of investigating. In Figure 11 concepts highlighted are related to *"Sustainability"*. It does not have a huge spread on other concepts which indicates that sustainability related skills are not widely spread into various professions.

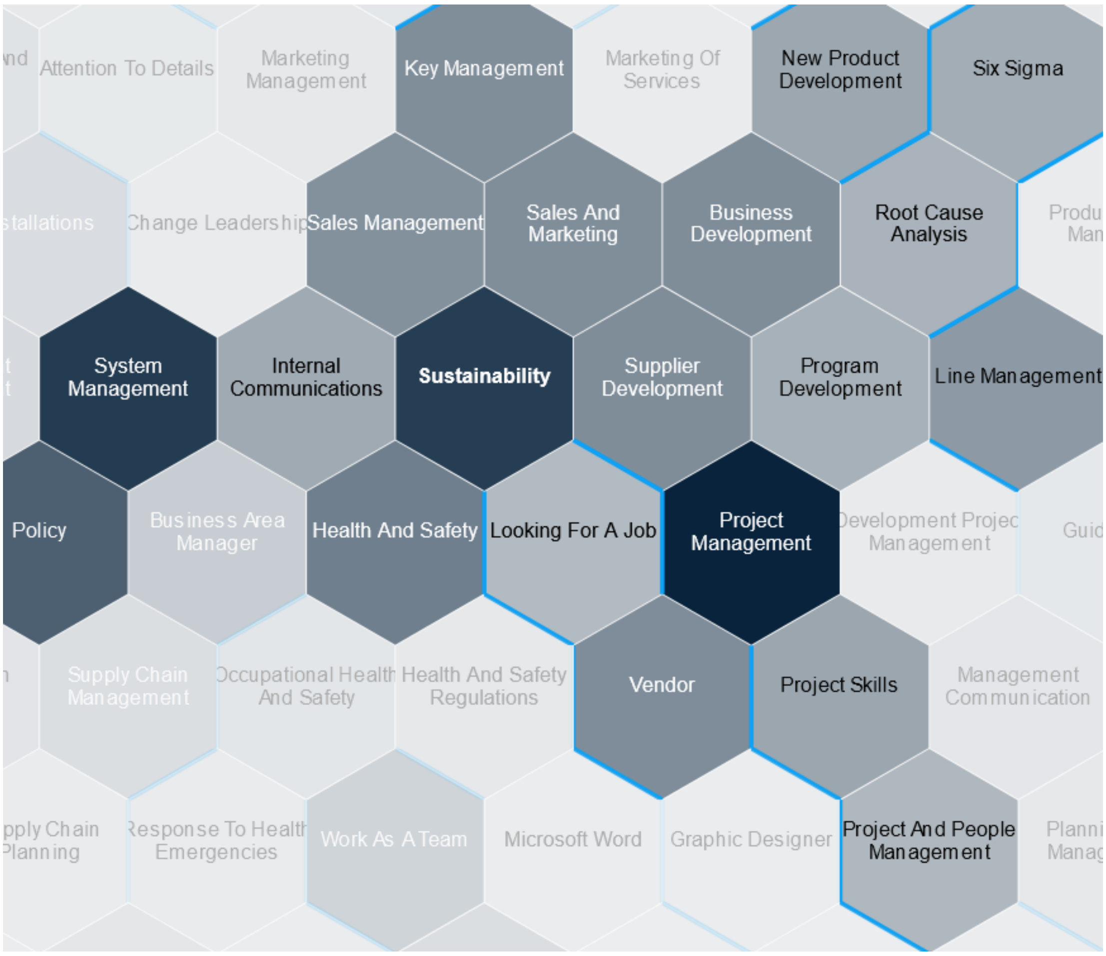

Figure 11. Skills and concepts related to "Sustainability"

However, *"Life Cycle"* has more connections to other skills and life cycle management is important for sustainable development in large manufacturing companies. Concepts especially related to life cycle management are presented in Figure 12. Though life cycle management and sustainability are related to each other as concepts the map does not show this connection.

Figure 12. *Skills and concepts related to "Life Cycle"*

Another interesting topic is data related concepts. According to literature it is one of the most important future skills. In Cargotec's skill mind map data is found in two different clusters, but it does not have as many connections as according to literature it should have. In Figure 13 skills and concepts related to *"Data Processing"* are highlighted and there are only a few connections. From the Figure 13 it could be concluded that sales data is the focus point of data analytics.

Figure 13. *Skills and concepts related to "Data Processing"*

However, the other skill cluster on the other side of the skill mind map is around *"Master Data Management"* shown in Figure 14. It relates to data quality, business support services, business process management and life cycle. It is interesting that these data clusters do not seem to be connected in the mind map, even though they are all related to working with data. Even though data related skills are found from the mind map some important data skills are still missing. For example, data interpretation and data visualization skills are not in the map even though in the literature review they were highlighted as one of the most important skills for the future.

Figure 14. *Skills and concepts related to "Master Data Management"*

Other very important skill brought up in literature was the ability to learn. *"Learning Skills"* and *"Competence Development"* are found from Cargotec's mind map, located quite near the center, and together have many connections. Figure 15 shows that learning skills are connected to health and safety related concepts which is an interesting connection. This indicates that learning and safety are often mentioned in the same context. In addition to health and safety, also business development, project skills and internal communications are shown to have connections with learning skills.

Figure 15. *Skills and concepts related to "Learning Skills"*

Competence development has much more connections to other concepts than learning skills as seen in Figure 16. It is more related to different kinds of development topics such as business development skills, work development. Learning skills and competence development do not seem to have a connection in the map, even though they are similar concepts. Though both concepts also have connections to blue-collar professions. Learning skills are connected to assembly and competence development has connection to maintenance and repair.

