MATCHING BUYING AND SELLING OF SKILLS

MATCHING BUYING AND SELLING OF SKILLS: AN OPTIMAL SKILL SELECTION PROBLEM IN AN ONLINE MARKET

BY

OLENA SKALIANSKA, B.A.

A THESIS

SUBMITTED TO THE SCHOOL OF COMPUTATIONAL SCIENCE AND ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2022) McMaster University

(Computational Science and Engineering) Hamilton, Ontario, Canada

TITLE:	Matching buying and selling of skills: An optimal skill selection problem in an online market
AUTHOR:	Olena Skalianska B.A. (Business Management), Wrocław University of Science and Technology, Wrocław, Poland
SUPERVISOR(S):	Dr. Sash Vaid and Dr. Yun Zhou
NUMBER OF PAGES:	ix, 96

Abstract

The central function of the job market is to match the available job vacancies with job candidates. For job candidates, it is critical for them to be equipped with the right skills for gaining competitive advantage. In this thesis, we obtain a dataset by scraping publicly available information from job postings for data science/analytics/engineering and similar positions on an online job marketplace. In the past few years, demand for those data-related jobs has been on the rise, and many job seekers change their career path to work in this area. For that purpose, it is important for them to understand the pattern of demand for skills in the labor market and to identify the best skills to acquire for maximizing the number of job vacancies they can apply for. We address these issues based on the real-life dataset. First, through exploratory data analyses, we examine the correlation between the size of a company and the types of its required skills, as well as the correlation between the salary level offered by a company and the rating it receives on the online job marketplace. Then, we develop a linear integer programming model to formulate the skill selection problem to maximize the number of jobs covered by the selected skills. We show that the problem is NP-hard, and then solve it using both the commercial solver CPLEX and greedy heuristics.

To my mom and dad, their constant love and support

Acknowledgements

I would like to express my deepest gratitude to Dr. Yun Zhou and Dr. Sash Vaid for their invaluable support, feedback, and patience. I would also like to thank them for seeing a potential in me, bringing and helped me to realize it. I also thank Dr. Kai Huang for reviewing my thesis and for his constructive comments and suggestions., which was very helpful for improving the work.

Additionally, this endeavor would not have been possible without the generous support from Dr. Romanko, who impacted and inspired me. Who believed in me from the first day we met and motivated me to apply at McMaster.

Lastly, I would be remiss in not mentioning my family, especially my parents. Their belief in me has kept my spirits and motivation high during this process.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Although the number of job openings is growing from year to year, landing a dream job seems to have become increasingly more difficult. Among the causes, imbalance (between the number of job positions and job seekers) and market frictions are two inevitable factors that hurts the welfare of both employers and job seekers. The latter factor, market frictions, is partly caused by mismatched supply and demand for job skills - the actual knowledge, abilities, competencies that the candidate has learned, consolidated, and can offer the employer from the first days of work in a company. When a recruiter looks for their ideal candidate, the first filter they apply to the list of applicants is often the job skills. To reduce frictions, it is crucial to facilitate better matches between job skills possessed by job seekers and those demanded by employers.

To that end, companies need to keep in step with the times, and more importantly, fill positions with the right people. To stay in step with the times means to collect massive amounts of data and analyze it with different methods and metrics, which, only several years ago, required a significant amount of manual work (Mbah *et al.*,

2017). Job seekers, on the other hand, need to understand which skills are (or will be) in demand to land a job. In particular, given that many technical skills (such as IT skills) become outdated at a fast pace (McLean, 2006), it is of critical importance for job seekers to update their skills or acquire new ones on a regular basis.

Understanding what the "righ" set of skills are benefit both job seekers and academic institutions as well. It is not uncommon that students spend years to learn some skills that are either irrelevant in practice or outdated in higher educational institutions (Santoso and Putra, 2017). It has been proposed that all stakeholders mentioned above, including universities, graduates entering the job market, and employers, work together to improve employability (Tran, 2015). Existing work has also explored how well the expectations of employers match the perceptions of near-graduate students about the skills necessary for the work. They reveal that there is a gap between employers' expectations and prospective employees' perception of what will be required from them, which makes it challenging for employers to fill open positions with the talent that has required levels of skills (Gibbs *et al.*, 2011). While this has been well acknowledged, a rigorous data-driven approach is required to analyze it.

Therefore, this situation creates the conditions of monitoring trends and identifying the optimal and most promising skills in relation to the desired employment.

The goal of the thesis is to analyze the labor market using modern approaches, such as data analytics and machine learning tools, with a focus on the study of the relationships between the set of skills, remuneration and the characteristics of employers, and in addition, to identify sets of skills that can increase a job seeker's chance of successfully landing a desired job.

Our analysis is enabled through text analytics, which is an automated process for extracting important, research-relevant information from unstructured text data. The research in this thesis is based on our collection of publicly available data on jobs, required skills, the level of salary and companies' characteristics that offer vacancies to test the models. This data is obtained primarily from job-search aggregator websites. At the same time arises an important question of proper collection, aggregation, and interpretation of the gathered information. Based on the data, we identify the most in-demand skills or set of skills that guarantee a salary close to the maximum on the market, or those skills that will make the candidate's portfolio more harmonious and increase the chances of getting the desired job. Then, the thesis addresses the following research issues.

The first goal is to examine the interrelation between the set of skills and their level with the level of salary and the characteristics of the employing company. In the mean time, the practical interest lays not only in studying regression dependencies, but also in the results of the exploration of aggregated data, which help to answer the questions about the most demanded skills, skills with the maximum demand on the market and scarce availability, as well as other questions that are typical for the open market. This will allow the job-seeker to determine which of the skills need to be developed as soon as possible. The importance of focusing on answering these questions is dictated by the constraints of time and human resources: during a limited period of time, it is not possible for a job seeker to acquire many skills.

Secondly, we build an optimization model to analyze the importance of certain skills and their combinations either for maximizing the number of jobs a job seeker is eligible to apply for, or for getting a job with the highest possible salary. The

model helps job-seekers to determine which of the skills need to be developed as soon as possible, subject to the constraints of individuals' time and energy, given that it is not possible to study or be at least familiar with all existing technologies. The scientific value of the study lays in building a model that will not only allow to identify the most significant skills, but also to use textual information in the form of natural language from open sources as an input data. From a practical point of view, this study can also be a guide for candidates to upgrade their skills to maximize their ability to get a job with maximum job coverage, as well as give professionals an idea of what skills they should focus on to remain relevant in the job market, and vice versa, which skills may be kept at their current levels.

To illustrate the models and methods in this thesis, we will be looking at job postings in the IT sector. On the one hand, the IT sector can be considered as the most sensitive to skill sets and the level of mastery of those skills; on the other hand, IT skills are more diverse (Rainie and Anderson, 2020). The narrowing of focus to the IT sector does not change the approaches used in this research but allows us to slightly reduce the complexity of the models and focus on the approaches for building an analytical system. There are several problems and restrictions that can arise during market analysis, but the algorithms can help, which will be described in the subsequent chapters.

The rest of the thesis is organized as follows.

In chapter 2, we review the relevant literature.

In chapter 3, we describe our process of collecting, cleaning, and preparing data for the analysis. We then conduct exploratory and regressional analyses to understand the patterns of the skills, such as how company features (e.g., size) impact on the

requirements of hard vs. soft skills. We find that soft skills become more important as company size increases. Average salary grows both with the growth of the share of technical requirements and with an increase of the company size. A candidate with more technical skills may have a higher chance to get a high salary when employed by a large company.

In chapter 4, we develop an optimization model to investigated which skills can be acquired by the candidate in order to be able to apply for the maximum number of job postings. In other words what skills you need to develop to be able to apply for as many jobs as you can and so that your skills would match the requirements for those jobs. We find that top 10 skills to select are be 'verbal', 'writing', 'sql', 'problem solving', 'communication', 'think analytically', 'excel', 'collaborate', 'detail oriented' and 'financial'. Having these skills allows you to apply for 83 jobs that were posted on indeed.ca but having them does not guarantee the higher salary and we will explain why in the following chapters.

In chapter 5, we conclude the thesis by summarizing the main results and discussing possible future research directions.

Chapter 2

Literature Review

In this chapter, we classify the relevant literature into several categories:

(i) value and demand analysis of skills;

(ii) text analysis;

(iii) the knapsack problem and the maximum coverage problem

We review the details about the current state of research for those topics.

Effect of labor skills on trades and economies

Keesing (1966) studies how differences in skill composition of the labor force in different countries shape the pattern of international trade. Tang (2012) develops an open-economy model to study how a country can influence its labor force's skill acquisition to affect the export pattern. Bombardini *et al.*(2012) find that countries with higher skill dispersion are more productive in industries of lower skill complementarity.

Value and demand analysis of skills

Every day new job postings are being advertised on the job market and job seekers need to invest in themselves and improve their skills (Burstein and Vogel, 2017). In the past, numerous researchers try to identify and calculate the value of job skills using different approaches. Many surveys confirmed positive correlation between the level of mastery of skills and job salaries (OECD, 2013, 2016). Some online recruitment services were created recently and become a good source of the job posting data (Yan *et al.*, 2019). Thus, they created a great opportunity for job market analysis, enabling researchers to identify skills in demand by employers (CEDEFOP, 2019, 2021; Boselli *et al.*, 2018);. Moreover, majority of the research papers focus on findings the top demanded skills (Lovaglio *et al.*, 2018; Colombo *et al.*, 2018).

The U.S. Employment and Training Administration created a competency model framework, which can be graphically represented by a pyramid with nine levels. Each level consists of blocks representing the skills, knowledge, and abilities, which are essential for good performance (ETA, 2012).

Gehrke, Lars and Kühn proposed the concept of a skillset for the worker of the future. Competences were divided into two categories: technical and personal. It was identified that "must have" technical skills are IT knowledge and ability to interact human-machine or human-robot for the future jobs (Gehrke *et al.*, 2015).

Moreover, Chryssolouris *et al.* (2013) introduce the competence model for future jobs. It includes technical, methodological, social and personal competencies. They find that top skills are media skills, coding skills, understanding IT security.

A report by The Organisation for Economic Co-operation and Development (OECD,

2016) finds that as the use of information and communications technology (ICT) permeates workplace, classrooms, auditoriums, homes, and social interactions in general, the expectation to utilize computers to manage information and to solve problems increases. Adults who are very familiar with the skills measured in the study may be able to take a full advantage of the opportunities, offered by the technological and structural changes, experienced by the modern society. Adults with a high level of literacy, computing, and problem-solving skills in a technology-rich environment tend to have better positions in the labor market than their peers. There is a high possibility for them to get a job or to earn higher wages if they already working. Employees who use information processing skills more intensively at work tend to earn more even after considering the difference in educational background and qualifications. Writing and problem solving are the skills, used most frequently in the profession.

Hershbein and Kahn (2018) found that during the Great Recession, skill requirements increased more in areas hit harder by the recession, than in other areas. Forsyth *et al.* (2020) show that job vacancies collapsed in the beginning of the Covid-19 pandemic, which was not caused only by the stay-at-home orders.

Starting from 1980, studies have found that if you are a male with college education, the probability to get high-paying position decrease. On the contrary, for women with college education these probabilities increase in comparison to men. This reflects the increasing importance of social skills for high paying jobs. Male and female salaries are also indicative of an increase in the demand for social skills (Cortes *et al.*, 2018).

Other studies have created a concept of task-specific human capital and explored gathered skills in the labour market. For university graduates, task-specific human

capital is an important source of individual salary growth (Gathmann and Schönberg, 2010).

Multiple studies have been conducted on skill requirements in job postings, and they showed that even for very specialized positions skills vary a lot. Focusing mostly on non-technical skills, the research identified positive correlations between skills, salary and company performance. Also, it was concluded that technical and non-technical skills complement each other (Deming and Kahn, 2018).

Nowadays non-tech skills are highly beneficial in the labour market. In the US labour market, levels of social relations increased by 12 percent in the period from 1980 to 2012. For jobs that require both technical and non-technical skills there was a significant salary increase during this time. Another finding was that modern labour market has more requirements for non-tech skills than it had 20 years ago, which is even more, compared to 30 or 40 years back (Deming, 2017).

Also, non-technical skills, especially people skills, were proven to have a high impact on the range of jobs and salary levels, accessible for the job-seeker. The importance of non-tech skills in the workplace was caused by organizational and technological changes (Borghans *et al.*, 2014).

In the past century employment and salaries for jobs that required high levels of both tech and non-tech skills has increased compared to jobs that require only one type of skill. And this trend in the labour market stays the same up to this day (Weinberger, 2014).

Sun *et al.* (2021) develops a neural network to predict the salary of a job position based on the required skills (and their levels) for the job, and they also estimate the value of each individual skill for the job.

Text analysis

Extensive amounts of data, being collected every day with blazingly fast, ever-accelerating speeds, cannot be handled by traditional data collection and data structuring; therefore, we cannot process and analyze the data using only traditional tools such as paper surveys or questionaries. Our research relies on modern quantitative text analysis methods (Zikopoulos *et al.*, 2011; Manyika *et al.*, 2020).

Quantitative text analysis is widely used to research job skills based on job postings data, including the descriptions of job postings (Roberts, 2000). Nowadays text analysis has become more popular and demanding in sciences. We can see this rise of interest from several books and articles, for example, Neuendorf (2002), Pennebaker *et al.* (2003), Weber (1990).

We use quantitative text analysis to analyze documents and find patterns in the texts (Humphreys and Wang, 2018). There are different kinds of analyses in the textual analysis. Content analysis, for one example, often implies creating categories and observing the presence or absence of the categories in a text or a set of texts.

It begins as a quantitative method. For example, answering the question, "How often does the word Python occur in job postings published between 2021 and 2022?" requires a researcher to count the frequency of the word in texts corpus. This method is recognized for its objectivity, but some scientists criticize it because establishing the categories involves human supervision. Even though content analysis is usually quantitative, it can become a qualitative task when findings are interpreted by a human (Rholetter, 2018).

Quantitative text analysis is widely used to investigate job skills from job postings and their descriptions and to discover skills (Meyer, 2019; Behpour *et al.*, 2021).

To conduct the research in this thesis, different methods from various fields are used, including computer linguistics (Natural language processing - NLP), information research and statistics. By means of natural language processing algorithms and statistical tools, text analytics allows to solve the problems of classification and clustering, tone analysis, selection, and aggregation of entities important for a particular study. After proper preprocessing, text and numerical information can be compared. During the processing the meaningful entities are extracted from large, complex, and unstructured texts. In such a way the unstructured texts are being transformed into structured data. It makes it possible not only to analyze the situation at a specific moment of time, but also to trace the trend and, therefore, to forecast it for the future periods of time accurately.

The knapsack problem and the maximum coverage problem

Our work in this thesis is similar to the classical knapsack problem. The knapsack problem chooses a number of items to fill a finite-capacity knapsack to maximize the total value of the selected items.

The knapsack problem is a combinatorial optimization problem. For a given set of items with given weights and values, the decision maker determines the number of each item to include in the collection (i.e., the "knapsack"), so that the total value of the selected items is maximized and the total weight does not exceed a certain limit. Similar in nature are many resource allocation problems, in which decision makers needs to choose from a given set of tasks under a finite budget/resource, with each task consuming a certain amount of budget/resource.

The knapsack problem has real-world applications such as cargo loading, project

selection, budget control, etc (Salkin and De Kluyver, 1975).

