MIE1624 Assignment 3 report
Jiekai Yin
1004706794

1.Introduction

The fields of data science, AI, big data, and business analytics have gained much popularity in recent years and have become a trending topic in the educational field. The purpose of the assignment is to design a course curriculum for the new “Master of Business and Management in Data Science and Artificial Intelligence” program at University of Toronto, combing both technical skills and soft skills the students need for future employment as a successful data scientist. The curriculum contains __ courses with the skills extracted and summarized from the job descriptions of the job vacancies posted on Indeed website. The clustering algorithms are applied to group relevant skills into courses.

2.Data collection and cleaning

The information of the job title, company, location, rating, date, salary, links, and descriptions are web-scraped into a table, which contains 1235 observations. We focus on data scientist in the location of USA, Canada and remote for this retrieval. The skills are extracted from the description of each job post.

3. Exploratory data analysis and feature engineering

A set of important technical and soft skills are presented and whether the posted job requires the certain skills are exanimated based on the job description. For example, the skill “Deep Learning” is checked by detecting whether the job description contains the word "deep learning", "Deep Learning", "DL", 'Neural Networks', 'ANN', 'MLP', 'CNN', which all represent the field of deep learning. The technical skills include "Python", "SQL/databases", "Excel", "R", "Finance/Risk Management", "Statistics”, “Java", "C/C++", "MATLAB", "SAS", "SPSS", "Tableau”, “Stata", "Power BI", "Hadoop", "Spark", "Mathematics", "Machine Learning", "natural language processing/NLP", "Analytics", "Computer science", "Big Data", "Data Mining", "Data Visualization", "AWS", "Artificial Intelligence", "Deep Learning", “GCP”, “Azure”, “Google Cloud”, “Algebra”, “Operations research”, “DevOps” and “Git”. The soft skills include "Communication", “Teamwork”, “Presentation”, “Problem Solving”, “Project Management”, “Consulting”, “Leadership” and “decision making”.

The occurrences of all the skills are recorded, and the importance of the skills are based on the frequency of occurrences. The word cloud of all tech skills and all business skills are displayed below. Larger front size indicates the more importance of the skills.

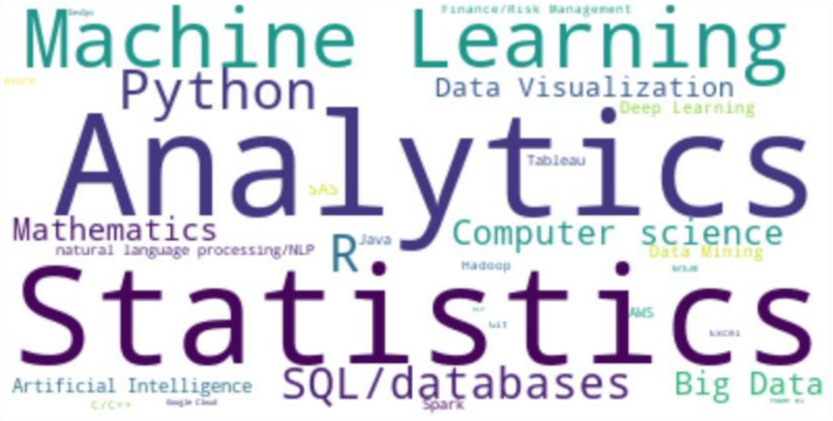


Figure 1:wordcloud of all tech skills

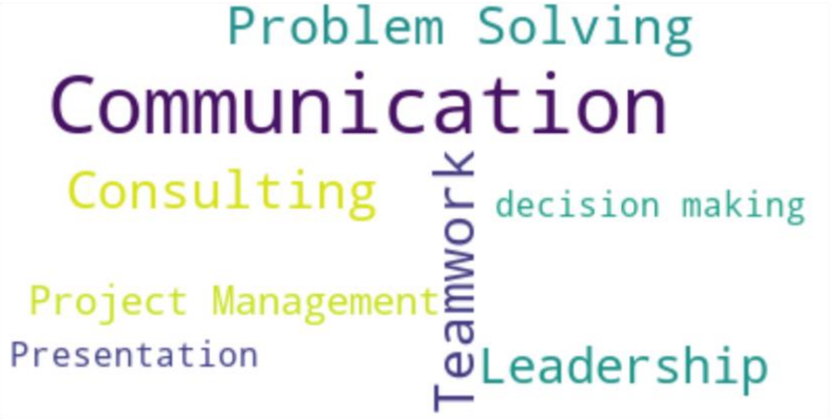


Figure 2:wordcloud of all soft skills

Based on the figure1 above, the most important hard skills for data scientist are analytics, statistics, machine learning and Python. Access to database/using SQL is also important. The most import soft skill is communication. Problem solving, teamwork, leadership and consulting have the similar importance.