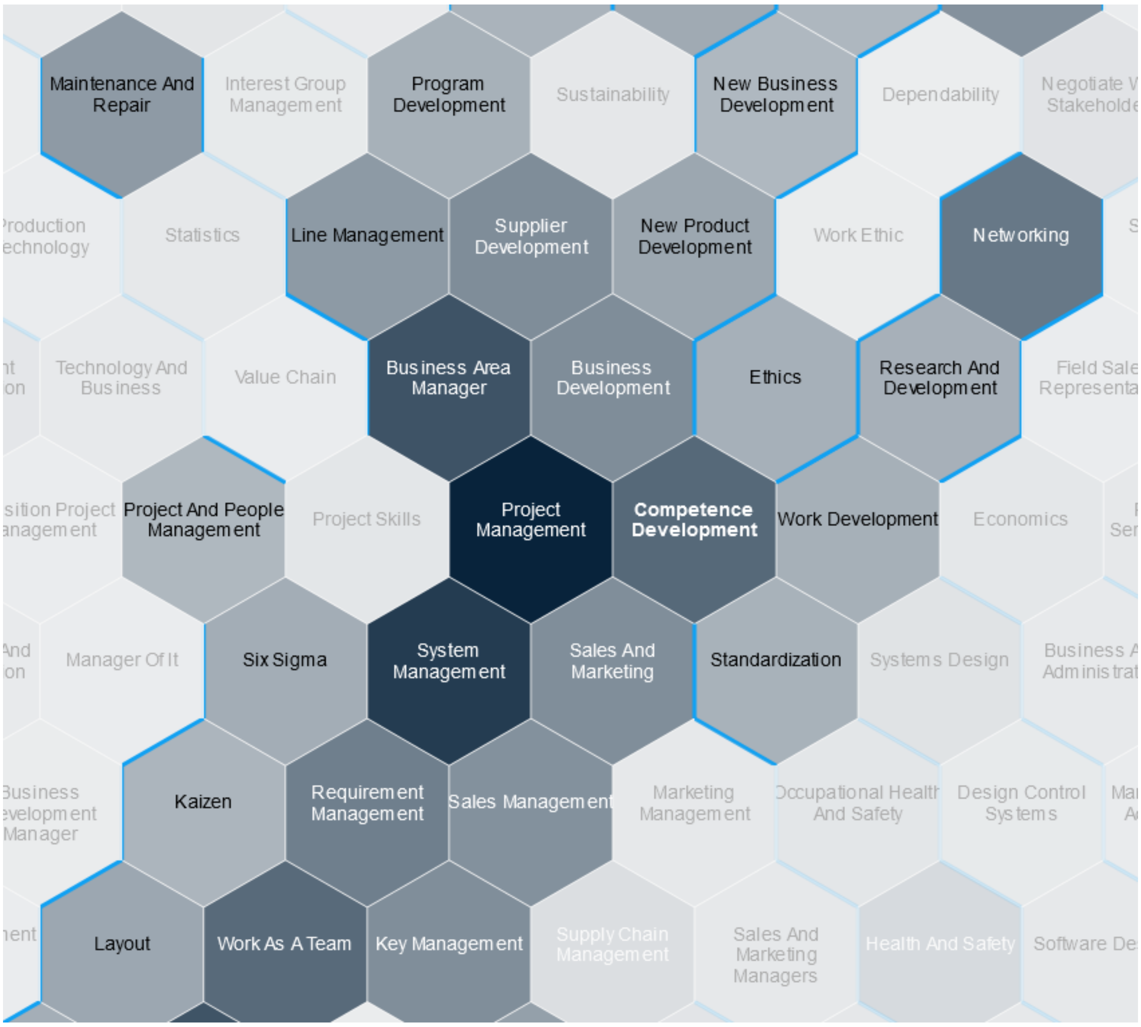

Figure 16. *Skills and concepts related to "Competence Development"*

Overall, Cargotec's mind map has similar skills that are mentioned in various literature sources. Technical skills are found from the sides of the map which means that they are not relevant skills for various professions, but they only exist in they own "technical clusters". Technical skills in the map include skills related to simulation, software, testing, automation, and machine learning. In the future, technical skills, especially data literacy skills, are more and more important for everyone because of digitalization of work and ways of working. It is important that all employees have the necessary technical skills to perform in the work of the future. Electronics and electrification related concepts were hardly found. These are skills that are not needed from everyone but due to Cargotec's high ambition with its eco-portfolio it would be useful if there would be more know-how related to electrification. Also, communication skills were lacking from Cargotec's mind map. However, it might be that communication is considered as a necessity and part of the teamwork skills and is therefore not mentioned in the job descriptions. All in all, Cargotec seems to have a good range of future skills but for the future it is important to maintain and further train personnel's technical skills.

5.3 Skill mind maps of the industry clusters

This section introduces the skill mind maps of three different industry clusters. First, the machine & metal industry's mind map is presented and compared to Cargotec's mind map then electronics & electrical industry's mind map, and lastly IT industry's mind map.

5.3.1 Machine & metal industry's skill mind map

Machine & metal industry (MMI) is closest to Cargotec's operations out of the three picked industries. The center of MMI's skill mind map is presented in Figure 17. In the center of MMI's skill mind map teamwork related skills appear to be the most relevant skills and mentioned the most of all skills. Other soft skills like positive attitude, innovation, working under pressure, and leadership skills are close to the centrum as well but there are some technical skills close to the center as well. For example, technical drawing skills, energy storage and market skills, design engineering and assembly. The soft skills have the same concepts as literature, but the technical skills close to the center are different and related more to energy than digitalization technologies.