The typical formulation of the problem is as follows maximize $\Sigma_{i=1}^{N} v_i x_i$, subject to $\Sigma_{i=1}^{N} w_i x_i \leq W$, $x_i \geq 0$, and x_i integer (or x_i binary). For $i = 1, 2...N$; where v_i, w_i and W are known integers and $w_i (i = 1, 2, ..., N)$ and W are positive.

A variety of techniques have been developed over the years to solve the knapsack problem. These methods can be classified in the following categories (Martello and Toth, 1990):

(i) dynamic programming or variants thereof;

(ii) integer programming methods;

(iii) heuristic search and Lagrangian methods;

(iv) network approaches.

Our research in Chapter 4 is similar to the above-described problems in the sense that we select skills to be acquired by a job-seeker for maximizing their job coverage. However, our problem is more challenging because, unlike the knapsack problem, skills are not independent from each other as employers require a certain combination of skills.

Another closely-related problem is the maximum coverage problem. With a number of sets that may share some comment elements, one is tasked with the selection of k of those sets in order to maximize the total number of elements included in the k selected sets. The problem has been shown to be NP-hard (Feige, 1998).

We show that our job selection problem in Chapter 4 is NP hard as it includes the maximum coverage problem as a special case.

Chapter 3

Description of the data and exploratory analyses

This chapter examines patterns in the relationship between the recruiting companies and the various job skills they require.

Before the mass penetration of the Internet in public life, job searches often were made through printed media, by sending letters, and asking around the people one knows. In the era of the Internet, an increasing number of job seekers are looking for vacancies online, using various job search aggregators such as LinkedIn Jobs, SimplyHired, CareerJet, and Monster, which led their constant growth since the middle of the first decade of this century. Like online stores and ad aggregators in the market of goods and services, job search aggregator sites are marketplaces for sellers and buyers of labor skills as a specific commodity, which is a person's abilities to perform different kinds of work. Sellers are workers who offer their labor skills and buyers are employers who seeks candidates with a suitable set of skills.

In general, there are two types of skills employers are interested in: "hard skills"

and "soft skills".

Hard skills (also known as technical skills) are professional skills that can be clearly demonstrated, proven, objectively assessed or measured. Examples include mastering a computer program, a programming language, a foreign language, profession-specific skills, etc. Those skills can be obtained from courses, college, university, or acquired with work experience. Employers often look for applicants with specific hard skills as they do not require the employer to make additional investments in training.

Equally important are *soft skills* (non-technical skills). These are general qualities that help individuals interact with each other in a team, regardless of the company's business sector. Usually soft skills are not specifically tied to a position. Rather, they are more about character personal traits related to communication, organization, etc. Examples of soft skills include responsibility, communication, discipline, ability to work in a team, ability for personal time-management, the emotional intelligence, leadership skills, and critical thinking.

It is critical for job seekers to understand the patterns of required skills and qualifications by potential employers, as well as the payoff of possessing/acquiring different types of skills, so that they can better be prepared for the job market. In this research we aim to answer the following questions. Is there a difference in salaries between companies of different sizes? Does one need to have more technical skills to have a higher salary? We will answer those questions empirically based on a real-world dataset we have collected.

In this chapter we start with the description of the data collection and data preparation. First of all, we describe the methods we use to collect job postings and to fetch additional information about companies, and after that, data preprocessing

and data analysis, where, as a part of our analysis, we identify correlation between skills in Section 3.2.2. Also, apply K-means and agglomerative clustering to classify the skills involved in the data and demonstrate results of these algorithms. Then we build linear and logistic regressions to predict the salary level in Section 3.3.

3.1 Describing data

Our research in this chapter is based on the publicly available data we collected from the website ca.indeed.com.

Indeed.com is a well-known world-wide famous job search website founded in 2004. It has a powerful search engine, and in October of 2010 it was declared as the most popular job search site in the United States of America. Today the site is available in more than 60 countries in 28 languages.

The main goal is building two datasets. The first dataset will contain Job ID, Job Title, Company Name, Salary, Location, Description of the job posting, and skills with their levels, which are extracted from the description. The second dataset will consist of the company url on indeed.ca, company name, employee rating, number of reviews, company founding date, company size (number of employees), revenue, industry, and salaries of jobs, which posted by the company.

The initial data collection of job postings and companies that have posted them, is done by using two spiders. The first one collects job postings by using http-requests on the corresponding site addresses. The second spider collects detailed information about the companies on indeed.ca.

Then collected by the spiders information is verified and aggregated as it is processed linearly page by page by extraction of data from specific fields, represented by

HTML blocks. The output of the aggregation process is formatted into a table, with a sample of it shown in Table 3.1.

Company name	Job title	Country	Location	Salary	Job description
Scarsin	Jr. Data	CA	Markham, ON	57-79	The Jr Data/Reporting Analyst will be responsible for maintaining our ...
[24]7.ai	Data Scientist	CA	Toronto, ON	80-80	Brief about the Role. The Platforms team of the Data Science Group at ...
1QBit	Applied Scientist	CA	Vancouver, BC	93-109	The Platforms team of the Data Science Group at [24]7.ai builds scalable ...
36Eight Technologies	Data Scientist	CA	Vancouver, BC	77-97	Company description36Eight Technologies is a bioinformatics/ ...
3v Geomatics	Data Scientist II	CA	Vancouver, BC	71-98	Are you looking to join a tech company that can challenge and ...
407 ETR	Commercial Accounts Analyst	CA	Woodbridge, ON	46-66	Job Purpose: Responsible for collecting, analyzing data and providing analytical ...
A2Z NETWORKS	Systems Analyst	CA	Vancouver, BC	81-81	Salary: $39.00/Hourly Job Type: Full Time, Permanent ...

Table 3.1: Aggregated data, collected by the spider

There is a total of 1032 job postings and their descriptions with locations.

It is worth noting that even though the collected data contains everything necessary for the analyses, some information is contained in the text blocks, from which it must be extracted first.

Figure 3.1: Job posting example from the website

Figure 3.1 shows a job post, where the required skills are listed in a separate block. However, job posts by different companies do not have a uniform structure, and the *required skills* may not be listed in a dedicated block of the job post. Instead, they could be included in the job descriptions. In the job post below, for example, another employer mentions the necessary skills directly in the text of the ad, without listing them in a separate section. This makes extraction and comparison during the analysis more complicated.

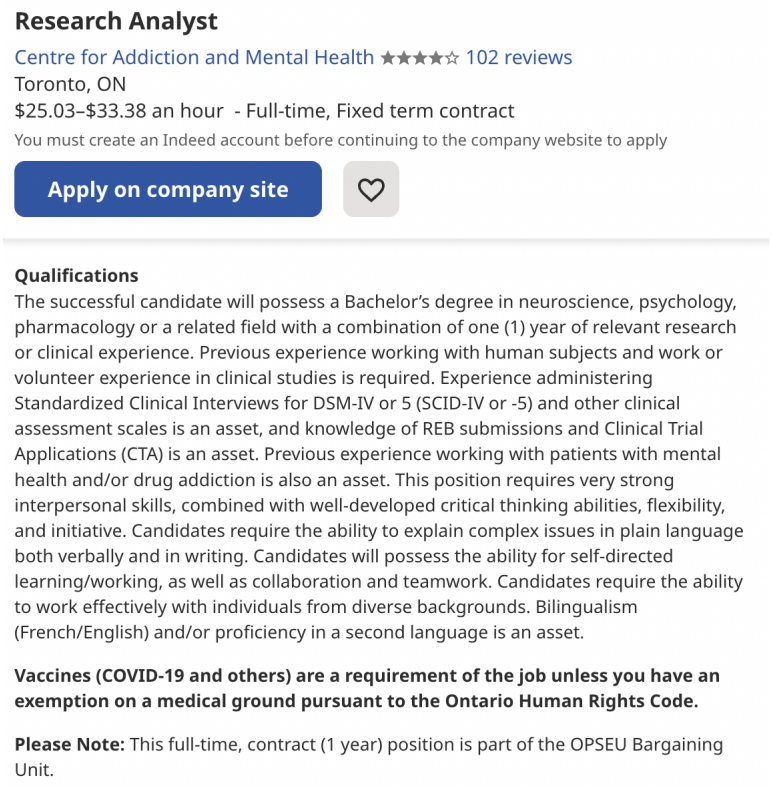


Figure 3.2: Illustration of a job posting with required skills directly in its description

Thus, one of the most important tasks of the text corpus preprocessing is to extract and group entities, which are necessary for the further analysis. These entities, as mentioned in the introduction, are the name and the skill level.

Most often, the term entity can be found when considering the patterns of the object-oriented programming paradigm. The term entity is used in the research because the project uses natural language processing libraries that use an object-oriented paradigm. In this context, entity instances are instances of the spaCy library class that describe tokens.

Another source of data will be extended information about the companies. In

order to analyze skills and their levels, we collect information about companies and will analyze if there are any interesting correlations which job seekers need to know about before applying for a job. After generating a list of unique company names in the initial dataset, the second spider can be launched. It collects detailed information about the companies on the same site. We will collect companies url, their names, ratings, number of reviews their have, their founding dates, sizes (number of employers), revenue, industry, and salaries they offer (Table 3.2).

url	Company name	Rating	Reviews count	Founded	Company size	Revenue	Industry	Salaries
ca.indeed...	Elections Ontario	4.3	209	1867	more than 10,000		Government & Public Administration	Business Analyst - 67,382 per year Senior Technical Analyst - 75,127 per year
ca.indeed...	IBM	3.9	31K	1911	more than 10,000	more than 12B (CAD)	Information Technology	Software Test Engineer - 70,523 per year Systems Administrator - 78,177 per year IT Technician - 80,916 per year
...								

Table 3.2: Aggregated companies data

The information about the number of employees in a company was used to further group all companies into several categories: from 501 to 1,000; from 1,001 to 10,000; more than 10,000 employees. We will not include companies that have fewer than 500 employees because there are not a significant amount of data for analysis, so there will be no relevant and reliable results. In total we have collected information about 306 companies and their salaries.

3.2 Data preprocessing and some summary statistics

After collecting the data, we can move on to cleaning and preparing the data for further analysis. As mentioned above, preprocessing text corpus implies doing specific standard steps common to text analytics. All necessary transformations can be performed using standard tools of the natural language processing libraries.

We process the text data by the following standard stages (Pathak, 2017; Bengfort *et al.*, 2018; Vasiliev, 2020)

(i) **filtration** (removal) of special symbols and punctuation marks;

(ii) **tokenization** - splitting text into words or combination of words – terms, used as information units for further analysis;
For example: "This is sample" → Tokenization → "This", "is", "a", "sample"

(iii) **stemming or lemmatization** - converting words to their original morphological form; we provide simple examples of stemming and lemmatization as follows.

Example of stemming:

leafs → leaf

leaves → leav

Example of lemmatization:

leafs → leaf

leaves → leaf

(iv) **removing stop-words** and low-frequency terms; processing negations.

When we create an NLP model, the complexity increases proportionally to the number of words contained in the term. If a term length is one, the created model will not consider relations between words. If we switch to bigrams or trigrams, a significant part of connections will be preserved, but the frequency of each term in the text corpus will decrease. Such a decrease may lead to deterioration of the model's quality.

The length of the token is determined by the objective of the analysis. Tokens are the result of graphematic analysis, i.e., highlighting sentences and word forms in the text, moving from symbols to words. In other words, if we extract all substrings from the text that do not contain separators (spaces, some punctuation marks, etc.), we will get an array of tokens. For each token, there is its initial (normal) form, which is also called a lemma.

In our study a particular skill is not necessarily defined by a single word, so the tokenization process identifies complete word combinations which define a particular skill. For example, skills "Python" or "Kubernetes" can be represented by a single word, while the token "think analytically" includes two words, whose separation into two terms "think" and "analytic" would distort the original meaning.

A term is an expression of a formal language (system) of a special kind. Similar to natural language, where a noun phrase refers to an object and a whole sentence refers to a fact, in mathematical logic a term refers to a mathematical object and a formula refers to a mathematical fact. In the literature and documentation, the "term" also occurs in relation to text search. In this context, the "term" can be considered synonymous with the "token". For example, in our research a *term* in a job posting could be: *Tableau, statistical analysis or Kafka.*

The task of extracting skills is quite simple, because it boils down to searching in the texts of job postings for tokens from certain sets. The training data set of skills and their levels was formed manually. This approach is also explained by the fact that there are no morphological or logical features that can make it possible to accurately highlight skills and levels. Also, there are no mentions anywhere of publicly available trained models for natural language processing libraries to label skills in job postings. So, we had randomly chosen 50 job postings. Then we have created manually a vocabulary of terms, denoting skills. First of all, an empty file is created, then the job description is read and the skills are identified. We have manually identified all the skills mentioned in the 50 job postings. Below we show an example that includes some of the identified skills.

Tableau	data migration	financial
statistical analysis	programming	collaborate
Kafka	DS	work under pressure
time management	NLP	communicating insights to non-technical
AI	GitHub	infrastructure technologies
technical communication	container	CMC

Table 3.3: Example of a part of the skill vocabulary

Then, according to the frequencies with which those skills are mentioned, we identified the 50 most frequent skills.

The levels of skills also require some certain formalization. A skill can be characterized by "ability" as well as "exceptional mastery", which changes the importance of the skill in the candidate's portfolio.

Skill level matters because a candidate who is familiar with NoSQL databases and a candidate who has knowledge of how to quickly set up a cluster of NoSQL databases cannot be evaluated equally. The next step was to standardize the level data. In order to do so, all levels were sorted and assigned numbers from 1 to 10, where 1 is the slightest knowledge required, 10 - expert knowledge required:

- 'canuse': 1,
- 'understand': 2,
- 'skill': 3, (ability)
- 'experience': 4,
- 'know': 4,

- 'familiar': 5,
- 'goodat': 5,
- 'strong': 6,
- 'effective': 6,
- 'proficient': 7,
- 'profound': 8,
- 'specialist': 8, ,
- 'excellent': 9,
- 'expertise': 9,
- 'exceptional': 10

Thus, we conclude that the list of candidate skills is a skill level pairs (proficiency level, skill). Therefore, the extraction of information about the skills of candidates should return data in this format. Below we give an example how we process data into a training dataset.

Text from job posting:

Strong command of querying and programming languages (**SQL, R and python - numpy, pandas, sklearn, TensorFlow, Keras**), and visualization tools (**Excel, Tableau, matplotlib or seaborn or plotly** in python packages

Experience applying various methods of numerical and categorical modelling techniques and supervised and unsupervised machine learning methods (OLS **regression** and GLMs, logistic regression, **KNN, SVM, decision trees or random forest, clustering** and cluster analysis, dimensionality reduction, **neural networks**)

The resulting set of skills (with their levels) is shown as follows.


```
[{"level":"strong,"skill":"SQL"},
{"level":"strong,"skill":"R"},
{"level":"strong,"skill":"Python"},
{"level":"strong,"skill":"numpy"},
{"level":"strong,"skill":"pandas"},
{"level":"strong,"skill":"sklearn"},
{"level":"strong,"skill":"Keras"},
{"level":"strong,"skill":"visualization"},
{"level":"strong,"skill":"excel"},
{"level":"strong,"skill":"Tableau"},
{"level":"strong,"skill":"matplotlib"},
{"level":"strong,"skill":"seaborn"},
{"level":"strong,"skill":"plotly"},
{"level":"experience","skill":"regression"},
{"level":"experience","skill":"KNN"},
{"level":"experience","skill":"SVM"},
{"level":"experience","skill":"decision trees"},
{"level":"experience","skill":"RF"},
{"level":"experience","skill":"clustering"},
```

{"level":"experience","skill":"ML"},
{"level":"experience","skill":"NN"}]

As we can see from this example instead of “neural network” we used “NN”. The same with “machine learning” - ”ML”, RF - random forest etc.