4. Hierarchical clustering implementation

The hierarchical clustering is implemented as the first clustering algorithm. It begins by treating each skill as a separate cluster and identify the two clusters that are closest to each other. It continues to merge the two maximum comparable clusters until all the clusters are merged. In this way, the data is grouped into a tree of clusters. The centroid-linkage is used to measure distance in hierarchical clustering, which is the distance between the centroids of two clusters. The dendrograms of all tech skills and all soft skills are displayed below. The number of clusters is decided based on the relevancy between each skill from the dendrograms and the common sense. All technical skills are separated into 8 courses and soft skills are formed to 1 course. Based on the word clouds, not important skills are removed from the curricula. The first two clusters we choose are “Computer science” combining “Mathematics” and “Machine Learning” combining “Python”, since they are the closest cluster. The third group is “SQL/databases” itself. Fourth is “Excel”, “Finance/Risk Management” and “Azure”, considering their closed position. Fifth is “natural language processing/NLP" and “Deep Learning” based on the position. Sixth is “Analytics” combining “Statistics” and seventh are “Data Visualization”, “Data Mining” and “SAS”. The soft skills are combined into 1 course.

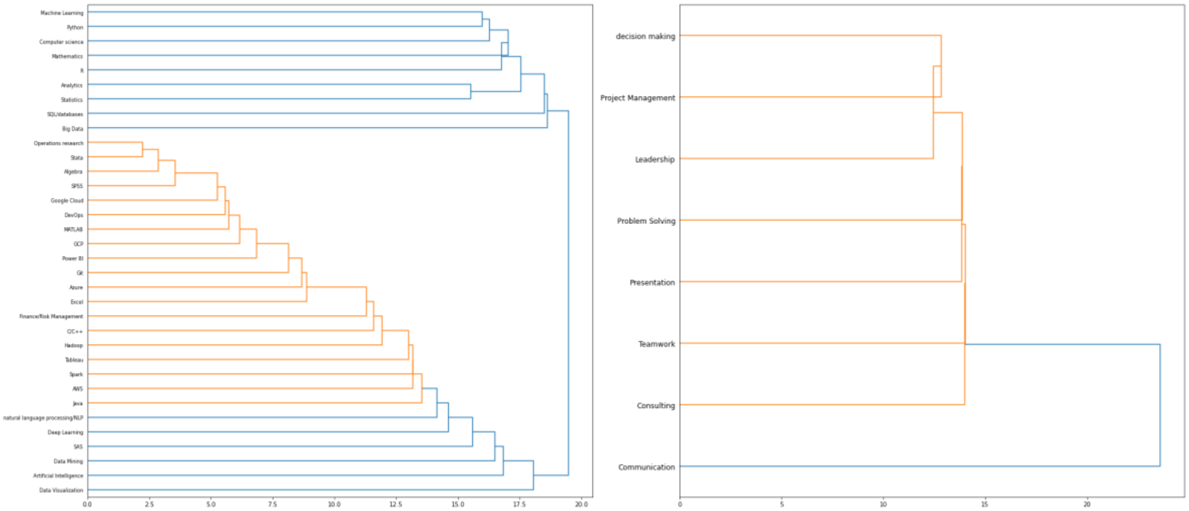


Figure 3: dendrograms of all tech skills and soft skills

5. K-means clustering implementation

K-means clustering is applied as the second algorithm for clustering. The optimal number of clusters is based on the elbow method, which is displayed in the figure below, where the elbow point is believed to be 9. The clustering results are displayed in the scatter plot below, where the data is transposed then dimensionality reduced to 2D using PCA. We can clearly see 9 clusters in the visualization below.