Figure 17. *The center of machine & metal industry's skill mind map*

Further from the middle, more technical skills start to appear but those are more related to energy than for example data or AI as well. Learning skills, curiosity, and critical thinking are near the center as well but do not have many connections to other concepts. Sustainability concepts such as limiting global warming and renewable energy have a lot of connections with other skills. Also, working in an international environment is relevant concept.

The soft skills in MMI mind map are quite similar compared to Cargotec. In both skill mind maps team, project, and learning skills appear in same areas of the map. Cargotec has a higher priority with sales related skills, but in both maps business skills, such as business development, have roughly the same relevancy. Biggest difference in these maps are the energy and electronics related skills. In Cargotec's skill mind map they are barely there whereas in MMI skills map they have a huge relevancy, a lot of connections, and were mentioned many times. Also, automation related skills such as control system engineering and robotic engineering are closer to the center than in Cargotec mind map.

Further away from the middle data and technology skills are starting to appear in MMI's skill mind map. From example concepts such as *"Linked Data", "Big Data", "Cloud Computing", "Artificial Intelligence",* and *"Data Strategy"* appear. These form a clear cluster of technical skills that are often mentioned in the same contexts. Cargotec's skill mind map did not have similar kind of clear clusters around these technical skills and all these same skills were not even found from the map.

Even though communication skills are not mentioned near the center, in MMI's skill mind map a bit further from the middle are a lot of different language skills. For example, Italian, German, Chinese, Spanish and French is mentioned. English language skills are not in the same language cluster which might mean that English is self-evident skill in many roles.

Overall, machine & metal industry's and Cargotec's skill mind maps have a lot of similarities but some differences as well. The soft skills have a lot of same concepts but MMI's mind map has mentions of various different soft skills that were missing from Cargotec's map. Technical skills have more differences. The energy and electrical skills were much more highlighted in MMI's skill mind map than in Cargotec's. MMI's map had also more modern technologies mentioned but Cargotec's mind map had more data skills, and the data skills were connected to many different concepts. Skills mentioned in the literature were found from both maps, but the order of importance varied between literature and the skill mind maps.

5.3.2 Electronics & electrical industry's skill mind map

Electronics & electrical industry's (EEI) skill mind map is related to Cargotec's operations due to Cargotec's expanding eco-portfolio. The EEI map has clearly two different electrical industries. The other is related to electrical car industry and the other has skills and concepts related to power grids. Figure 18 shows the center of the EEI's skill mind map. The most relevant and common concept is battery which is not surprising since batteries are used in many different areas of this industry cluster.

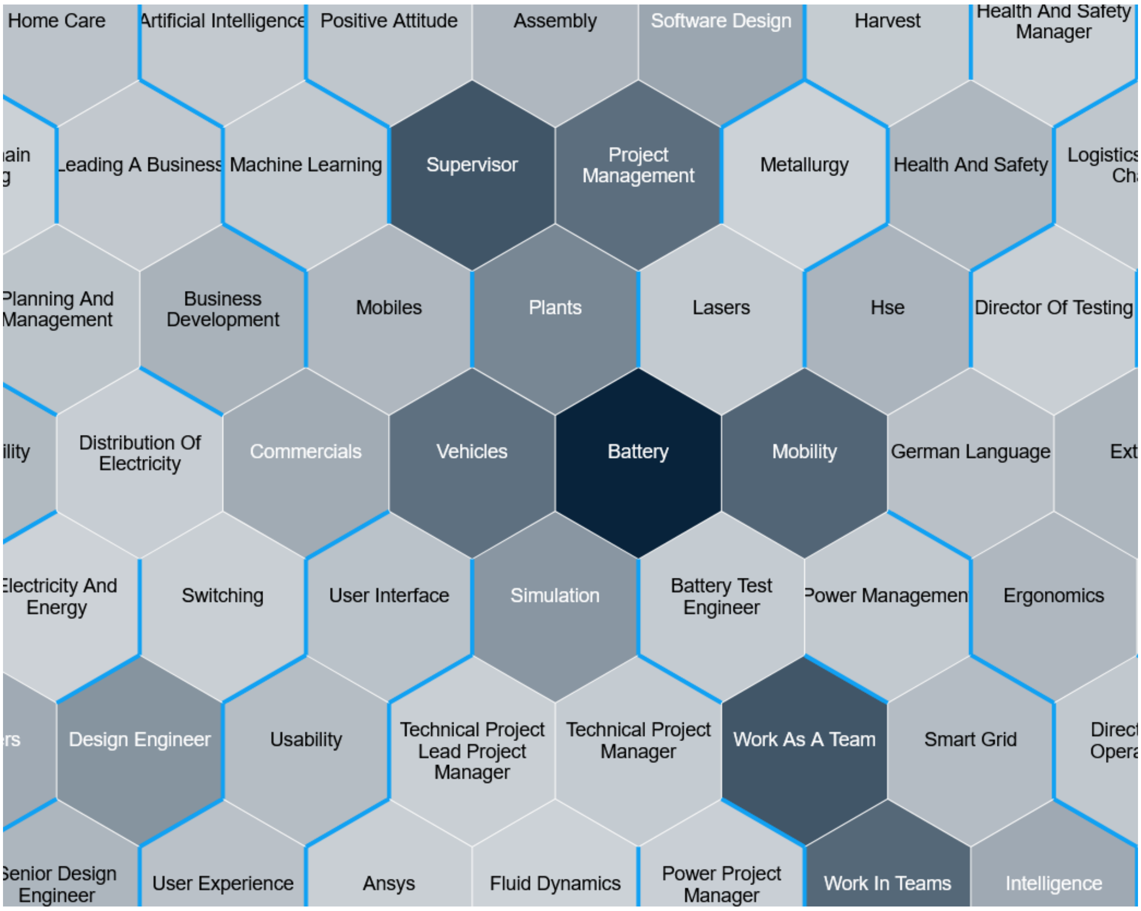

Figure 18. *The center of electronics & electrical industry's skill mind map*

This map has a lot more technical skills and technology concepts in the middle than the previous maps. It is quite clear that the car industry is hiring a lot of people who have knowledge on electronics. It is also interesting how close to the center technologies such as AI and machine learning are compared to the previous maps. The only two skills that can be purely categorized as soft skills in Figure 18 are positive attitude and teamwork skills. Although, leadership skills are also quite noticeable and connected to different concepts.