After labelling a small dataset, we need to apply this approach to the whole dataset. The process is time-consuming by hand, so we create a model and train it based on the training dataset.

We use the rule-based matching engine, provided by the spaCy library, which is an open-source library for advanced natural language processing (NLP - Natural Language Processing) in Python, for extraction of skills along with predefined levels from all the job postings of a complete dataset.

We remark that although the skill level pairs (proficiency level, skill) is a finite state machine, the number of such unique pairs is quite large. There are 981 pairs of values in our data.

The result of tokenization of the texts with job postings is shown in Figure 3.3. Each element from the set corresponds to one job from the text corpora at our disposal. In the model, each job posting is characterized by a set of skills along with their levels and an offered salary.

	skill_set	salary
0	[{"level":"excellent","skill":"management skills"},{"level":"understand","skill":"software development life cycle"},{"level":"understand","skill":....	99-130
1	[{"level":"skill","skill":"data architecture"},{"level":"skill","skill":"data governance"},{"level":"skill","skill":"data integration"},{"level":"...	99-117
2	[{"level":"strong","skill":"communication"},{"level":"strong","skill":"leadership"},{"level":"strong","skill":"writing"},{"level":"proficient","sk...	99-111
3	[{"level":"proficient","skill":"excel"},{"level":"strong","skill":"writing"},{"level":"strong","skill":"think analytically"},{"level":"strong","sk...	99-111
4	[{"level":"familiar","skill":"Java"},{"level":"familiar","skill":"Python"},{"level":"familiar","skill":"C#"},{"level":"familiar","skill":"Ruby"},{...	99-108
5	[{"level":"specialist","skill":"SQL"},{"level":"specialist","skill":"data warehousing"},{"level":"specialist","skill":"ETL"}]	99-106
6	[{"level":"familiar","skill":"complex software systems"},{"level":"know","skill":"software development life cycle"},{"level":"familiar","skill":"c...	99-104
7	[{"level":"proficient","skill":"SQL"},{"level":"familiar","skill":"ETL"},{"level":"familiar","skill":"regression"},{"level":"familiar","skill":"pr...	98-112
8	[{"level":"familiar","skill":"decision making"},{"level":"proficient","skill":"think analytically"},{"level":"proficient","skill":"financial"},{"l...	98-108
9	[{"level":"strong","skill":"data analytics "},{"level":"excellent","skill":"problem solve"},{"level":"specialist","skill":"excel"},{"level":"stron...	98-107

Figure 3.3: Text tokenization result

We now conduct exploratory data analysis to better understand the relationships the between the job skills required by the job postings, the employers in the postings, and the offered salaries for the posted positions.

For subsequent analysis it is necessary to encode (vectorize) the lemmatized corpus of job postings texts. First, we will apply the simplest vectorization method. For the selected group of skill level pairs, we will assign “1” if this skill is present in the given job description, and a “0” if it is absent.

The result of such encoding, similar to the OneHotEncoder from sklearn library, is shown in Figure 3.4.

	writing	verbal	sql	python	problem solve	communication	think analytically	excel	collaborate	detail oriented	...	communicating insights to non-technical	access	leadership	data management
0	1	1	1	1	0	0	0	1	1	0	...	1	0	0	0
1	1	1	0	1	0	0	0	0	1	0	...	1	0	0	0
2	1	1	1	1	0	1	1	0	0	0	...	0	0	0	0
3	0	1	1	1	0	0	1	0	0	0	...	0	0	0	0
4	1	1	0	0	0	0	0	0	1	0	...	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
394	1	1	1	1	0	1	0	0	0	0	...	0	0	0	0
395	0	0	0	0	1	1	0	0	0	0	...	0	0	0	1
396	1	0	0	1	0	0	0	0	1	0	...	0	0	0	0
397	1	1	0	0	0	0	0	0	0	0	...	0	0	0	0
398	0	0	1	0	0	0	0	1	0	0	...	0	0	0	0

Figure 3.4: The matrix - The result the texts of jobs coding

Also, some transformations were done with the companies data.

- Company of size “11 to 50 ”, “51 to 200 ”, “201 to 500” are dropped (due to insufficient number of companies of those sizes).
- “501 to 1,000” - group 1, small companies.
- “1,001 to 5,000 ”, “5,001 to 10,000” - group 2, middle size companies.
- “More than 10,000 ” - group 3, big size companies.

For some calculation the company size descriptor will be replaced by the mean value.

- “501 to 1,000 ” $\rightarrow$ 750;
- “1,001 to 5,000 ”, “5,001 to 10,000 ” $\rightarrow$ 5500s;
- “More than 10,000 ” $\rightarrow$ 20000.

Similarly, salaries will be transformed in this way.

- "46-66 " $\rightarrow$ 56
- "10-110 " $\rightarrow$ 110
- "105-157 " $\rightarrow$ 131
- "112-152 " $\rightarrow$ 132

For further analysis, we complement the dataset presented in Figure 3.3 with the additional data – information about the size of the company, and the name of the company. We get the dataset as shown in Table 3.4.

ID	Company name	Skills	Salary	Job title	Company size
0	407 ETR	[{'level': 'exceptional', 'skill': 'Python'}, ...	110-110	Data Scientist	201 to 500
1	407 ETR	[{'level': 'exceptional', 'skill': 'Python'}, ...	46-66	Commercial Accounts Analyst (12 month contract)	201 to 500
2	Accenture	[{'level': 'exceptional', 'skill': 'Spark'}, ...	112-152	Data Engineering Senior Manager	more than 10 000
3	Accenture	[{'level': 'familiar', 'skill': 'BigData'}, ...	105-157	Data Engineering Consultant	more than 10 000
...					
374	Zynga	[{'level': 'proficient', 'skill': 'AWS'}, ...	66-73	Data Analyst 3	1 001 to 5 000
375	Zynga	[{'level': 'proficient', 'skill': 'verbal'}, ...	122-122	Data Scientist	1 001 to 5 000

Table 3.4: Initial dataset for the exploratory data analysis

3.2.1 Top demanded skill

Before we start an advanced analysis, let us generate a WourdCloud (Figure 3.5) for all job postings we have collected. As we may see, the most important words in all jobs' descriptions are **"experience", "team", "support"**. We can also notice words such as **"will", "client", "including"**, etc. because we used full job descriptions to generate clouds. Later in the research, we will extract keywords (skills) for the analysis. Of course, these are very general definitions, but they give us an idea that non-technical skills are among the required skills.

Figure 3.5: Word cloud for all job postings

Moreover, we will create the WordCloud (Figure 3.6) for Data Analyst job postings. Our prior expectation was to see words "data," "analysis," "business," and some similar keywords. After generating WordCloud, we see the most frequent words are **"data", "experience", "business", "team", "work"**. It gives us another confirmation that not only technical skills are required for the technical position of data analyst. Let us look in more details.

Figure 3.6: Word cloud for the Data Analyst position

First of all, we identify the **top 10 demanded skills** by frequency, as well as the 10 most common skill - level pairs in the text corpus (skill level, skill).

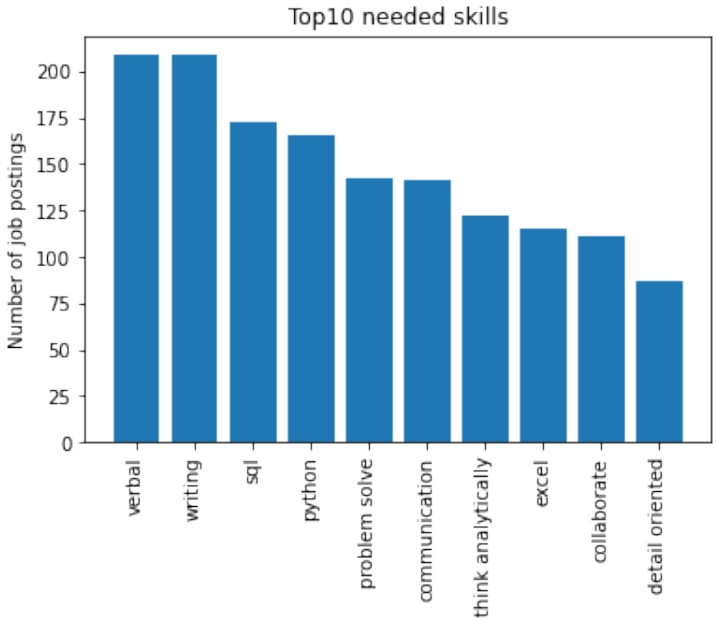


Figure 3.7: Top 10 skills

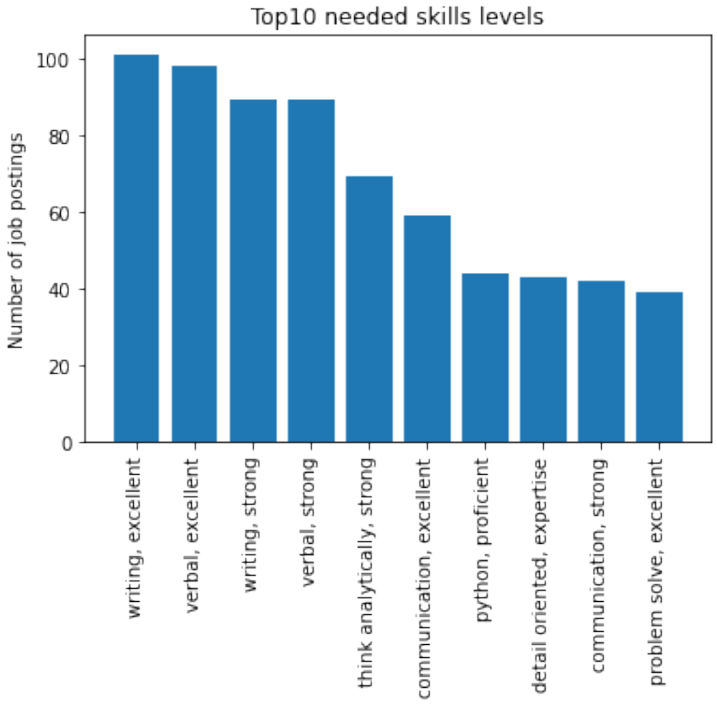


Figure 3.8: Top 10 skill level pairs describing skills

We can observe from Figure 3.8 that the skills "writing" and "verbal" both appeared twice in the top 10 list with different levels. The total amount of the top skills is 50. Among them, 37 are the hard skills and the rest (13 skills) are the soft skills (Figure 3.9)

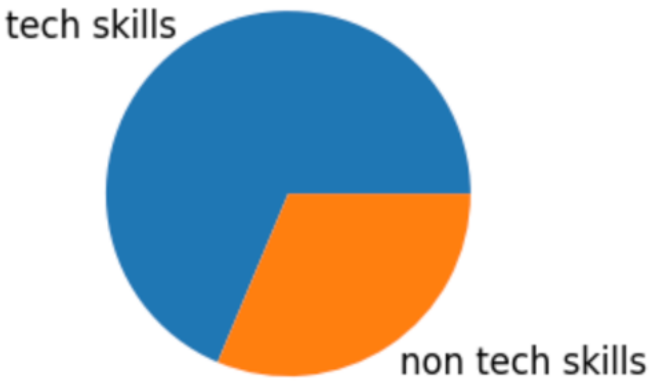


Figure 3.9: Hard vs Soft skills in top 50

Figure 3.10 shows the distribution of skills in our job postings. In average we would required around 5-6 skills per one position.

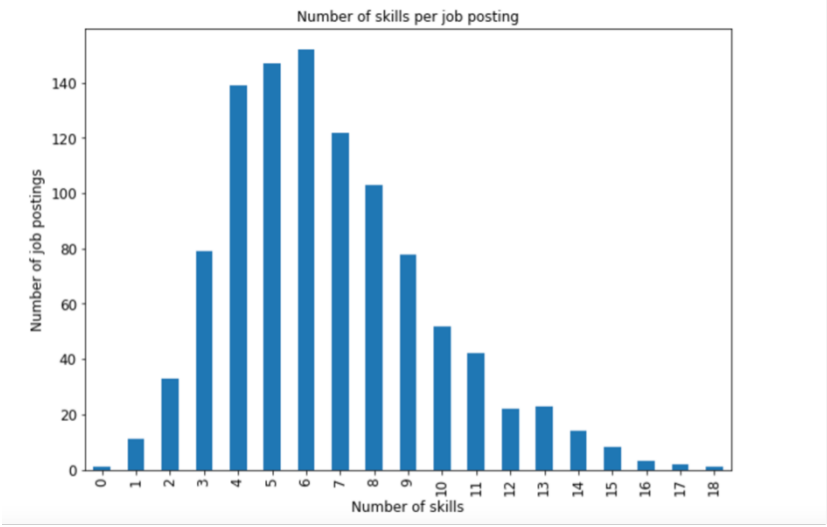


Figure 3.10: Distribution of skills

3.2.2 Correlation matrix

One of the most common and simple techniques to analyze data is to create a matrix where we can see how often different skills appear together in Table 3.5. In this table, each entry of the matrix shows the frequency of co-occurrence of two skill in the same job posting.

Similarly, in job postings where the skill "powerpoint" is included, 98% of them include "Excel" (referring to twenty fourth row and fifth column). This is intuitive because often in job description you will see Microsoft office which includes the whole suite of software, such as Word, Excel, PowerPoint, etc. As expected, many complementary skills are required by the same job. In the entry at twenty second row and fourth column we see that 90% of the job postings that require "Pytorch" also have "Python". The fourth column in eighteenth row shows us that when the job description has NLP, then 93% of the time "Python" is also mentioned. In the recent years Python has become one of the most fundamental tools in machine learning, and NLP is not an exception here. In 86% of the jobs that require "tensorflow", there is also a requirement of "Python". (e.g., this observation is demonstrated in the fourth column in the twelfth row). Also, referring to the entry in the third column of at the fourteenth row, we see that 70% of the job postings with tableau also includes "SQL".

	verbal	writing	sql	python	excel	collaborate	r	...
verbal	1.00	0.99	0.33	0.30	0.37	0.31	0.14	
writing	0.97	1.00	0.33	0.30	0.37	0.31	0.14	
sql	0.44	0.44	1.00	0.54	0.24	0.28	0.25	
python	0.42	0.42	0.57	1.00	0.16	0.26	0.33	
problem solve	0.53	0.53	0.34	0.29	0.38	0.32	0.11	
communication	0.58	0.59	0.36	0.34	0.33	0.34	0.18	
think analytically	0.52	0.51	0.29	0.25	0.32	0.36	0.09	
excel	0.53	0.54	0.26	0.17	1.00	0.31	0.14	
collaborate	0.48	0.48	0.32	0.29	0.33	1.00	0.12	
detail oriented	0.59	0.61	0.33	0.25	0.43	0.34	0.08	
r	0.49	0.50	0.67	0.83	0.34	0.27	1.00	
tensorflow	0.33	0.35	0.33	0.86	0.01	0.19	0.11	
ml	0.31	0.33	0.46	0.74	0.07	0.25	0.27	
tableau	0.50	0.52	0.70	0.56	0.37	0.26	0.37	
data analysis	0.49	0.50	0.34	0.29	0.39	0.34	0.14	
time management	0.59	0.61	0.38	0.25	0.45	0.34	0.11	
power bi	0.46	0.46	0.63	0.46	0.49	0.18	0.31	
nlp	0.36	0.39	0.21	0.93	0.00	0.32	0.11	
visualizations	0.54	0.55	0.70	0.55	0.26	0.30	0.30	
spark	0.44	0.44	0.59	0.75	0.09	0.21	0.29	
azure	0.52	0.55	0.65	0.54	0.16	0.19	0.19	
pytorch	0.45	0.45	0.19	0.90	0.00	0.21	0.10	
aws	0.43	0.44	0.55	0.62	0.13	0.25	0.20	
powerpoint	0.59	0.60	0.21	0.12	0.98	0.31	0.10	
interpersonal	0.62	0.61	0.24	0.20	0.54	0.40	0.08	
...								