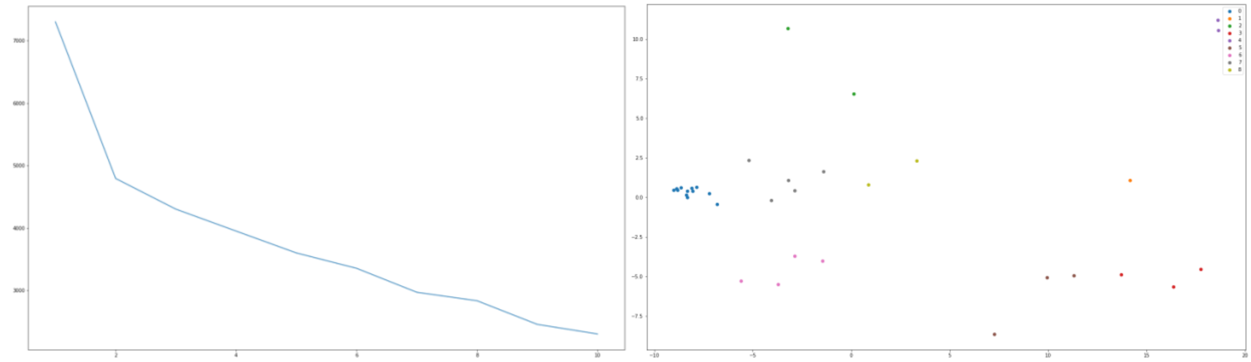
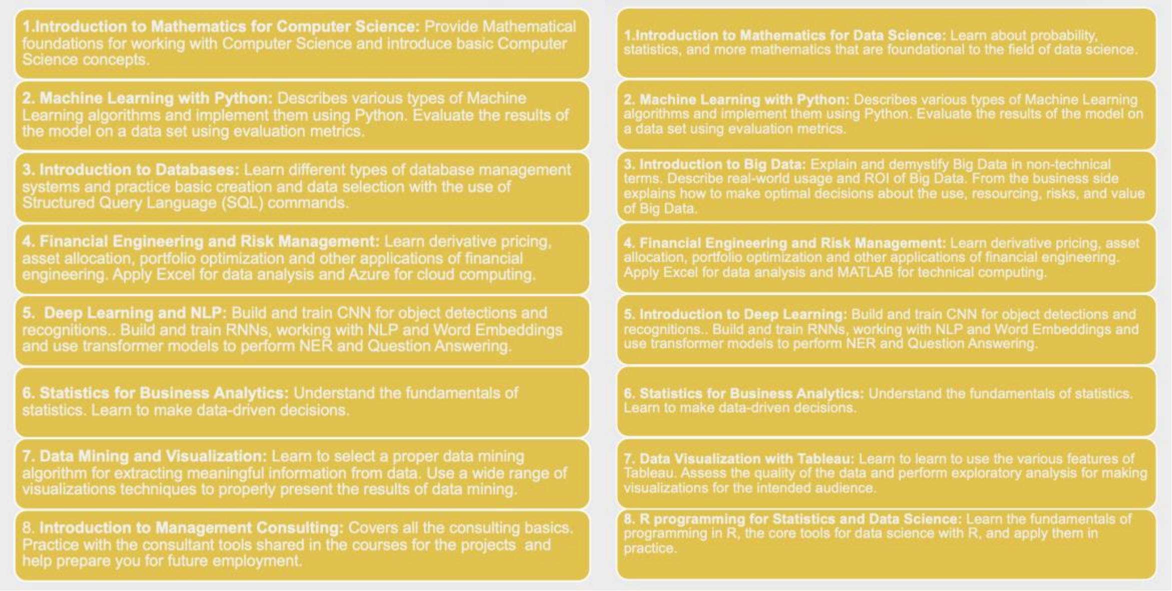


Figure 4:plots of elbow method and clustering resultYs

The choice of clustering is based on the algorithm’s output, importance of the skills and common sense. The first cluster is decided to be “Statistics” and “Analytics”. Second is chosen to be “Deep Learning” since Java and C/C++ are shown to be not very important based on the word cloud. The following 3 clusters are all single skill, which are “Mathematics”, “R” and “Big Data” respectively. The sixth group is “Excel”, “Finance/Risk Management” and “MATLAB”, with removing of some not important skills. The seventh group is “Machine Learning” and “Computer science” and eighth group is “Tableau” and “Data Visualization”, since Hadoop is not very important from the word cloud and NLP is a different field compared to others.

6. Interpretation of results, discussion, and final course curriculum

The course curriculums for the two algorithms are displayed in the graph below. The left table is grouped by hierarchical clustering and the right one is by k-mean clustering.



1.Introduction to Mathematics for Computer Science: Provide Mathematical foundations for working with Computer Science and introduce basic Computer Science concepts.

2. Machine Learning with Python: Describes various types of Machine Learning algorithms and implement them using Python. Evaluate the results of the model on a data set using evaluation metrics.

3. Introduction to Databases: Learn different types of database management systems and practice basic creation and data selection with the use of Structured Query Language (SQL) commands.

4. Financial Engineering and Risk Management: Learn derivative pricing, asset allocation, portfolio optimization and other applications of financial engineering. Apply Excel for data analysis and Azure for cloud computing.

5. Deep Learning and NLP: Build and train CNN for object detections and recognitions.. Build and train RNNs, working with NLP and Word Embeddings and use transformer models to perform NER and Question Answering.

6. Statistics for Business Analytics: Understand the fundamentals of statistics. Learn to make data-driven decisions.

7. Data Mining and Visualization: Learn to select a proper data mining algorithm for extracting meaningful information from data. Use a wide range of visualizations techniques to properly present the results of data mining.

8. **Introduction to Management Consulting:** Covers all the consulting basics. Practice with the consultant tools shared in the courses for the projects and help prepare you for future employment.

1.Introduction to Mathematics for Data Science: Learn about probability, statistics, and more mathematics that are foundational to the field of data science.