Compared to Cargotec's skill mind map, business skills do not have a such a high priority even though for example business development is in the center of the EEI's map. Similar phenomenon happens with sales skills, but they are even further from the middle. The reason might be the dominance of vehicle sectors in this cluster, since cars are usually sold to people and not for other businesses and therefore the whole sales process is different which would explain the lack of sales skills. There are some skills related to supply chains, but they are a bit further from the middle.

Even though the EEI's skill mind map has a lot of technical skills and technology concepts, data is not visible. Data skills are far from the middle, and they do not have any connections to AI or machine learning. AI is connected to automation skills, software development, and new technology concept. Science related skills are as well much more relevant in EEI's map than in other maps presented previously.

EEI’s skill mind map has a lot more technical skills with higher relevancy than in Cargotec’s skills mind map but is lacking many important soft skills. Learning skills are mentioned in the EEI’s map but they are much further from the middle than in Cargotec’s map. Even business skills are less relevant than technical skills, but project skills are almost as relevant in EEI’s map as in Cargotec's map.

5.3.3 IT industry’s skill mind map

IT industy’s (ITI) skill mind map was included for this study because even more software expertise is needed in the manufacturing industry as well. The main skills in ITI skill mind map differs quite a lot from the other presented mind maps due to its technologies and clearly defined working methods. Figure 19 shows the center of ITI’s skill mind map which has DevOps right in the middle with the darkest color as well. It is surrounded with open-source technologies, programming languages, test automation, and teamwork. In fact, the only soft skills in the center of the ITI map are teamwork and learning skills. Almost everything else is directly related to specific technologies or programming languages except business model skills and headhunter concept.

Figure 19. The center of information technology industry’s skill mind map

Different programming languages and technologies continue further from the middle as well. Some business and project skills start to appear around seven cells away from the middle but only a few. Compared to Cargotec's skill mind map, ITI's skill mind map is almost completely different. The only unifying factor is teamwork and learning skills that appear around the same area in both maps. Many of the different technologies in ITI map are connected with each other and teamwork is connected to various words on the map. Data as a word is not mentioned that often but programming languages used in data processing and data visualizations are found from the map. ITI seems to have much more defined techniques and skills than other areas and for example understanding the logic of programming is not enough.

The ITI map has a lot less business and sales related skills than Cargotec but there are a few marketing related skills further away from the center. Also, language skills or communication skills are not specifically mentioned anywhere on the ITI skill mind map, and it has only a few soft skills. Overall, ITI skill mind map is completely different from Cargotec's skill mind map and there are only a few similarities in this huge sea of skills and concepts.

6. CONCLUSION

The aim of the thesis was to find relevant future skills and to see if NLP can pick skill-related words from job descriptions and create a skill profile. The effects of digitalization for future work and future skills were studied using literature review in the first phase of this thesis. The second part of this thesis consisted of introduction to NLP and a practical implementation using NLP to form skill profiles for an industrial company and three other industry fields. The results from the second part were compared to the findings of the first part of the thesis. This thesis was able to answer to the two research questions presented below.

- What are the future skills needed during the digital revolution in industrial companies? and;
- Can NLP pick skill related words from job descriptions and job advertisements and create a skill profile from text data?

It is unclear what digitalization will do to job structures, but it is certain that other jobs will remain while others will vanish. It is predicted that mid-skilled jobs would diminish while the amount low-skilled and high-skilled jobs would remain the same or even increase. This and the change in ways of working caused by digitalization is forcing employees to learn new skills.

Future skills can be roughly divided into two categories: soft skills and technical skills. The literature review was used to find out these skills and in many different sources the same skills seemed to emerge. The most presented future soft skills were ability to learn, teamwork, problem-solving, leading others, self-leadership, critical thinking, and communication. Technical skills had a bit more variety depending on the source, but the most popular technical skills were still quite similar across multiple sources. The most important technical skills include data literacy skills, such as data analytics and data visualization, programming skills, cyber security, and overall Industry 4.0 technology skills, such as cloud computing, machine learning, and IoT. What also was discussed in various sources, was that also employees in non-technical roles need to have some technical skills, especially related to data. The same applied to soft skills and how they are especially important for all employees, regardless of role.

For companies to succeed in digital transformation, it is important to have the workforce with the correct skills. However, it is hard to identify all the relevant existing skills in the organization. To form a learning strategy, it is important to know the current skill profile of the organization in order to make a development plan for the future. Reading different text sources to define the organization's skill profile would require a vast amount of resources especially in large organizations. Therefore, the second research question tried to answer if the skill profile could be done using natural language processing.