Table 3.5: Skills matrix

	verbal, sql	verbal, communication	writing, sql	writing, communication	writing, sql, communication	sql, communication	...
verbal	1.00	1.00	0.98	0.98	0.97	0.56	
writing	0.98	0.99	1.00	1.00	1.00	0.58	
sql	1.00	0.34	1.00	0.35	1.00	1.00	
python	0.47	0.29	0.47	0.29	0.49	0.58	
problem solve	0.43	0.34	0.44	0.35	0.38	0.33	
communication	0.42	1.00	0.43	1.00	1.00	1.00	
think analytically	0.33	0.35	0.32	0.34	0.36	0.32	
excel	0.29	0.33	0.29	0.33	0.18	0.22	
collaborate	0.31	0.31	0.32	0.32	0.36	0.32	
detail oriented	0.27	0.26	0.28	0.26	0.17	0.17	
r	0.24	0.17	0.23	0.17	0.29	0.33	
ml	0.11	0.12	0.11	0.12	0.14	0.24	
tableau	0.30	0.16	0.30	0.16	0.32	0.30	
data analysis	0.15	0.18	0.14	0.18	0.21	0.15	
time management	0.15	0.11	0.15	0.12	0.08	0.12	
power bi	0.17	0.12	0.16	0.12	0.21	0.22	
visualizations	0.23	0.16	0.23	0.16	0.31	0.24	
...							

Table 3.6: Skills matrices after a few iterations

Table 3.6 shows the frequencies of three skills required in the same job posting. For example, referring to the entry at the first row and the fourth column, we see that 98% of the job postings requring "verbal" will also require both "writing" and "SQL". Also, we can see that if the job description requires "writing", "SQL", and "communication", then with a 97% chance it will ask for verbal skills (referring to the first row and fifth column).

The correlations shown in Tables 3.5 and 3.6 is useful for job seekers to identify skills that are valuable as a combination. But it does not give us any information for how many jobs one can apply for if equipped with those skills. We will try to find an answer in the Chapter 4.

3.2.3 Clustering of skills

We now conduct clustering of the skills, which can be useful for finding a group of skills that could be learnt together and would be beneficial for job seekers compared to learning different unconnected skills. Clustering of skills is an alternative way of finding a winning combination of skills, which would help cover as many jobs as possible.

K-means model

Using the K-means model we do the clustering of the skill vectors. Firstly, we provide a short overview of this algorithm.

The algorithm initializes the k clusters by placing one input point in each cluster Then it places each of the remaining points into the clusters one at a time. It places each point in the cluster, whose centroid is closest to the point A. One further step could be at the end of the algorithm to fix the centroids and start over the algorithm

of assigning all points to the centroids for robustness:

Algorithm 1 The k-Means Algorithm (see, e.g., Hartigan (1975))

Input: Point set $P \subseteq \mathbb{R}^d$
number of centers k
Choose initial centers $c_1, ..., c_k$ of from $\mathbb{R}^d$
repeat
$P_1, ..., P_k \leftarrow \emptyset$
for each $p \in P$ **do**
Let $i = argmin_{i=1,...,k} \|p - c_i\|^2$
$P_i \leftarrow P_i \cup \{p\}$
for $i = 1 \rightarrow k$ **do**
if $P_i \neq \emptyset$ **then** $c_i = \frac{1}{|P_i|} \sum_{p \in P_i} p$
until the centers do not change

We use the elbow plot (Figure 3.11) is used to determine the number of clusters for K-means clustering model.

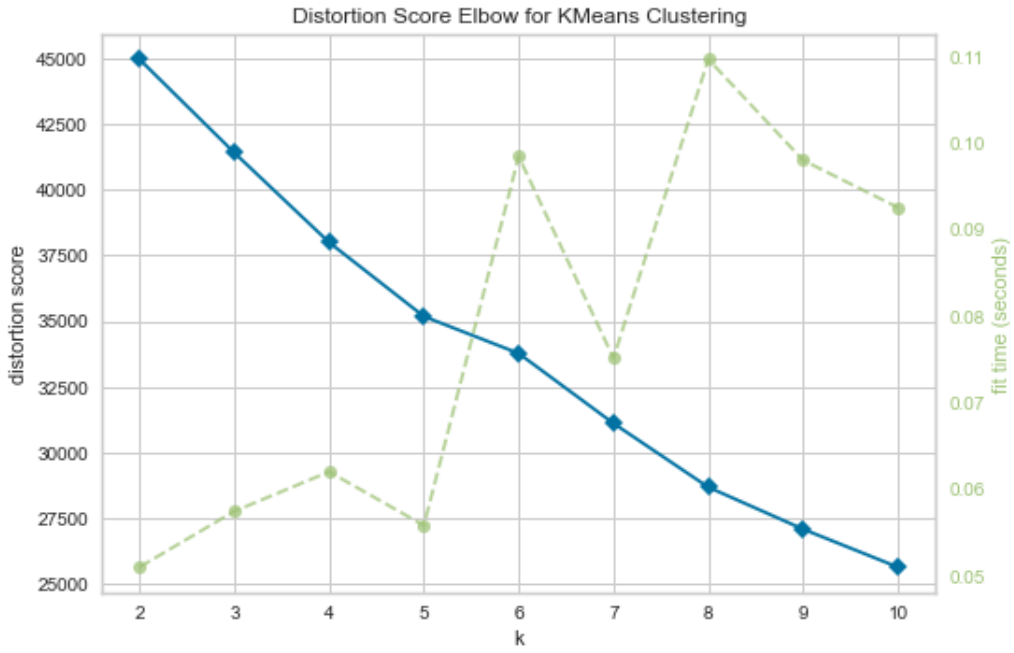


Figure 3.11: Elbow plot for the K-means clustering model

The elbow plot shows how the goodness-of-fit improves as one increases the number of clusters. Based on the plot, we determine the number of clusters as 5, at which the improvement of the goodness-of-fit slows down as we further increase the number of clusters. In this case, we will be grouping skills this way.

Cluster 1 includes the following skills excel, powerpoint, access, in Cluster 2 includes: r, ml, ai. Cluster 3 will contain: detail oriented, data analysis, time management, interpersonal, presentation, databases, data analytics, financial, project management, agile, communicating insights to non-technical, pytorch, leadership, data management, research, sas. In the Cluster 4 we include: tableau, power bi, visualizations, azure, spark, aws, hadoop, java, cloud software services, modeling, tensorflow, statistical analysis, c++, bigdata, oracle, think critically, scala, hive, ds, nlp. And in the last Cluster 5 are - writing, verbal, sql, python, problem solve, communication, think analytically, collaborate.

The obtained clustering results can be interpreted as follows. Each cluster contains skills specific to a particular group of jobs. The table below shows the correspondence between clusters and groups of job postings.

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Operational analysts	Artificial Intelligence	Financial analysts and managers	Business analysts and Data engineers	Junior analysts

Table 3.7: Initial dataset of the feature ranking of companies

Agglomerative Clustering

The agglomerative clustering is one of the clustering models which are used often for categorical features' set. K-Means is simple and gives good results while applying

on numerical data. We show the elbow plot for the agglomerative clustering model in Figure 3.12, which help us determine the number of clusters as 4. We also present the clusters resulting from this model.

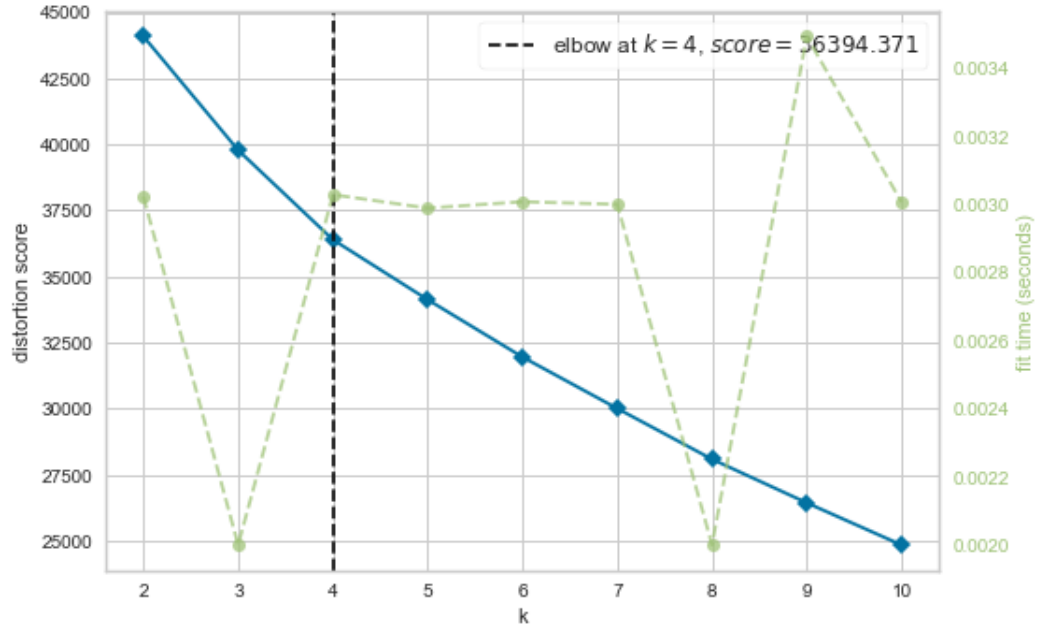


Figure 3.12: Elbow plot for Agglomerative clustering model

Cluster 0 includes the following skills: communication, think analytically, excel, r, ml, data analysis, visualizations, spark, azure, interpersonal, databases, java, cloud software services, modeling, tensorflow, c++, agile, communicating insights to non-technical. In Cluster 1 there are financial, hadoop, oracle, think critically, access, scala. Cluster 2 includes: sql, tableau, power bi, presentation, ai, project management, statistical analysis, pytorch, leadership, hive, research, sas, ds, nlp. Cluster 3 includes verbal, writing, python, problem solve, collaborate, detail oriented, time management, aws, powerpoint, data analytics, big data, data management.

Below we can see (Table 3.8) frequencies of all skills calculated for the job postings

corpus. It is evident that many non-tech skills have high frequencies, and it can have a strong effect on the clustering results. This observation led to the modeling of clusters without non-tech skills.

Skill	Frequency
writing	499
verbal	496
sql	496
problem solve	378
communication	346
python	345
collaborate	324
excel	312
think analytically	306
detail oriented	222

Table 3.8: Skills frequencies

Finally, the clustering into 5 clusters without non-tech skills will be used. The clustering result will be used later in this research in Chapter 4.

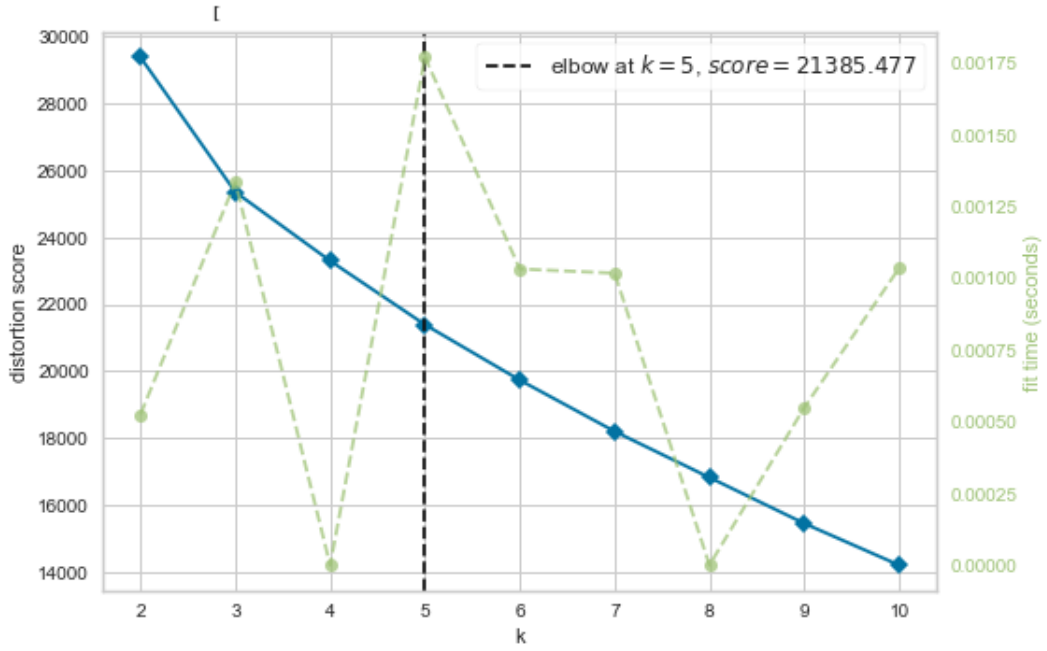


Figure 3.13: Elbow plot for Agglomerative clustering model without non-technical skills

Cluster 0 includes the following skills: power bi, visualizations, data analytics, oracle, access, research. Cluster 1 has: excel, azure, aws, databases, java, tensorflow, pytorch, data management, hive, ds. Cluster 2 includes: scala, sas. In Cluster 3 there are sql, python, r, ml, data analysis, spark, powerpoint, c++. And in the last Cluster 4 there are tableau, financial, hadoop, ai, cloud software services, modeling, statistical analysis, bigdata, nlp.

There are many important differences between K-means and hierarchical clustering. They lie in how these two algorithms are implemented to how the results are interpreted.

The K-means algorithm uses k value, representing the number of clusters to create. The algorithm starts by creating k centroids. It then repeats the assignment step

(each sample is assigned to the nearest centroid) and the update step (each centroid is updated to be the average of all the samples assigned to it). This iteration continues until it meets the stopping criteria, such as when the sample is not reassigned to another centroid.

Agglomerative hierarchical clustering, instead, builds clusters incrementally, producing a dendrogram. The algorithm begins by assigning each sample to its own cluster. At the next steps, the two clusters that are the most similar are merged. Unlike K-means, there is no need to specify a k parameter: once the dendrogram is created, you can select a layer in the tree to see how many clusters are best suited for your particular application.

As we observe, the two different clustering methods give different results for the same data set. Therefore, obtaining different results when using clustering methods of two different basic types (non-hierarchical and hierarchical) is not extraordinary.

Moreover, even when applying the K-Means method sequentially several times, different results may be obtained, as the method is sensitive to the initial selection of points and centroids.

3.3 Technical vs. non-technical skills for companies of different sizes

We will now answer the following question: does the size of a company impact the type of skills required by the company? Will a bigger company have more focus on technical or non-technical skills, or the size of a company does not affect skill requirements?

Each job posting in our list is characterized by the categorical feature of the company size, the offered salary, and skills (from the top 50 technical and non-technical skills list) it requires.