2. Machine Learning with Python: Describes various types of Machine Learning algorithms and implement them using Python. Evaluate the results of the model on a data set using evaluation metrics.

3. Introduction to Big Data: Explain and demystify Big Data in non-technical terms. Describe real-world usage and ROI of Big Data. From the business side explains how to make optimal decisions about the use, resourcing, risks, and value of Big Data.

4. Financial Engineering and Risk Management: Learn derivative pricing, asset allocation, portfolio optimization and other applications of financial engineering. Apply Excel for data analysis and MATLAB for technical computing.

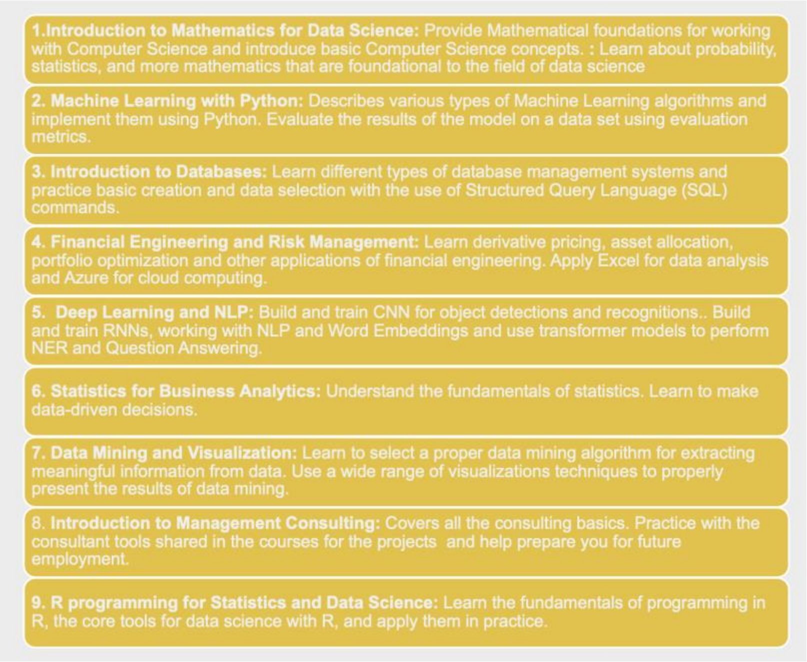
5. Introduction to Deep Learning: Build and train CNN for object detections and recognitions.. Build and train RNNs, working with NLP and Word Embeddings and use transformer models to perform NER and Question Answering.

6. Statistics for Business Analytics: Understand the fundamentals of statistics. Learn to make data-driven decisions.

7. Data Visualization with Tableau: Learn to learn to use the various features of Tableau. Assess the quality of the data and perform exploratory analysis for making visualizations for the intended audience.

8. R programming for Statistics and Data Science: Learn the fundamentals of programming in R, the core tools for data science with R, and apply them in practice.

Figure 5: course curriculums for two algorithms



1.Introduction to Mathematics for Data Science: Provide Mathematical foundations for working with Computer Science and introduce basic Computer Science concepts. : Learn about probability, statistics, and more mathematics that are foundational to the field of data science

2. Machine Learning with Python: Describes various types of Machine Learning algorithms and implement them using Python. Evaluate the results of the model on a data set using evaluation metrics.

3. Introduction to Databases: Learn different types of database management systems and practice basic creation and data selection with the use of Structured Query Language (SQL) commands.

4. Financial Engineering and Risk Management: Learn derivative pricing, asset allocation, portfolio optimization and other applications of financial engineering. Apply Excel for data analysis and Azure for cloud computing.

5. Deep Learning and NLP: Build and train CNN for object detections and recognitions.. Build and train RNNs, working with NLP and Word Embeddings and use transformer models to perform NER and Question Answering.

6. Statistics for Business Analytics: Understand the fundamentals of statistics. Learn to make data-driven decisions.

7. Data Mining and Visualization: Learn to select a proper data mining algorithm for extracting meaningful information from data. Use a wide range of visualizations techniques to properly present the results of data mining.

8. **Introduction to Management Consulting:** Covers all the consulting basics. Practice with the consultant tools shared in the courses for the projects and help prepare you for future employment.

9. R programming for Statistics and Data Science: Learn the fundamentals of programming in R, the core tools for data science with R, and apply them in practice.

Figure 6: final course curriculum

In the final step, we choose to merge the results from two algorithms to the final curriculums. We remove "Introduction to Big Data", since it is from the business side and the student from this program has the science background to dig deeper. The only Math courses for both sides are merged into “Introduction to Mathematics for Data Science”. The courses on the right side with the similar contents are replaced by the ones from the left side. The final output is displayed above. We have 9 courses in total for this program, combining soft and tech skills for data science, and in both theoretical and practical aspects.