The practical part analyzed over 1500 job descriptions from Cargotec and a vast amount of job advertisements from three different industries. The NLP solution was able to pick skill related words and form a skill profile but not perfectly. The skill mind maps contained words that were not exactly skill words but rather concepts. Also, it was not able to connect all the same skills with a bit different wording as a same skill. This made it harder to analyze the results and notice skill-connections. The number of skills and concepts presented in the skill mind maps made it impossible to go through all the results, but the most important and most relevant skills were easy to find from the skill mind maps.

The most important skills in the skill mind maps reflected to some extend the skills found in the literature review. However, the soft skills were not as visible in the results and there were far fewer soft skill words in the results than in the literature review. This might be due to many different reasons but one reason might be that some skills are thought to be included in one skill. The other reason might be that since job descriptions and job advertisements are done by HR professionals, they ask contributions for the technical requirements from hiring managers but list the technical skills by themselves. This can result in many job descriptions using the same words to describe soft skills because the soft skill requirements are written by one person into many different job texts.

Technical skills had a lot more variety in the results than in literature review. This is of course partly influenced by the fact that research tries to find universal technical skills whereas the job descriptions and advertisements must describe accurately the needed skills in specific role. Still, same concepts and technologies were found both form the literature and from the skill mind maps. Though, in the skill mind maps the technical skills seemed to be still connected with other technical skills, which would mean that technical skills are not yet required in non-technical roles. The relevancy of technical skills varied a lot between industries, but in each industry at least a few technical skills were among the most found skills.

A human interpretation of the results provided by the NLP solution is still much needed. For better results, the semantic computing needs further improvements which would result in easier interpretation of the results for humans. Also, a search option in the skill mind maps might make it easier to recognized different concepts from the large skill mind maps. The large size of the mind maps required a lot of computing power and processing them in the browser was slow. Based on the results however, it can be said that NLP makes it easier to recognize the most relevant and common skills found in an organization but for easier interpretation further development is still needed.

It is also important to note what kind of data was used to make the skill mind maps. Even though job descriptions and job advertisements try to list all the necessary skills needed in the role, they do not guarantee that the hired employee would have all the required skills. In turn, employees might have a lot of other skills that are not listed in the job texts. Therefore, the results do not give a completely accurate picture of the organization's skills, but they can nevertheless be considered quite reliable due to large amount of data.

In the light of the improvement ideas, it would be important to delve deeper in the semantic computing as a further development of this thesis. Also, it would be interesting to follow how these skill mind maps change overtime and analyze the reasons behind the changes. A good extension for the NLP solution would be a prediction of skills. This would help users to think about different scenarios that might happen and how to prepare for them. Another further research topic could what kind of data would give more accurate results of the organization’s existing skills.

REFERENCES

Accenture. (2017). New Skills Now: Inclusion in the digital economy. Available at: https://www.accenture.com/_acnmedia/pdf-63/accenture-new-skills-now-inclusion-in-the-digital.pdf [Accessed 22.10.2022]

Aggarwal. (2018). Machine learning for text. Springer. Available at: https://andor.tuni.fi/permalink/358FIN_TAMPO/1j3mh4m/alma9910393524205973 [Accessed 28.10.2022]

Alcácer, V. & Cruz-Machado, V. (2019). Scanning the Industry 4.0: A Literature Review on Technologies for Manufacturing Systems. Engineering Science and Technology, an International Journal, 22(3), 899–919. Available at: https://doi.org/10.1016/j.jestch.2019.01.006 [Accessed 22.8.2022]

Almeida, R., Behrman, J., & Robalino, D. (2012). The right skills for the job? rethinking training policies for workers. World Bank. Available at: https://ebookcentral.proquest.com/lib/tampere/detail.action?docID=978133 [Accessed: 18.10.2022]
Andersson, C., Haavisto, I., Kangasniemi, M., Kauhanen, A., Tikka, T., Tähtinen, L. & Törmanen, A. (2016). Robotit töihin: Koneet tulivat – mitä tapahtuu työpaikoilla? EVA raportti 2/2016. Available at: http://www.eva.fi/wp-content/uploads/2016/09/Robotit-t%C3%B6ihin.pdf [Accessed 22.10.2022]

Altınel, B. & Ganiz, M. C. (2018). Semantic text classification: A survey of past and recent advances. Information Processing & Management, 54(6), 1129–1153. Available at: https://doi.org/10.1016/j.ipm.2018.08.001 [Accessed 29.10.2022]

Antikainen, M., Uusitalo, T., & Kivikytö-Reponen, P. (2018). Digitalisation as an Enabler of Circular Economy. 10TH CIRP CONFERENCE ON INDUSTRIAL PRODUCT-SERVICE SYSTEMS, IPS2 2018, 73, 45–49. Available at: https://doi.org/10.1016/j.procir.2018.04.027 [Accessed 11.8.2022]

Auktor, G. (2020). Green Industrial Skills for a Sustainable Future. United Nations Industrial Development Organization. Available at: https://lkdfacility.org/wp-content/uploads/LKDForum-2020_Green-Skills-for-a-Sustainable-Future.pdf [Accessed 22.10.2022]

Autor, D. H. (2015). Why Are There Still So Many Jobs? The History and Future of Workplace Automation. The Journal of Economic Perspectives, 29(3), 3–30. Available at: https://doi.org/10.1257/jep.29.3.3 [Accessed 16.8.2022]

Beysolow II. (2018). Applied Natural Language Processing with Python Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing (1st ed. 2018.). Apress. Available at: https://doi.org/10.1007/978-1-4842-3733-5 [Accessed 25.10.2022]

Birgün, S. & Ulu, M. (2020). Site Selection for a Training Centre Focused on Industry 4.0 by Using DEMATEL and COPRAS. In Digital Conversion on the Way to Industry 4.0 (pp. 37–50). Springer International Publishing. Available at: https://doi.org/10.1007/978-3-030-62784-3_4 [Accessed 18.8.2022]

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016). fastText. Available at: https://research.fb.com/blog/2016/08/fasttext/ [Accessed 1.11.2022]

Cammack, R., Atwood, T., Campbell, P., Parish, H., Smith, A., Vella, F. and Stirling, J. (2006) ‘Natural language processing’, in Oxford Dictionary of Biochemistry and Molecular Biology.