First, for companies of different sizes, we investigate the share of required technical skills in the set of all required skills, as opposed to the fraction of required non-technical skills. In the following figure, we group companies into three categories based on their sizes and calculate the fractions of technical and non-technical skills required for the jobs posted by companies in each category. For example, among the 68 skills the job postings posted by companies of size 501 to 1000, 63.24% of them are technical skills and the remaining 36.76% are non-technical skills.

company size	**number of skills**	**technical skills**	**non-technical skills**
501 to 1 000	68	0.6323	0.3676
1001 to 10 000	647	0.6290	0.3709
more than 10 000	1786	0.6131	0.3868

Table 3.9: The ratios of technical and non-technical required skills

From the results, we observe that in all three categories, the technical skills have higher weight than non-technical skills. However, large companies require non-technical skills more frequently compared to small and medium-sized companies. This may be an indication that larger companies are more likely to emphasize non-technical skills due to their higher need in effective communication, teamwork, etc.

We further investigate the correlation between the size of a company, the average salary it offers in its job postings, and the share of technical skills among all required skills.

First of all, let us build a correlation matrix of the attributes.

	company size	**salary average**	**share of technical skills**
company size	1.0000	0.0832	-0.0183
salary average	0.0832	1.0000	0.2038
share of technical skills	-0.0183	0.2038	1.0000

Table 3.10: Correlation matrix of the constructed feature

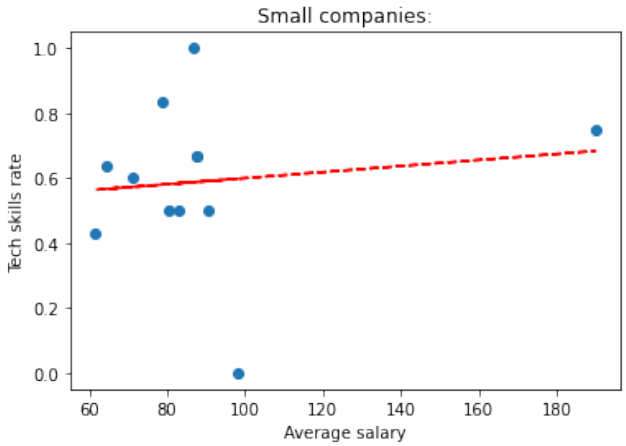
As we can see from the resulting correlation matrix, the share of technical skills falls in company size, as we observed earlier. Average salaries, by contrast, grow not only in company size, but also with the share of technical skills, implying that companies are willing to pay more to specialists with existing technical skills.

More formally, we fit a linear regression model to quantify the above-mentioned correlation. The model and its estimated coefficients are shown in the following equation. We see that the average wage is correlated with the share of technical skills to a higher magnitude.

$$Tech\ skills\ share = 0.462 - 1.24e^{-6} * \ Company\ size - 1.89e^{-3} * Avg\ salary \quad (3.1)$$

We further analyze the correlation between salary level and the technical skills share for each group of companies controlled by their sizes in Figure 3.14.

We can observe that in all three segments the average salary increases as share of technical skills in the requirements for candidate increases. At the same time, the angles of slope for the regression line are different for different size groups. We observe



(a) Small companies

(b) Mid-size companies

(c) Big companies

Figure 3.14: Segmentation results by the company size

that the share of technical skills is positively correlated with the level of wages to a greater (smaller) extent for small (larger) companies.

3.3.1 Company ratings vs. salary level

In this section, we will answer the following question: Is there a correlation between company rating and its salary level?

To perform the analyses, we consider yet another numerical attribute that can be obtained from the originally collected data – the rating of the company by its employees and examine how it is correlated with the salary level offered by the company.

The rating is expressed as a number from 1 to 5 with one decimal place. We rank the salaries by assigning them a level from 1 to 4. The level 1 corresponds to the category of low wages, while level 4 corresponds to the category of high wages. The grouping of the target variable will allow us to study the dataset not only by regression analysis, but also by constructing a classifier. In the classifier the following features are included

- Features: rating, tech skills, non - tech skills
- Target: average salary

The format of the initial dataset for the described study is shown in Table 3.11

	rating	number of technical skills	number of non-technical skills	salary avarage
0	3.8	4	56.0	1.low
1	3.8	6	110.0	4.high
2	4.0	4	131.0	4.high
3	4.0	2	132.0	4.high
4	4.0	5	37.0	1.low
...	...	...	...	...

Table 3.11: Initial dataset of the feature ranking of companies

We first use a linear regression model to examine the correlation between the average salary level and the attributes in the first three columns of Table 3.11.

Symbolically, the linear regression equation is written in the following form, with dependent variable y representing the average salary level, and the x_i's representing the rating, the number of technical skills, and the number of non-technical skills.

$$y = \Sigma i = 1^n a_i x_i \tag{3.2}$$

We fit the linear regression model using sklearn, and show that a 1-point increase in a company's rating corresponds to an increase in salary by an average of 14,000 CAD per year. An increase in the number of technical skills by 1 increases the salary by an average of 1680 CAD per year, and an increase in the number of non-technical skills by 1, on the contrary, reduces the salary by an average of 1110 CAD per year. We must clarify the last point. Even though an increase of a non-technical skills by 1 showed average salary decrease, this is related to the fact that many entry-level

positions require mostly non-technical skills, and the more non-technical skills are present in the ad, the more likely it is to be an entry-level position.

Given that we have categorized the average salary offered by companies into four classes (i.e., "1. low", "2. below average", "3. above average", and "4. high"), the prediction is essentially a classification problem, for which methods such as logistic regression can be applied. Thus, we now build a multinomial logistic regression model to predict the salary level based on the attributes in the first three columns of Table 3.11.

The multinomial logistic regression predicts the probability for the salary to be belong to the four classes as follows:

$$\Pr(y = 1) = \frac{1}{1+e^{-w^T x}}$$

$$\Pr(y = 2) = \frac{w_1 x_1}{1+e^{-W^T X}}$$

$$\Pr(y = 3) = \frac{w_2 x_2}{1+e^{-W^T X}}$$

$$\Pr(y = 4) = \frac{w_3 x_3}{1+e^{-W^T X}}$$

In the above equations, the independent variable $X = (x_1, x_2, x_3)$ is the vector of the three attributes in the first three columns of Table 3.11., and $W = (w_1, w_2, w_3)$ is the vector of weight parameters to be fitted.

We fit the mode using the sklearn library and obtain the fitted parameters W as follows.

$$W = \begin{pmatrix} -0.824 \\ -0.084 \\ +0.089 \end{pmatrix}$$

Thus, the fitted probability of a low salary (i.e., the first class of salary levels) is given as follows. for the first class's belonging probability is

$$\Pr(y = 1) = \frac{1}{1+e^{0.824x_1 - 0.0884x_2 + 0.089x_3}}$$

According to the fitted parameters of the logistic regressions, we find the following correlations. An increase in a company's rating by 0.1% is concurrent to a reduction of the probability of a low salary by 6.3%, but at the same time is concurrent to an increase of the probability of a high salary by 1.7%; Increasing the number of technical skills by 1% is concurrent to a reduction of the chance of a low salary by 10% and an increase of the chance of a high salary by 14%; Increasing the number of non-technical skills by 1% is concurrent to an increase of the chance of a low salary by 13% and a decrease if the chance of a high salary by 15%.

We now conclude this chapter by summarizing the results.

1. The share of technical skills in candidate requirements drops as company size increases see Table 3.9.

2. Average salary grows with both the growth of the share of technical requirements and with the growth of company size see Table 3.10.

3. Increase in the number of technical skills by one unit will give a candidate a larger salary increase in a smaller company compared to larger company see Figure 3.14).

4. The decreasing number of required technical skills as company size increases see

Table 3.10) can be explained by the fact that large companies in the IT sector more often advertise non-technical positions compared to small companies.

5. From the regression model (3.1) in Section 3.3 it can be concluded that a candidate with more technical skills is more likely to get the highest possible salary when employed by a large company.

6. From linear regression in Section 3.3.1 we can say that salaries are higher in companies that have higher review ratings.

Chapter 4

An optimal skill selection problem

As the economies in North America and other parts of the world recover from the Covid-19 lockdowns, many of them have experienced significant labor shortage. In Canada record number of 915,500 job postings was registered in the fourth quarter of 2021 (Statistics Canada, 2021). Despite of the labor shortage, many still struggles to find a job. Such a phenomenon, sometimes referred to as a "jobless job boom", is caused by mismatches in demand and supply of skills, geographic locations (of employers and job seekers), and expectations (in salary, flexibility, etc. Lichtenberg (2021)). In this chapter, we focus on the mismatches in skills and ask the question of how to reduce mismatches from a job seeker's perspective.

To be better prepared for entering the job market and eventually landing a job, it is critical for job seekers to acquire skills related to the jobs they hope to get. But how should a job seeker decide which skills to learn to achieve a good outcome? To answer this question, we develop a data-driven model that takes advantage of the abundant data from online marketplaces to study the problem of optimally selecting skills to acquire. The objective of our model is to maximize a job seeker's exposure to

the number of jobs they are eligible to apply for. Given that an individual can only add a few skills to their portfolio during a limited period of time and usually possesses a relatively small number of skills at the same time, we will consider a cardinality constraint on the number of skills one can have. We will show that our problem is NP-hard.

Our skill-selection model in this chapter is also applicable to curriculum design problems by educational institutions, which often have the same goal of better preparing learners for the job market as the individual learners. When developing a course or a specialization, educators need to determine the materials to include, which is often based on the skills they intend to let learners learn.

4.1 Model Formulation

We now formulate the problem as an optimization model from the perspective of a job seeker. Let $S = \{s_1, s_2, ...s_n\}$ be a set of n skills related to the job positions that the job seeker is interested in.

Let $I = \{J_1, J_2, \ldots, J_m\}$ be the set of job positions interesting to the job seeker. We view each job position $J_i \in I$ as a set of skills; i.e., $J_i \subseteq S$. Each skill $s \in J_i$ is a required skill for the job position J_i. To be eligible to apply for a position J_i, the job seeker must acquire all skills required by J_i, in which case we say that J_i is covered.

Suppose that the sets S and I, and which skills are required for each position are known. The job seeker needs to select up to a total number of K skills from S, with the objective of maximizing the total number of job positions that are covered. Equivalently, the job seeker may select the job positions to cover, which will prescribe the set of skills to acquire. Suppose that they intend to cover a subset

$V = \{J_{i_1}, J_{i_2}, \ldots, J_{i_l}\} \subseteq I$ of job positions. Then all the skills in $J_{i_1} \cup J_{i_2} \cup ... \cup J_{i_l}$ must be selected. Let us denote $|C|$ as the cardinality of a set C. Then we have the following (combinatorial) formulation of the problem.

$$V = \max_{\{J_{i_1}, J_{i_1}, \ldots, J_{i_l}\} \in I} |V| \tag{4.1}$$

$$|J_{i_1} \cup J_{i_2} ... \cup J_{i_l}| \leq K \tag{4.2}$$

Despite being compact, the above formulation is non-trivial to solve. To be able to solve the model efficiently, in the following we will reformulate it as a linear integer programming model.

We define an incidence matrix $A = (a_{is})_{m \times n}$ with its entry a_{ik} equal to 1 if job position J_i requires skill s and 0 otherwise.

Let the binary variable $x_s \subseteq \{0, 1\}$ to represent whether skill s is selected and define $x = (x_1, \ldots, x_n)$.

If J_i is covered, then we must have $x_s = 1$. Since $a_{is} = 1$ for any skill s required for job position J_i, we have $x_s \geq a_{is}$ if J_i is covered. On the other hand, if $x_s \geq a_{is}$ for all s, then for any skill s required for J_i we must have $x_s = 1$ (since $a_{is} = 1$ and $x_s \geq 1$ will imply that $x_s = 1$). Therefore, we see that J_i is covered if and only if

$$x_s \geq a_{is} \; for \; all \; s = 1; , \ldots, n \tag{4.3}$$

If we let $y_i \in \{0, 1\}$ represent whether job J_i is covered, then the above constraint can be written as

$$x_s \geq a_{is} y_i \ for\ all\ s = 1;, \ldots, n \tag{4.4}$$

We note that the above constraint is also satisfied by any J_i that is not covered, since in that case the right-hand-side of the constraint will be 0 and the constraint holds trivially.

We also have the cardinality constraint on the number of skills to include

$$\Sigma_{s=1}^{n} x_s \leq K \tag{4.5}$$

The objective is to maximize the total number of matching job postings

$$max\ \Sigma_{i=1}^{m} y_i \tag{4.6}$$

Thus, we obtained the following optimization model

$$max\ \Sigma_{i=1}^{m} y_i$$

$$\Sigma_{s=1}^{n} x_s \leq K \tag{4.7}$$

$$x_s \geq a_{is} y_i \ for\ all\ s = 1;, \ldots, n\ and\ all\ i = 1, \ldots, m.$$
$$x_s, y_i \in \{0, 1\}\ for\ all\ s = 1;, \ldots, n\ and\ all\ i = 1, \ldots, m.$$

We establish the equivalence of the linear integer programming mode and the combinatorial formulation (4.2) in the following proposition. The proof of this proposition is apparent from the above analysis and is, thus, omitted.

Proposition 4.1 The formulation (4.7) is equivalent to the formulation (4.2).

We further show the hardness of the skill-selection problem.

Proposition 4.2 The skill-selection problem (4.7) is NP-hard.

Proof. We will show that a special case of the problem becomes the maximum coverage problem (see, e.g., Hochbaum and Pathria (1998)), which is known as an NP-hard problem. Consider the case in which each job position only requires one skill. In this case, each skill s can cover one or more jobs in the set of jobs I and therefore can be viewed as the subset of jobs it can cover, which we denote as I^s. Then, the problem can be described as to select K subsets from $I^1, \ldots, I^n$, say $I^{j_1}, \ldots, I^{j_K}$, such that the total number of elements (i.e., job positions) in those subsets, $|I^{j_1} \cup, \ldots, \cup I^{j_K}|$, is maximized. This is exactly the maximum coverage problem. ■

4.2 Solving the skill-selection problem

We will solve the skill-selection problem using IBM's CPLEX solver (specifically, we call CPLEX in Python using the library "cvxpy"). Several numerical experiments based on our real-life data are conducted to illustrate how our model can help a job seeker select skills optimally and understand the value of possessing/acquiring more skills.

4.2.1 Selecting from the 50 most common skills

We conduct our first set of numerical experiments based on our dataset of 1032 job postings and 50 most common skills (see Figure 3.4 in Chapter 3 for an illustration of how the dataset looks like). In this part of the research, we will not take into account the level of skills (in other words, we consider the same skill of two different levels

simply as one skill).

We solve the problem for different values of K. In Figure 4.1, we report the maximum number of job postings that can be covered by selecting K skills

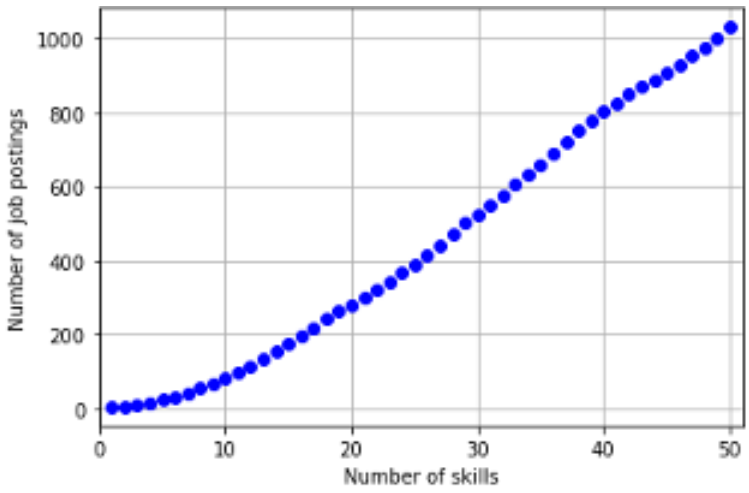


Figure 4.1: Number of skills vs Number of Job postings

If the job seeker can select up to 5 skills, the optimal set of five skills will allow them to cover 24 job postings. An increase of the skill number to 10 increases the number of job postings to 83. These numbers fully correspond to the dependence shown in Figure 4.1. Overall, the number of jobs that can be covered grows with the number of skills K in an approximately linear manner. However, initially, the slope of the curve is quite small, but gradually increases, keeping the growth up to 16 skills. After 16 skills the rate decreases slightly but it is still positive.