Chandrasekaran, D. & Mago, V. (2021). Evolution of Semantic Similarity-A Survey. ACM Computing Surveys, 54(2), 1–37. Available at: https://doi.org/10.1145/3440755 [Accessed 28.10.2022]

Church, K. W. (2017). Emerging Trends: Word2Vec. Natural Language Engineering, 23(1), 155–162. Available at: https://doi.org/10.1017/S1351324916000334 [Accessed 29.10.2022]

Cohen, & Demner-Fushman, D. (2014). Biomedical natural language processing. J. Benjamins Publishing Company. Available at: https://ebookcentral.proquest.com/lib/tampere/detail.action?docID=3016032 [Accessed 25.10.2022]

Constantinou, S. (2021) New National Electrification Skills Framework and Forum could put the UK at the forefront of the green revolution. The Faraday Institution. Available at: https://www.faraday.ac.uk/electrification-skills-framework-and-forum/ [Accessed 22.8.2022]

Drath, R. & Horch, A. (2014). Industrie 4.0: Hit or Hype? [Industry Forum]. IEEE Industrial Electronics Magazine, 8(2), 56–58. Available at: https://doi.org/10.1109/MIE.2014.2312079 [Accessed 18.8.2022]

Degryse, C. (2016). Digitalisation of the Economy and its Impact on Labour Markets. SSRN Electronic Journal. Available at: https://doi.org/10.2139/ssrn.2730550 [Accessed 22.8.2022]

Deng, L. & Liu, Y. (2018). Deep learning in natural language processing. Springer. Available at: https://ebookcentral.proquest.com/lib/tampere/detail.action?docID=5401147 [Accessed 25.10.2022]

De Smet, A., Reich, A. & Schaninger B. (2021). Getting skills transformations right: The nine-ingredient recipe for success. McKinsey & Company. Available at: https://www.mckinsey.com/business-functions/people-and-organizational-performance/our-insights/the-organization-blog/getting-skills-transformations-right-the-nine-ingredient-recipe-for-success [Accessed: 15.8.2022]

Flach. (2012). Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press. Available at: https://ebookcentral.proquest.com/lib/tampere/detail.action?docID=1025072 [Accessed 30.10.2022]

Fonseca, L. M. (2018). Industry 4.0 and the digital society: concepts, dimensions and envisioned benefits. PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON BUSINESS EXCELLENCE, 12(1), 386–397. Available at: https://doi.org/10.2478/picbe-2018-0034 [Accessed 11.8.2022]

Fonseca, P. & Picoto, W. N. (2020). The competencies needed for digital transformation. The Online journal of applied knowledge management. [Online] 8 (2), 53–70.

Garman, A. N., Erwin, T. S., Garman, T. R. & Kim, D. H. (2021). Developing competency frameworks using natural language processing: An exploratory study. The Journal of Competency-Based Education, 6(3). Available at: https://doi.org/10.1002/cbe2.1256 [Accessed 16.8.2022]

Goldenstein, J., Poschmann, P., & Händschke, S. G. M. (2015). Linguistic Analysis: The Study of Textual Data in Management and Organization Studies with NLP. Academy of Management Annual Meeting Proceedings, 2015(1), 10882–. Available at: https://doi.org/10.5465/ambpp.2015.10882abstract [Accessed 16.8.2022]

Hallamaa, T. (2022). Suomalaisnuorten koulutustaso putosi OECD-keskitason alapuolelle. Yle Uutiset (engl. Yle News) 3.10.2022. Available at: https://yle.fi/uutiset/74-20000622 [Accessed 16.10.2022]

Harper, T., Drury, W. & Greenwood, D. (2021). The Opportunity for a National Electrification Skills Framework and Forum. Catapult, The University of Warwick & The Faraday Institution. Available at: https://hvm.catapult.org.uk/wp-content/uploads/2021/11/National-Electrification-Skills-Forum-Brochure-FINAL.pdf [Accessed 22.8.2022]

Heideman Lassen, A. & Waehrens, B. V. V. (2021). Labour 4.0: developing competences for smart production. JOURNAL OF GLOBAL OPERATIONS AND STRATEGIC SOURCING, 14(4), 659–679. Available at: https://doi.org/10.1108/JGOSS-11-2019-0064 [Accessed 22.10.2022]

Hirschi, A. (2018). The Fourth Industrial Revolution: Issues and Implications for Career Research and Practice. The Career Development Quarterly, 66(3), 192–204. Available at: https://doi.org/10.1002/cdq.12142 [Accessed 16.8.2022]

Hirsch-Kreinsen, H. 2016. Digitization of industrial work: development paths and prospects. Zeitschrift für Arbeitsmarktforschung. s. 4. Available at: https://link.springer.com/content/pdf/10.1007/s12651-016-0200-6.pdf [Accessed 16.8.2022]

Ilmarinen, V. & Koskela, K. (2015). Digitalisaatio : yritysjohdon käsikirja . Helsinki: Talentum. s. 22-23.