The second inflection point is observed at the amount of 40 skills. After this amount, growth slows down further. It suggests that out of the selected 50 skills, 16 are highest frequencies, 40 of the skills have higher frequencies than the other 10.

When selecting 5 skills, the largest number of positions can be covered with the following skill-set:

writing, problem solve, communication, think analytically, verbal

The above set of skills is a set of soft skills, which mainly corresponds to the position of an entry-level data analyst. This can be explained by the fact that there are quite a lot of job postings for entry-level data analysts, which do not require much technical skills.

	verbal	writing	sql	python	problem solve	communication	think analytically	excel	collaborate	detail oriented	...	communicating insights to non-technical	pytorch	leadership	data management
18	1	1	0	0	0	1	0	0	0	0	...	0	0	0	0
30	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0
183	1	1	0	0	0	0	0	0	0	0	...	0	0	0	0
252	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
253	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
254	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
255	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
333	0	0	0	0	1	1	1	0	0	0	...	0	0	0	0
387	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0
409	1	1	0	0	0	0	0	0	0	0	...	0	0	0	0
443	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0
482	1	1	0	0	1	1	0	0	0	0	...	0	0	0	0
485	1	1	0	0	1	1	0	0	0	0	...	0	0	0	0
598	0	0	0	0	1	1	1	0	0	0	...	0	0	0	0
603	1	1	0	0	0	1	1	0	0	0	...	0	0	0	0
604	1	1	0	0	0	1	1	0	0	0	...	0	0	0	0
703	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0
728	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
729	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
730	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
979	0	0	0	0	1	0	1	0	0	0	...	0	0	0	0
1010	1	1	0	0	1	1	0	0	0	0	...	0	0	0	0
1011	1	1	0	0	1	1	0	0	0	0	...	0	0	0	0
1027	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0

24 rows × 50 columns

Figure 4.2: Jobs that can be covered by 5 selected skills

Selecting 10 skills gives the following result:

verbal, writing, sql, problem solve, communication, think analytically, excel, collaborate, detail oriented, financial

Soft skills still prevail here. Direction of data analysis remains the same, now supplemented by financial analysis skills. Java or Python are still not among these skills. It also allows us to conclude that specific hard-skills are less common than soft-skills. A proper extension of the analysis is to build an optimization model based on a set of hard skills only. This research can produce a straightforward manual for a candidate. A set of hard skills plus a set of professional soft skills can also be used.

	verbal	writing	sql	python	problem solve	communication	think analytically	excel	collaborate	detail oriented	...	communicating insights to non-technical	pytorch	leadership	data management
18	1	1	0	0	0	1	0	0	0	0	...	0	0	0	0
30	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0
183	1	1	0	0	0	0	0	0	0	0	...	0	0	0	0
252	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
253	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
254	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
255	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
333	0	0	0	0	1	1	1	0	0	0	...	0	0	0	0
387	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0
409	1	1	0	0	0	0	0	0	0	0	...	0	0	0	0
443	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0
482	1	1	0	0	1	1	0	0	0	0	...	0	0	0	0
485	1	1	0	0	1	1	0	0	0	0	...	0	0	0	0
598	0	0	0	0	1	1	1	0	0	0	...	0	0	0	0
603	1	1	0	0	0	1	1	0	0	0	...	0	0	0	0
604	1	1	0	0	0	1	1	0	0	0	...	0	0	0	0
703	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0
728	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
729	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0

Figure 4.3: Sample of jobs that can be covered by 10 selected skills

4.2.2 Selecting from all skills

For solving the optimization problems described above we have used only the top 50 skills. Will be there any difference if we include all unique skills from our dataset?

So, let us run this model, including all of the skills.

If model is constrained to return a maximum of 5 skills, the result is the same as for the dataset with 50 skills:

writing, problem solve, communication, think analytically, verbal

But when constrain the optimization model to 10 skills, the result is:

detail oriented, collaborate, communication, think analytically, problem solve, writing, excel, financial, verbal, data analysis

It covers 51 jobs. Interestingly, there is only one difference – skill 'data analysis' replaces 'sql' (compared to dataset with top 50 skills), and the covered jobs reduces from 83 job postings to only 51, since by considering a larger set of all skills, more skills are needed to cover a job.

Again, we also consider the case where skill levels are considered. At the skill levels of not less than 4, 5 and 6 nothing changes, compared to dataset, which includes 50 skills.

Another finding is that after changing the threshold to skill levels no lower than 7, we received the following 10 skills:

detail oriented, collaborate, communication, time management, think analytically, problem solve, writing, excel, financial, verbal

One difference you can see is the addition of the 'time managent' skill. Moreover, these 10 skills cover 181 job postings. That is because a candidate with the higher level of skills could cover both junior and middle positions, while time management skill is a good precursor for the candidate to be fit for the top-level positions.

Our results in this section is only slightly different from the result in Section 4.2.1 This shows that by focusing on the top 50 common skills, our model selects a near-optimal set of skills for maximizing the number of jobs covered.

4.2.3 Selecting hard skills

Returning to the labor market trends, let us remember that hard skills allow hiring an employee who will immediately start working and does not need an extensive training, full compliance with soft-skills cannot be sufficient for the majority of IT job postings. Let us further filter the dataset under consideration by hard-skills and repeat the same steps that were done for solving the optimization problem in 4.2.1 (where it was done without filtering out any types of skills, hard or soft).

We now let the universe of skills include only all the hard skills. By optimally selecting 5 skills from all hard skills, we get the following set:

spark, hadoop, scala, hive, nlp

The resulting sample, unlike the ones shown above, can be a direct guide to action. Three of the five skills refer to the implementation of distributed processing of unstructured and weakly structured data. Hive and Hadoop refer to the Hadoop platform, and Spark is an alternative framework to solve similar problems. Thus, three of the five skills relate to distributed computing. This allows us to make an

unambiguous conclusion about the demand and good perspectives of employment for candidates with these skills. Inclusion of those skills are consistent with the general trends of the IT sector toward big data analytics, thus proving the relevance of our optimization model.

Another one of the selected skills, is Scala, which is a programming language with a growing share in software development and a high level of satisfaction by developers using it. The last selected skill is NLP - natural language processing, which is also one of the modern trends. Natural language processing is used everywhere, from distributed text search, which has already been discussed, to chat bots, autoresponders, and voice assistants. Therefore, it is also a practical recommendation.

Thus, we can conclude that the selection of hard-skills by our model reflects the demand of the IT-sector of the job market.

The dependency of the number of job postings suitable for the candidate on the number of skills in the case of selection by hard skills will be different, as can be seen in Figure 4.4. In contrast with Figure 4.1, we may observe the diminishing return to scale effect for the number of jobs that can covered by increasingly many hard skills. Intuitively, this is because many of the job positions in the IT sector share a common set of hard skills. As K becomes larger, the model will be able to select most of common skills. By further increasing K, only the less demanded skills will be added, limiting the additional increments in job coverage.

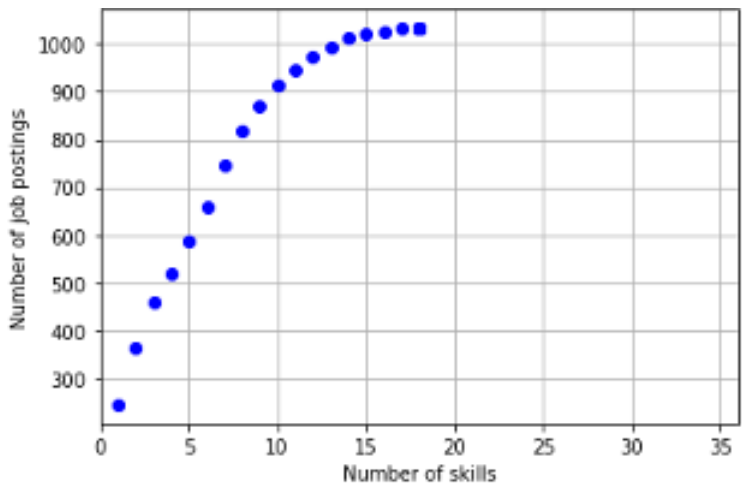


Figure 4.4: Number of skills vs Number of Job postings (hard skills selection)

That being said, we observe that even with less than 5 hard skills a candidate can be qualified for many jobs. The top 5 hard skills are as follows:

sql, python, excel, nlp, data analysis

This quite accurately reflects the trend of data analysis from different angles: data collection and grouping using sql, data analysis using python or excel, and the skill of data analysis in general. NLP is still in the trends.

	sql	data analytics	python	excel	powerbi	database structure	non-relational databases	analytical	quantitative research	tableau	...	ms o365	data analysis	oracle	vba	word	statistical	google analytics
0	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
4	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
5	1	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1025	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
1026	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
1027	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
1028	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
1031	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0

587 rows × 35 columns

Figure 4.5: Jobs that can be covered by 5 selected skills

We can conclude that employers specify hard-skills much more precisely, compared to soft-skills. Indeed, if we are not talking about DevOps or senior level positions, the requirements for the candidates contain a small number of skills related to specific tools and technologies. Soft skills are more numerous. As shown in Figure 4.6, most jobs mention only one or two soft skills, although different jobs could ask for quite different soft skills. Therefore, it is necessary to monitor trends and update skills mainly based on hard skills, but without forgetting completely about the basic soft skills.

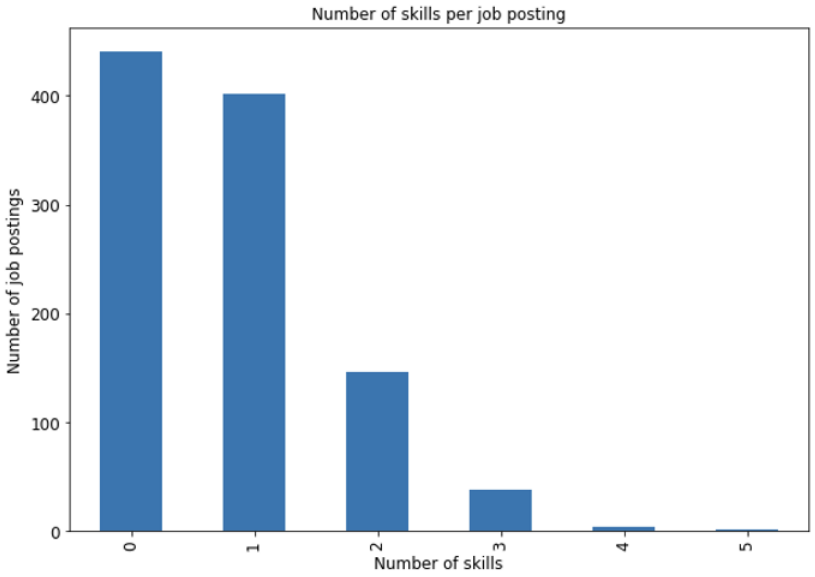


Figure 4.6: Number of soft skills mentioned in the job postings' descriptions

4.2.4 Selecting soft skills

We now focus on selecting soft skills. The dependency of the number of matching (covered) job postings on the number of available soft skills also confirms the pattern in Figure 4.6.

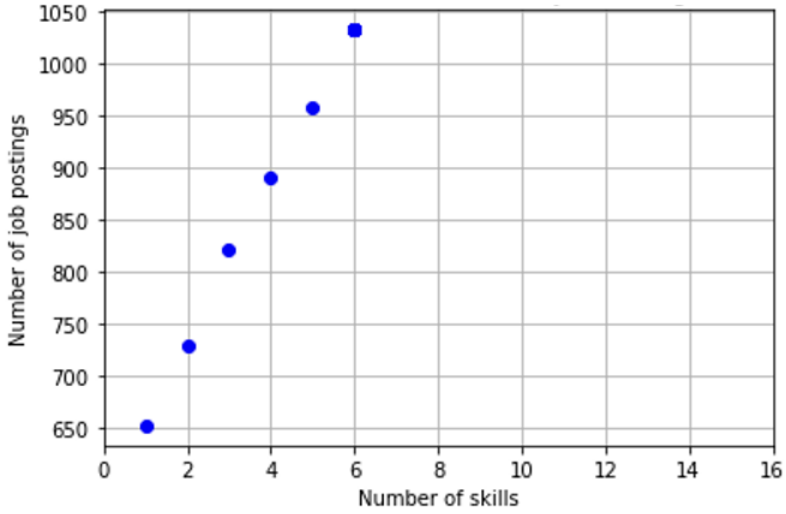


Figure 4.7: Dependence of the number of closed positions on the number of skills

With only two skills a candidate matches more than 700 job postings, and with four skills a candidate matches slightly less than 90% of all the job postings in the corpus under consideration.

Thus, the recommendation for the candidates is to focus on the hard skills, keeping the finger on the pulse of trends in tooling and technologies. Focusing on developing 2-3 soft skills is recommended.

4.2.5 Selecting skills with maximum coverage weighted by salaries

Finally, let us consider another problem. Previously we wanted to find 5 or 10 skills that will help a candidate to apply for as many job postings as possible. But what if the candidate hopes to acquire new skills that will help them to get a higher salary? For this purpose, we have formulated another optimization problem with the salary

weight. The objective is to maximize the total number of matching job postings and weighted by their salary:

$$max\ \Sigma_{i=1}^{m} v_i \times y_i \tag{4.8}$$

As before, we have the cardinality constraint on the number of skills to include:

$$\Sigma_{s=1}^{n} x_s \leq K \tag{4.9}$$

So, at the end it can be formulated as the following optimization model:

$$max\ \Sigma_{i=1}^{m} v_i \times y_i$$

$$\Sigma_{s=1}^{n} x_s \leq K \tag{4.10}$$

$$x_s \geq a_{is} y_i\ for\ all\ s = 1;, \ldots, n\ and\ all\ i = 1, \ldots, m.$$
$$x_s, y_i \in \{0, 1\}\ for\ all\ s = 1;, \ldots, n\ and\ all\ i = 1, \ldots, m.$$

As done before, let us solve the problem for different values of K. In Figure 4.1, we can see the maximum number of job postings, covered by selecting K skills. In the Figure 4.8 below blue dots represent previous optimization model without salary weights and ret dots represents last model with the salary weights. As expected, salary weight did not impact the number of skills and their coverage of the job postings. The number of jobs that can be covered grows with the number of skills K.