ILO (2018). World employment and social outlook 2018: Greening with jobs. Geneva: ILO. Available at: https://www.ilo.org/global/about-the-ilo/newsroom/news/WCMS_628644/lang--en/index.htm [Accessed 22.10.2022]

ILO (2019). Skills for a greener future: A global view based on 32 country studies. Geneva: ILO. Available at: https://www.ilo.org/wcmsp5/groups/public/---ed_emp/documents/publication/wcms_732214.pdf [Accessed 22.10.2022]

Indurkhya, N., Damerau, F. J., & Damerau, F. J. (Frederick J. (2010). Handbook of natural language processing (Second edition.). Taylor & Francis. Available at: https://doi.org/10.1201/9781420085938 [Accessed 28.10.2022]

Islam, A. (2022). Industry 4.0: Skill set for employability. Social Sciences & Humanities Open, 6(1), 100280–. Available at: https://doi.org/10.1016/j.ssaho.2022.100280 [Accessed 22.10.2022]

Kadir, B. A., Broberg, O., & Conceição, C. S. da. (2019). Current research and future perspectives on human factors and ergonomics in Industry 4.0. Computers & Industrial Engineering, 137, 106004. Available at: https://doi.org/10.1016/j.cie.2019.106004 [Accessed 29.9.2022]

Kang, Y., Cai, Z., Tan, C.-W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics, 7(2), 139–172. Available at: https://doi.org/10.1080/23270012.2020.1756939 [Accessed 25.10.2022]

Kao, A. & Poteet, S. R. (2007). Natural Language Processing and Text Mining (Kao & S. R. Poteet, Eds.; 1st ed. 2007.). Springer London. Available at: https://doi.org/10.1007/978-1-84628-754-1 [Accessed 25.10.2022]

Kohtamäki, M., Parida, V., Patel, P. C., & Gebauer, H. (2020). The relationship between digitalization and servitization: The role of servitization in capturing the financial potential of digitalization. Technological Forecasting & Social Change, 151, 119804–. Available at: https://doi.org/10.1016/j.techfore.2019.119804 [Accessed: 15.8.2022]

Khurana, D., Koli, A., Khatter, K. & Singh, S. (2022). Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications, 1–32. Available at: https://doi.org/10.1007/s11042-022-13428-4 [Accessed 25.10.2022]

Laaksonen, J. & Honkela, T. (2011). Advances in Self-Organizing Maps 8th International Workshop, WSOM 2011, Espoo, Finland, June 13-15, 2011. Proceedings (J. Laaksonen & T. Honkela, Eds.; 1st ed. 2011.). Springer Berlin Heidelberg. Available at: https://doi.org/10.1007/978-3-642-21566-7 [Accessed 29.10.2022]

Langley, D. J., van Doorn, J., Ng, I. C. L., Stieglitz, S., Lazovik, A., & Boonstra, A. (2021). The Internet of Everything: Smart things and their impact on business models. Journal of Business Research, 122, 853–863. Available at: https://doi.org/10.1016/j.jbusres.2019.12.035 [Accessed 22.8.2022]

Leonardi, P. M. (2021). COVID-19 and the New Technologies of Organizing: Digital Exhaust, Digital Footprints, and Artificial Intelligence in the Wake of Remote Work. Journal of Management Studies, 58(1), 247–251. Available at: https://doi.org/10.1111/joms.12648 [Accessed 22.8.2022]

Marr, B. (2016). Why Everyone Must Get Ready For The 4th Industrial Revolution. Forbes. Available at: https://www.forbes.com/sites/bernardmarr/2016/04/05/why-everyone-must-get-ready-for-4th-industrial-revolution/?sh=3274ee1b3f90 [Accessed 15.8.2022]

Marr, B. (2022). The Top 10 Most In-Demand Skills For The Next 10 Years. Forbes. Available at: https://www.forbes.com/sites/bernardmarr/2022/08/22/the-top-10-most-in-demand-skills-for-the-next-10-years/?sh=5acee8f917be [Accessed 22.10.2022]

Maisiri, W. Darwish, H., & van Dyk, L. (2019). An investigation of industry 4.0 skills requirements. South African Journal of Industrial Engineering, 30(3), 90–105. Available at: https://doi.org/10.7166/30-3-2230 [Accessed 22.8.]

Martinez, A. R. (2010). Natural language processing. Wiley Interdisciplinary Reviews. Computational Statistics, 2(3), 352–357. Available at: https://doi.org/10.1002/wics.76 [Accessed 28.10.2022]

Montelisciani, G., Gabelloni, D., Tazzini, G., & Fantoni, G. (2014). Skills and wills: the keys to identify the right team in collaborative innovation platforms. Technology Analysis & Strategic Management, 26(6), 687–702. Available at: https://doi.org/10.1080/09537325.2014.923095 [Accessed 16.8.2022]

Müller, A. C. & Guido, S. (2017). Introduction to machine learning with Python a guide for data scientists (1st ed.). O'Reilly Media. Available at: https://andor.tuni.fi/permalink/358FIN_TAMPO/1j3mh4m/alma9910689272305973 [Accessed 30.10.2022]

Nelson M.M. & Illingworth, W.T. (1991). A Practical Guide to neural nets.