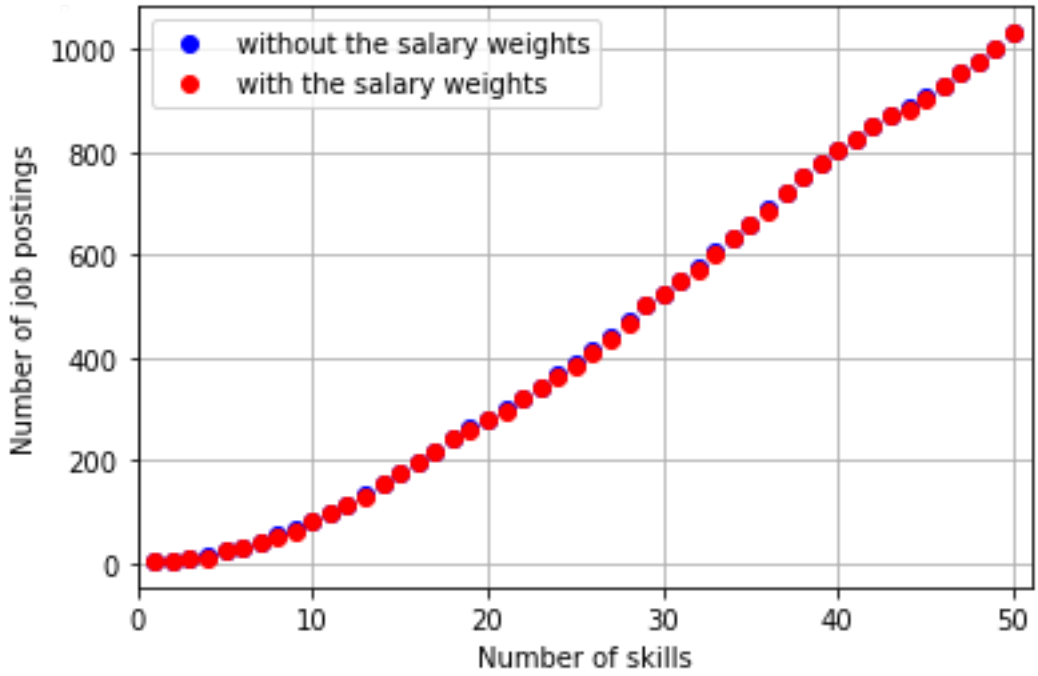


Figure 4.8: Number of Skills vs. Number of Job Postings with and without salary weight

Under the first model, by selecting 10 skills we can cover 83 job postings. Under the second model with salary weights. 10 skills cover 79 job posting, but the highest possible salary will be awarded to the when they possess the following skills:

verbal, writing, sql, python, problem solve, communication, think analytically, excel, collaborate, detail oriented

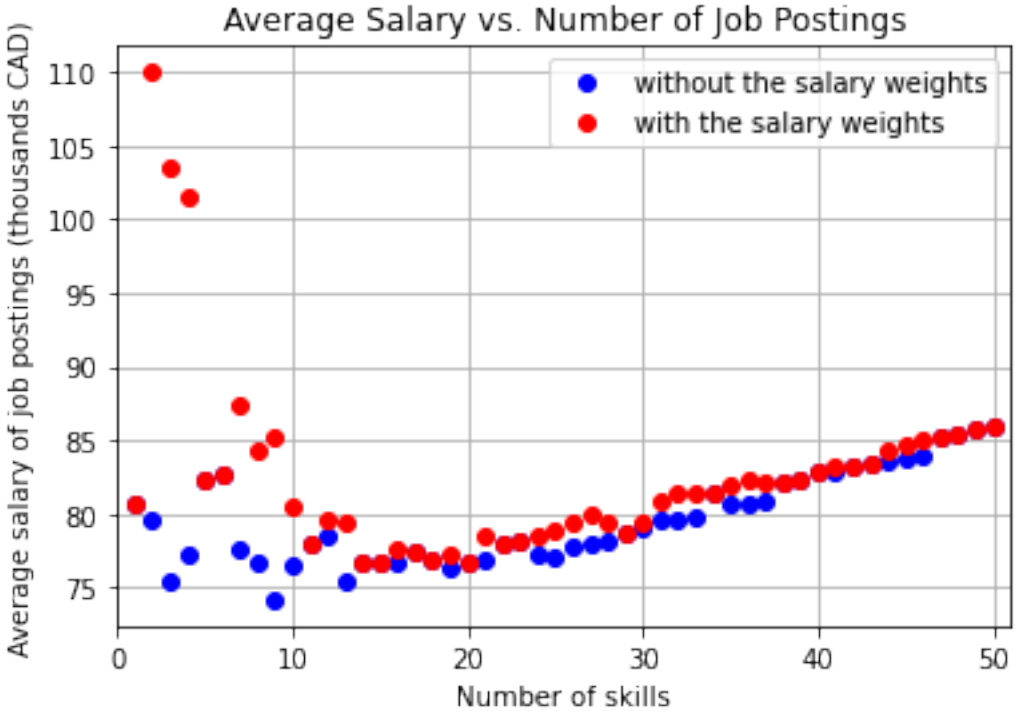


Figure 4.9: Average Salary vs. Number of Job Postings

In Figure 4.9, we show that the average salary of the covered jobs under the model with salary weights (red dots in Figure 4.9) has certain fluctuations for when the number of skills to be selected is small. With increasing number of required skills the slope of the curve is also increase.

This can be explained as follows. For the possession of certain rare skills, the employer is willing to pay an amount significantly higher than the market average. On the other hand, as already noted, entry-level positions contain only a few requirements for hard skills. On the Figure 4.9 we can see a dramatic decrease in salary under the model with salary weights because if more jobs are covered, they will include both high-paying and lower-paying ones. It brings down the average. High salaries are offered to specialists with unique skills and hard to learn.

When the number of skills is large, the average salaries under the two models become very close. This is because if one is allowed to select many skills, they will be able to cover many jobs under both models, which include the higher-paying ones.

Comparing Figures 4.8 and 4.9, we can conclude that including the salary weights in the model impacts the salary level, but to a much lesser extent it impacts the number of job postings a candidate can apply.

4.3 Greedy heuristics

In this part we will consider greedy heuristics for solving the skill selection problem. In the first method, we will add one skill at a time, with the new addition being the skill that leads to the greatest incremental coverage. In the third method, we first cluster the skills, and then use a cluster to cover job postings. In the second method, we add a batch of skills at a time, with each batch maximizing the total number of jobs covered given the batch size.

4.3.1 The basic greedy algorithm

First of all, we start from an empty set and then select the single skill that covers the most jobs.

Single skill	**Number of job postings with the skill**
think analytically	3
collaborate	1
databases	1
sql	1
ds	1
aws	1
communication	1
data analysis	1
agile	1

Table 4.1: Job postings with the single skill

The largest number of vacancies in which only one skill is required is 3. This skill is **think analytically**. The mentioned skill will be the starting point for the greedy algorithm.

In each iteration of the algorithm, we select a skill such that if this skill is added to the current set of skills, the total number of covered jobs will increase by the greatest amount. There may be several such skills, in which we break the tie arbitrarily. After the first iteration of our algorithm, we found that maximum number of vacancies that we can be added is 2. Thus, we can add either 'sql', or 'data analysis' skill.

We use a directed graph to visualize the greedy algorithm. The vertices of the graph are skills. Each edge is assigned a weight that equals the number of vacancies by which coverage increases when a skill is added. We also include a root node as the

starting point, when no skill has been selected.

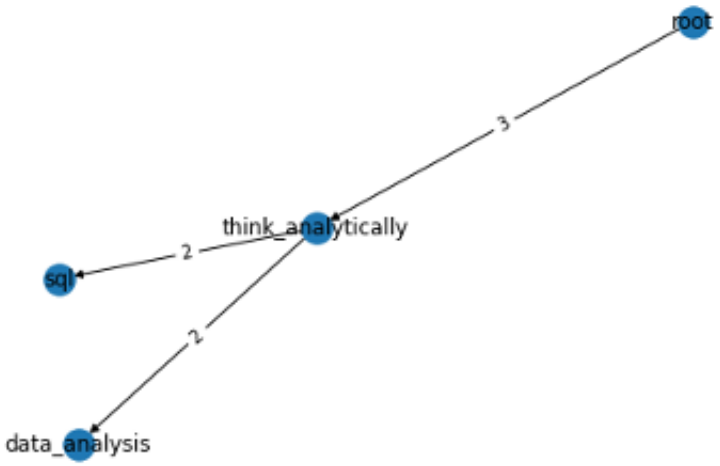


Figure 4.10: Directed graph to model the greedy algorithm

From the graphical representation of the graph (Figure 4.10), it is also clear that when the first skill - think analytically is added to the set, 3 vacancies are covered. By adding one of the following two skills, the number of vacancies covered increases to 5.

Then we continue running the greedy algorithm, which decides the next skill to add.

In the third iteration, we have already selected: [‘think analytically’, ‘data analysis’] or [‘think analytically’, ‘sql’], which covers 5 jobs. Our next skill to choose is [‘collaborate’]. That will give us coverage of another 3 jobs. So, a total number of jobs covered increases to 8.

In the beginning of iteration 4, we have already selected the skills: [‘think analytically’, ‘sql’, ‘collaborate’]. Next skill to add is ‘financial’ with the 4 job posting coverage. That increases the total coverage to 12.

Result of performing more iterations is presented in the Table 4.2 below

Iteration	**Already selected skills**	**Current coverage**	**Next skill to choose**
5	[‘think analytically’, ‘sql’, ‘collaborate’, ‘financial’]	12	[‘excel’], 4
6	[‘think analytically’, ‘sql’, ‘collaborate’, ‘financial’, ‘excel’]	16	[‘problem solve’, ‘data analysis’, ‘data analytics’] 3
7	[‘think analytically’, ‘sql’, ‘collaborate’, ‘financial’, ‘excel’,] ‘data analytics’]	19	[‘problem solve’, ‘data analysis’], 3

Table 4.2: Results of the Greedy algorithm

Also, the same results are represented on the Figure 4.11.

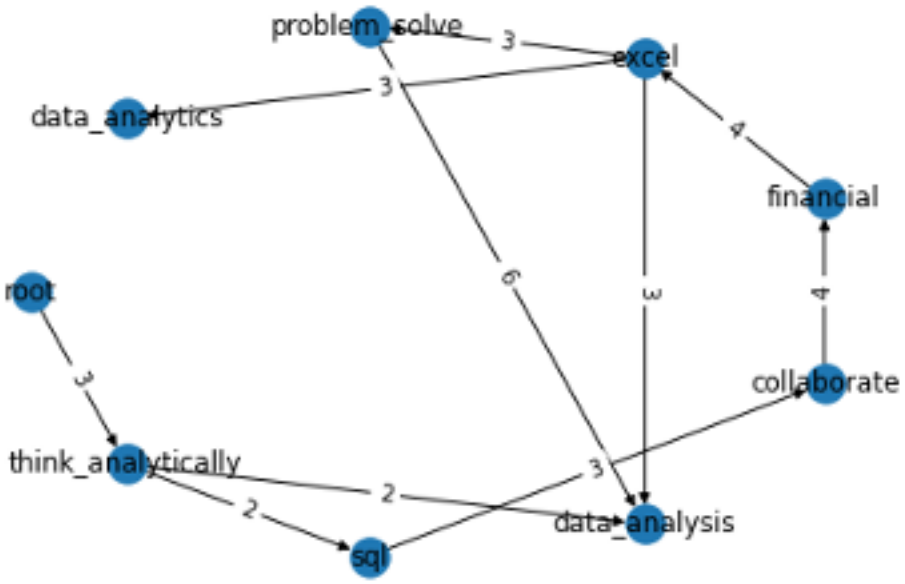


Figure 4.11: Final graph of the Greedy algorithm

If we are to select 5 skills in total, our skills set will be:

think analytically, sql, collaborate, financial, excel

That covers 16 vacancies.

If we compare these two solutions by selecting 5 skills, we cover 24 jobs using the solution from CPLEX and 16 jobs with greedy method. As we can see the greedy method can achieve 66.67% of the optimal value.

4.3.2 The batch greedy algorithm

Another way is to break down the job selection problem into several iterations. Specifically, in order to select K skills, we may performance m iterations and select $k_1, k_2, ..., k_m$ in those m iterations, where $k_1 + k_2 + ... + k_m = K$.

In this section, we will consider selecting a batch of 5 skills in each iteration. For example, if we would like to select 10 skills in total, we may select 5 first in the first iteration by solving the skill selection problem with a cardinality constraint of 5 using CPLEX. Then, after removing those 5 skills and the job posting they can cover, we further selection another 5 skills. For example, form the first iteration of the optimization problem with skills:

verbal, writing, problem solve, communication, think analytically

That means that we will delete these skills in our data before running the next iteration.

	verbal	writing	sql	python	problem solve	communication	think analytically	excel	collaborate	detail oriented	...	communicating insights to non-technical	pytorch	leadership	data management
18	1	1	0	0	0	1	0	0	0	0	...	0	0	0	0
30	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0
183	1	1	0	0	0	0	0	0	0	0	...	0	0	0	0
252	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
253	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
254	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
255	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0
333	0	0	0	0	1	1	1	0	0	0	...	0	0	0	0
387	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0
409	1	1	0	0	0	0	0	0	0	0	...	0	0	0	0
443	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0
482	1	1	0	0	1	1	0	0	0	0	...	0	0	0	0
485	1	1	0	0	1	1	0	0	0	0	...	0	0	0	0
598	0	0	0	0	1	1	1	0	0	0	...	0	0	0	0
603	1	1	0	0	0	1	1	0	0	0	...	0	0	0	0
604	1	1	0	0	0	1	1	0	0	0	...	0	0	0	0
703	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0
728	1	1	0	0	1	0	1	0	0	0	...	0	0	0	0

Figure 4.12: Deleting columns with skills after running Greedy algorithm

After that we will also delete all 24 rows (which represent job postings, fully covered by those skills).

After second iteration we received the following 5 skills:

sql, excel, collaborate, detail oriented, financial

These skills cover 59 job postings. After repeating these steps over and over again, we get:

Skills	**Number of job postings**
[‘data analysis’, ‘time management’, ‘powerpoint’, ‘interpersonal’, ‘data analytics’]	93
[‘python’, ‘databases’, ‘project management’, ‘think critically’, ‘research’]	105
[‘tableau’, ‘presentation’, ‘access’, ‘communicating insights to non-technical’, ‘leadership’]	104
[‘r’, ‘power bi’, ‘visualizations’, ‘statistical analysis’, ‘data management’]	141
[‘azure’, ‘aws’, ‘cloud software services’, ‘modeling’, ‘agile’]	134
[‘ml’, ‘java’, ‘c++’, ‘oracle’, ‘sas’]	142
[‘ai’, ‘tensorflow’, ‘bigdata’, ‘pytorch’, ‘ds’]	109

Table 4.3: Results of the Greedy algorithm with CPLEX

If we compare this solution with the regular CPLEX method in Section 4.2.1 by selecting 10 skills, we would see that this method covers 59 job postings, while the optimal solutin by CPLEX would give us 83 job positions. Thus, the batch greedy

algorithm that selects 5 skills at a time achieves 71.08% of the optimal coverage, which is 48.97% better result.

4.4 Clusters coverage

In this section we want to find out how our clusters from the Chapter 3 cover job postings. First of all, we split dataset into several subsets, based on the skills from the clusters.

Recall that, our clusters from Agglomerative Clustering in Section 3.2.3 were:

0: [‘power bi’, ‘visualizations’, ‘data analytics’, ‘oracle’, ‘access’, ‘research’]

1: [‘excel’, ‘azure’, ‘aws’, ‘databases’, ‘java’, ‘tensorflow’, ‘pytorch’, ‘data management’, ‘hive’, ‘ds’]

2: [‘scala’, ‘sas’]

3: [‘sql’, ‘python’, ‘r’, ‘ml’, ‘data analysis’, ‘spark’, ‘powerpoint’, ‘c plus’]

4: [‘tableau’, ‘financial’, ‘hadoop’, ‘ai’, ‘cloud software services’, ‘modeling’, ‘statistical analysis’, ‘bigdata’, ‘nlp’]

Initial number of vacancies was 1032. After splitting we became with the number of 194 job postings. It’s 18.80% from the initial dataset. One of subset is empty, for other clusters we have:

Cluster	**Number of jobs in the subset**
0	24
1	63
3	67
4	40

Table 4.4: Number of job postings cover by clusters

We note that the skills in the second cluster alone do not cover any jobs. Rather, one needs them combined with skills in other clusters to cover any jobs. So, after all, if we want to select skills from one group of clusters, Table 4.5 reports the number of job vacancies covered by selecting 5 skills from one of the clusters. We observe that only a small number of jobs can be covered. We also report the covered jobs in Figures 4.13 - 4.16.

Our method in this section serves as a benchmark to our optimal skill selection model. By using a "divide and conquer" approach, the scale of the subproblems (i.e., selecting from a cluster of skills) can be reduced significantly. However, as we have observed, the total number of jobs that can be covered is much smaller than our model in the previous sections.

Cluster	**Number of vacancies for 5 skills**	**Skills**
0	4	[‘think analytically’, ‘collaborate’, ‘data analytics’, ‘project management’, ‘communicating insights to non-technical’]
1	6	[‘problem solve’, ‘think analytically’, ‘excel’, ‘collaborate’, ‘databases’]
3	7	[‘sql’, ‘python’, ‘think analytically’, ‘collaborate’, ‘data analysis’]
4	4	[‘verbal’, ‘writing’, ‘think analytically’, ‘tableau’, ‘financial’]

Table 4.5: Group of skills in each cluster

Job_ID	verbal	writing	sql	python	problem_solve	communication	think_analytically	excel	collaborate	detail_oriented	...	communicating_insights_to_non-technical	p
829	0	0	0	0	0	0	1	0	1	0	...	0	
835	0	0	0	0	0	0	1	0	1	0	...	0	
836	0	0	0	0	0	0	1	0	1	0	...	1	
837	0	0	0	0	0	0	1	0	1	0	...	1	

4 rows × 50 columns

Figure 4.13: Selected jobs for the 5 selected skills from the Cluster 0

Job_ID	verbal	writing	sql	python	problem_solve	communication	think_analytically	excel	collaborate	detail_oriented	...	communicating_insights_to_non-technical	p
80	0	0	0	0	0	0	0	1	1	0	...	0	
81	0	0	0	0	0	0	0	1	1	0	...	0	
172	0	0	0	0	1	0	0	1	0	0	...	0	
279	0	0	0	0	0	0	0	0	0	0	...	0	
704	0	0	0	0	0	0	1	1	0	0	...	0	
800	0	0	0	0	1	0	1	1	1	0	...	0	

6 rows × 50 columns

Figure 4.14: Selected jobs for the 5 selected skills from the Cluster 1

Job_ID	verbal	writing	sql	python	problem_solve	communication	think_analytically	excel	collaborate	detail_oriented	...	communicating_insights_to_non-technical	p
24	0	0	1	0	0	0	1	0	0	0	...	0	
262	0	0	1	0	0	0	1	0	1	0	...	0	
277	0	0	1	1	0	0	0	0	0	0	...	0	
390	0	0	1	0	0	0	1	0	1	0	...	0	
651	0	0	0	0	0	0	0	0	0	0	...	0	
688	0	0	0	0	0	0	1	0	0	0	...	0	
987	0	0	1	0	0	0	0	0	0	0	...	0	

7 rows × 50 columns

Figure 4.15: Selected jobs for the 5 selected skills from the Cluster 3

Job_ID	verbal	writing	sql	python	problem_solve	communication	think_analytically	excel	collaborate	detail_oriented	...	communicating_insights_to_non-technical	p
119	1	1	0	0	0	0	0	0	0	0	...	0	
120	1	1	0	0	0	0	1	0	0	0	...	0	
438	1	1	0	0	0	0	0	0	0	0	...	0	
716	0	0	0	0	0	0	1	0	0	0	...	0	

4 rows × 50 columns

Figure 4.16: Selected jobs for the 5 selected skills from the Cluster 4

We conclude this chapter by summarizing some main findings and statistics.

1. The optimal set of skills of 5 is: writing, problem solve, communication, think analytically, verbal, that will cover 24 job positions.

2. Best set of 10 skills is verbal, writing, sql, problem solve, communication, think analytically, excel, collaborate, detail oriented, financial. It gives a candidate a chance to apply to 83 job postings.

3. The top optimal technical 5 skills in demand are sql, python, excel, nlp, data analysis that will cover 587 jobs.

4. Employers are more specific when describing required technical skills than non-technical skills. Non-technical skills are generally identical in job postings, so it is necessary to monitor technical trends, but it is also important to not forget about improving non-technical skills.

5. From the perspective of analyzing only non-technical skills, a job seeker would be eligible to apply for more than 700 job postings if they have only 2 non-technical skills. With 4 non-technical skills a candidate fits approximately 90% of all jobs.

6. Employers are willing to pay an amount significantly higher than the market average for rare skills.

7. Entry-level positions contain only a few requirements for hard skills.

8. Most of job postings require skills from different clusters.

Chapter 5

Conclusion

This thesis presents exploratory analyses and an optimal skillset selection problem. The proposed model and analyses can be useful to potential job seekers (e.g., new graduates, professionals seek to advance or change their careers, etc.) They can also be useful to by colleges, universities, and continuing education programs to adapt their curriculums by adding courses that cover the most in-demand skills

In chapter 3, we explain the collection, pre-processing, and analyses of the data. Through quantitative text analysis, we examine patterns in the job-postings texts. The results show that most common skills in job descriptions are verbal, writing, sql, python, problem solve. Among the top 50 skills 37 are technical skills and the rest (13 skills) are non-technical skills. On average, a candidate needs around 5-6 skills to qualify for one position. The highest correlation among all skills is between PowerPoint and Excel (97%). The frequencies of 3 skills required in the same job posting show that 97% of the job postings requiring verbal will also require writing and sql. Moreover, clusters are identified based on the K-means method and the agglomerative algorithm. We also investigate the correlation between the size of

a company and the types of its required skills, as well as the correlation between the salary level offered by a company and the rating it receives on the online job marketplace.

In chapter 4, we focus on the optimization problem of selecting skills to maximize the number of jobs covered by those skills. We solve several instances of the problem by CPLEX. In the first skill-selection problem instance, we focus on the most 50 common skills as an input. For the job seeker who is ready to master up to 5 skills to qualify for jobs, an optimal set of 5 skills are selected, which cover 24 jobs. Increasing the skills number to 10 increases the number of job postings to 83. In the second skill-selection problem instance, where all skills are considered, only one difference was discovered – skill 'data analysis' replaces skill 'sql' (compared to considering the top 50 skills only). Thus, using top 50 skills for analysis produces near-optimal results. When focusing only on technical skills (as opposed to non-technical or soft skills), an optimal set of 5 was identified that covers 587 jobs, sql, python, excel, nlp, and data analysis. We also consider the objective of the maximum coverage of the skills weighted by their corresponding salaries. We find that employers are ready to pay for rare skills. Entry-level positions contain very few requirements for technical skills. Given that the problem is shown to be NP-hard, we also use the greedy methods to solve the problem heuristically. In the first greedy algorithm, we add one skill at a time. The method selects "think analytically" as the first skill. The same algorithm selected top 5 skills that cover only 16 jobs. In the second greedy method, we add skills in batches of 5 in each iteration. When comparing this solution with the optimal solution by CPLEX for selecting 10 skills, we saw that it covers 59 job postings, while the optimal solution from CPLEX covers 83. Although our model uses a cardinality

constraint on the number of skills to select, it can be readily generalized to include a budget constraint, in which selection of each skill requires a distinct amount of effort/cost and the total effort/cost cannot exceed a given budget.

We conclude the thesis by mentioning a few possible extensions of the current work for future research. One extension could be investigating the competitive ratio of the greedy methods in chapter 4 and exploring other heuristics with a theoretical performance guarantee. Second, we could extend our optimization model by weighting x_s by the effort or time you need to spend on learning new skills to acquire the skill model and to find out the cost for every skill. Third, instead of selecting skills, a similar framework can be developed to select predefined batches of skills. For example, we could consider the problem of selecting/recommending courses on Massive Open Online Course (MOOC) platforms (such as Coursera) that cover the skill requirements of job vacancies. Finally, in this thesis, we apply our models to the Canadian job market due to our choice of data. It would also be interesting to apply our models to the U.S. or the global market and compare those results with the results of this research.

Bibliography

Behpour, S., Hawamdeh, S., and Goudarzi, A. (2021). Employer's perspective on data science; analysis of job requirement.

Bengfort, B., Bilbro, R., and Ojeda, T. (2018). *Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning*. O'Reilly Media, Inc., 1st edition.

Bombardini, M., Gallipoli, G., and Pupato, G. (2012). Skill dispersion and trade flows. *American Economic Review*, **102**(5), 2327–48.

Borghans, L., Weel, B. T., and Weinberg, B. A. (2014). People skills and the labor-market outcomes of underrepresented groups. *ILR Review*, **67**(2), 287–334.

Boselli, R., Cesarini, M., Marrara, S., Mercorio, F., Mezzanzanica, M., Pasi, G., and Viviani, M. (2018). Wolmis: A labor market intelligence system for classifying web job vacancies. *J. Intell. Inf. Syst.*, **51**(3), 477–502.

Burstein, A. and Vogel, J. (2017). International trade, technology, and the skill premium. *Journal of Political Economy*, **125**(5), 1356 – 1412.

CEDEFOP (2019). *The Online Job Vacancy Market in the EU*. Research paper. Publications Office of the European Union.

CEDEFOP (2021). Online job vacancies and skills analysis.

Chryssolouris, G., Mavrikios, D., and Mourtzis, D. (2013). Manufacturing systems: skills & competencies for the future. *Procedia CIRp*, **7**, 17–24.

Colombo, E., Mercorio, F., and Mezzanzanica, M. (2018). Applying machine learning tools on web vacancies for labour market and skill analysis. *Terminator or the Jetsons? The Economics and Policy Implications of Artificial Intelligence*.

Cortes, M., Jaimovich, N., and Siu, H. (2018). The "end of men" and rise of women in the high-skilled labor market.

Deming, D. (2017). The growing importance of social skills in the labor market. *Quarterly Journal of Economics*, **132**, 1593–1640.

Deming, D. and Kahn, L. B. (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals. *Journal of Labor Economics*, **36**(S1), S337–S369.

ETA (2012). Competency model clearinghouse.

Feige, U. (1998). A threshold of ln n for approximating set cover. *JOURNAL OF THE ACM*, **45**, 314–318.

Forsyth, P., Guiomard, C., and Niemeier, H.-M. (2020). Covid -19, the collapse in passenger demand and airport charges. *Journal of air transport management*, **89**, 101932.

Gathmann, C. and Schönberg, U. (2010). How general is human capital? a task-based approach. *Journal of Labor Economics*, **28**(1), 1–49.

Gehrke, L., Kühn, A., Rule, D., Moore, P., Bellmann, C., Siemes, S., Dawood, D., Singh, L., Kulik, J., and Standley, M. (2015). A discussion of qualifications and skills in the factory of the future: A german and american perspective.

Gibbs, S., Steel, G., and Kuiper, A. (2011). Expectations of competency: The mismatch between employers' and graduates' views of end-user computing skills requirements in the workplace. *Journal of Information Technology Education: Research*, **10**(1), 371–382.

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc., USA, 99th edition.

Hershbein, B. and Kahn, L. B. (2018). Do recessions accelerate routine-biased technological change? evidence from vacancy postings. *American Economic Review*, **108**(7), 1737–72.

Hochbaum, D. S. and Pathria, A. (1998). Analysis of the greedy approach in problems of maximum k-coverage. *Naval Research Logistics (NRL)*, **45**(6), 615–627.

Humphreys, A. and Wang, R. J. H. (2018). Automated text analysis for consumer research. *Journal of Consumer Research*, **44**(6), 1274–1306.

Keesing, D. B. (1966). Labor skills and comparative advantage. *The American Economic Review*, **56**(1/2), 249–258.

Lichtenberg, N. (2021). 3 'mismatches' that explain the labor shortage.

Lovaglio, P. G., Cesarini, M., Mercorio, F., and Mezzanzanica, M. (2018). Skills in demand for ict and statistical occupations: Evidence from web-based job vacancies.

Statistical Analysis and Data Mining: The ASA Data Science Journal, **11**(2), 78–91.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2020). Big data: The next frontier for innovation, competition, and productivity.

Martello, S. and Toth, P. (1990). *Knapsack problems: algorithms and computer implementations*. John Wiley & Sons, Inc.

Mbah, R. B., Rege, M., and Misra, B. (2017). Discovering job market trends with text analytics. In *2017 International Conference on Information Technology (ICIT)*, pages 137–142. IEEE.

McLean, C. (2006). A foot in the door: It job-search strategies. *Certification Magazine*, **8**(4), 38–40.

Meyer, M. A. (2019). Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings. *Journal of the American Medical Informatics Association*, **26**(5), 383–391.

Neuendorf, K. (2002). *The Content Analysis Guidebook*.

OECD (2013). *OECD Skills Outlook 2013: First results from the survey of Adult Skills*. OECD Publishing.

OECD (2016). *Skills Matter: Further Results from the Survey of Adult Skills*. OECD Publishing.

Pathak, N. (2017). *Artificial Intelligence for .NET: Speech, language, and search*. Apress.

Pennebaker, J., Mehl, M., and Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, **54**, 547–77.

Rainie, L. and Anderson, J. (2020). The future of jobs and jobs training.

Rholetter, Wylene, P. (2018). Textual analysis. *Salem Press Encyclopedia*.

Roberts, C. W. (2000). A conceptual framework for quantitative text analysis. *Quality and Quantity*, **34**(3), 259–274. 00083.

Salkin, H. M. and De Kluyver, C. A. (1975). The knapsack problem: a survey. *Naval Research Logistics Quarterly*, **22**(1), 127–144.

Santoso, H. B. and Putra, P. O. H. (2017). Bridging the gap between it graduate profiles and job requirements: A work in progress. In *2017 7th World Engineering Education Forum (WEEF)*, pages 145–148. IEEE.

Statistics Canada (2021). Job vacancies, fourth quarter 2021.

Sun, Y., Zhuang, F., Zhu, H., Zhang, Q., He, Q., and Xiong, H. (2021). Market-oriented job skill valuation with cooperative composition neural network. *Nature Communications*, **12**, 1992.

Tang, H. (2012). Labor market institutions, firm-specific skills, and trade patterns. *Journal of International Economics*, **87**(2), 337–351.

Tran, T. T. (2015). Is graduate employability the 'whole-of-higher-education-issue'? *Journal of Education and Work*, **28**(3), 207–227.

Vasiliev, Y. (2020). *Natural language processing with python and spacy: A practical introduction*. No Starch Press, Inc.

Weber, R. P. (1990). *Basic content analysis*. Sage Publications.

Weinberger, C. J. (2014). The increasing complementarity between cognitive and social skills. *The Review of Economics and Statistics*, **96**(5), 849–861.

Yan, R., Le, R., Song, Y., Zhang, T., Zhang, X., and Zhao, D. (2019). Interview choice reveals your preference on the market: To improve job-resume matching through profiling memories. pages 914–922.

Zikopoulos, P., Eaton, C., and IBM (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media, 1st edition.