Nousiainen, R. (2020). Osaaminen mahdollistaa muutoksen, muutos synnyttää osaamisvaatimuksia. STTK ry. Available at: https://www.sttk.fi/2020/12/21/osaaminen-mahdollistaa-muutoksen-muutos-synnyttaa-osaamisvaatimuksia/ [Accessed: 16.10.2022]

Oliver, D., Freeman, B., Young, C., Yu, S. & Verma, G. (2014). Employer satisfaction survey report for the Department of Education. Available at: https://www.researchgate.net/publication/272746861_Employer_satisfaction_survey_report_for_the_Department_of_Education [Accessed: 18.10.2022]

Parviainen, P., Federley, M., Seisto, A., Koponen, J., Annala, M., Korhonen, O., & Harjunen, V. (2017a). Digimuutoksessa onnistumisen eväät. Prime Minister's Office Finland. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja Vol. 54 Available at: http://urn.fi/URN:ISBN:978-952-287-436-8 [Accessed 16.10.2022]

Parviainen, P, Federley, M., Grenman, K. & Seisto, A. (2017b). Osaaminen ja työllisyys digimurroksessa. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja, vol. 24/2017, Prime Minister's Office Finland, Helsinki. S. 24 Available at: https://tietokayttoon.fi/julkaisu?pubid=17806 [Accessed: 15.8.2022]

Perna, M. C. (2022). Why 78% Of Employers Are Sacrificing Employee Trust By Spying On Them. Forbes 15.3.2022. Available at: https://www.forbes.com/sites/markcperna/2022/03/15/why-78-of-employers-are-sacrificing-employee-trust-by-spying-on-them/?sh=14eaa9ec1659 [Accessed: 16.10.2022]

Pitkänen, M. & Silvén, O. (2019). Tekoälyn haasteet – Koneoppimisesta ja konenäöstä tunnetekoälyyn. Oulun yliopisto: Konenäön ja signaalianalyysin keskus. Available at: http://jultika.oulu.fi/files/isbn9789526224824.pdf [Accessed 30.10.2022]

Porter, M. E. & Heppelmann, J. E. (2014). How Smart, Connected Products Are Transforming Competition. Harvard Business Review, 92(11), s. 64–74, 84. Available at: https://hbr.org/2014/11/how-smart-connected-products-are-transforming-competition [Accessed: 15.8.2022]

Saarikoski, J. (2014). On text document classification and retrieval using self-organising maps. Tampere University Press. Available at: https://urn.fi/URN:ISBN:978-951-44-9627-1 [Accessed 29.10.2022]

Skender, F. & Ali, I. (2019). DIGITALIZATION AND INDUSTRY 4.0. Vision Journal Vol. 4 No. 2. s. 47-62. Available at: https://www.researchgate.net/publication/343639352_DIGITALIZATION_AND_INDUSTRY_40 [Accessed 15.8.2022]

Stock, T., & Seliger, G. (2016). Opportunities of Sustainable Manufacturing in Industry 4.0. Procedia CIRP, 40, 536–541. Available at: https://doi.org/10.1016/j.procir.2016.01.129 [Accessed 16.8.2022]

Suleman, F. (2018). The employability skills of higher education graduates: insights into conceptual frameworks and methodological options. Higher Education, 76(2), 263–278. Available at: https://doi.org/10.1007/s10734-017-0207-0 [Accessed: 18.10.2022]

Technology Industries of Finland (fin. Teknologiateollisuus ry), (2022a). Skills Data Playbook. Available at: https://teknologiateollisuus.fi/sites/default/files/inline-files/Osaamisdatan-Playbook--ENG-03-aukeamittain_0.pdf [Accessed 16.8.2022]

Technology Idustries of Finland (fin. Teknologiateollisuus ry). (2022b). Teknologiateollisuuden viestit budjettiriiheen 2022. Available at: https://teknologiateollisuus.fi/en/node/28384 [Accessed 22.8.2022]

Uysal, & Gunal, S. (2014). The impact of preprocessing on text classification. Information Processing & Management, 50(1), 104–112. Available at: https://doi.org/10.1016/j.ipm.2013.08.006 [Accessed 28.10.2022]

Wang, S., Wan, J., Li, D., & Zhang, C. (2016). Implementing Smart Factory of Industrie 4.0: An Outlook. International Journal of Distributed Sensor Networks, 2016(1), 3159805–. Available at: https://doi.org/10.1155/2016/3159805 [Accessed 15.8.2022]

Waschull, S., Bokhorst, J. A. C., Wortmann, J. C., & Molleman, E. (2022). p The redesign of blue- and white-collar work triggered by digitalization: collar matters. Computers & Industrial Engineering, 165. Available at: https://doi.org/10.1016/j.cie.2021.107910 [Accessed 17.8.2022]

Wellener, P., Dollar, B., Ashton, H., Monck, L. & Hussain, A. (2020). The future of work in manufacturing. Deloitte. Available at: https://www2.deloitte.com/us/en/insights/industry/manufacturing/future-of-work-manufacturing-jobs-in-digital-era.html [Accessed 16.8.2022]

Wilson, R. (2013). Skills anticipation—The future of work and education. International Journal of Educational Research, 61, 101–110. Available at: https://doi.org/10.1016/j.ijer.2013.03.013 [Accessed 18.10.2022]

Wingard, J. & Farrugia, C. A. (2021). The great skills gap: optimizing talentfor the future of work (Wingard & C. A. Farrugia, Eds.). Stanford Business Books. Available at: https://doi.org/10.1515/9781503628076 [Accessed 25.10.2022]

World Economic Forum (WEF). (2018). Insight Report: The Future of Jobs Report 2018. Available at: https://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf [Accessed 16.10.2022]

Zavyiboroda, M. (2022). A guide to future-oriented skills: skills in demand to watch in the next five years. HR Forecast. Available at: https://hrforecast.com/a-guide-to-future-oriented-skills-skills-in-demand-to-watch-in-the-next-five-years/ [Accessed 22.10.2022]

Zhu, G. & Iglesias, C. A. (2017). Computing Semantic Similarity of Concepts in Knowledge Graphs. IEEE Transactions on Knowledge and Data Engineering, 29(1), 72–85. Available at: https://doi.org/10.1109/TKDE.2016.2610428 [Accessed 28.10.2022]