Ayushi Pitchika | Curriculum design project

The purpose of this project is to design a course curriculum for a new "Master of Business and Management in Data Science and Artificial Intelligence" program at University of Toronto with focus not only on technical but also on business and soft skills. To achieve this goal, I had to extract skills that are in demand at the job market from job vacancies posted on indeed web-portal and apply clustering algorithms to group/segment skills into courses.

1. Data collection and cleaning:

To start off, I web-scrapped job postings from the indeed.com website for the positions of 'data scientist' and 'data analyst' located either remotely or in the USA using beautifulsoup. A total of 1482 jobs and their important attributes were scrapped and saved to be processed in the main notebook.

2. Exploratory data analysis and feature engineering:

For the pre-processing steps, to clean and prepare data for machine learning, the unstructured web-scrapped job descriptions were transformed by removing all tabulations, numbers, and punctuations, changing everything to lower case and removing all stopwords from the nltk stopwords corpus.

A dictionary was manually defined to identify important technical/hard and business/soft skills that are expected to be in the job postings for the positions of Data Analyst and Data Scientist including 21 technical skills and subskills like Python, Neural Networks, Big Data and 7 business skills like presentation, leadership, project management etc. Each job posting was then run over with this skills dictionary and new features were created for each skill as a binary column. This dataframe gave a good overview of which skills are most frequent in the job postings. These skills were also visualized using wordclouds, first to include all skills and their respective frequencies[Figure 1a] and then all soft skills and their respective frequencies[Figure 1b].

Figure 1: Wordcloud of a) All skills, b) Business/Soft Skills

3. Hierarchical clustering implementation:

First clustering algorithm that was used was hierarchical clustering algorithm with one feature, where that feature represents a distance between each pair of skills. The idea is that if a pair of skills is found together in many job postings, those two skills would be required together on the job, and it makes sense to teach those skills (topics) together within the same course (cluster). To achieve this, a distance matrix was created by looping over each combination of a pair of skills. For each job posting, if a pair of skills occurred together, the distance was determined to be 0. If they did not occur together, the distance was 1. This resulted in a nxn matrix with n being the total number of skills, where pairs of skills that frequently showed up

together had a smaller total distance and skills that did not occur together had a substantially large distance. This was then used to create a 'weighted' linkage and plot a dendogram from hierarchal clustering algorithm[Figure 2].

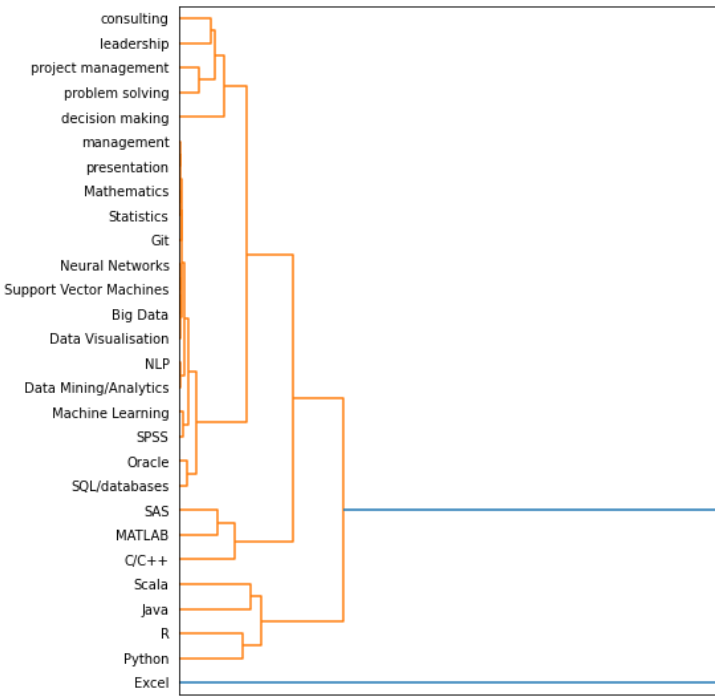

Figure 2: Dendogram of all skills

It was decided to slice the dendogram to obtain 8 clusters that would be the course curriculum obtained through hierarchal clustering.

4. K-means clustering implementation:

For the second clustering algorithm K-Means was chosen. 10 new features were created to be used for fitting the K-Means algorithm. The first 6 features were based on the title of the job posting i.e., 'ML Engineer', 'Consultant', 'Analyst', 'Data Analytics', 'Data Engineer' and 'Data Scientist' binary columns with the 28 skills in the skills dictionary as index. 4 more features were manually encoded, namely 'Hard Skills', 'Soft Skills', 'Coding Needed' and 'Advanced Skill' based on the properties of each of the skills in the skill dictionary. Because the values were of different scales, MinMaxScaler() was implemented to standardize the features. Next the elbow method was used to determine the optimum number of clusters. Although the elbow method indicated 7 as the optimum number, 8 was used to use the corresponding clusters as the 8 required course curriculums. The clustering results were visualized by using PCA to reduce them to 2 features[Figure 3].

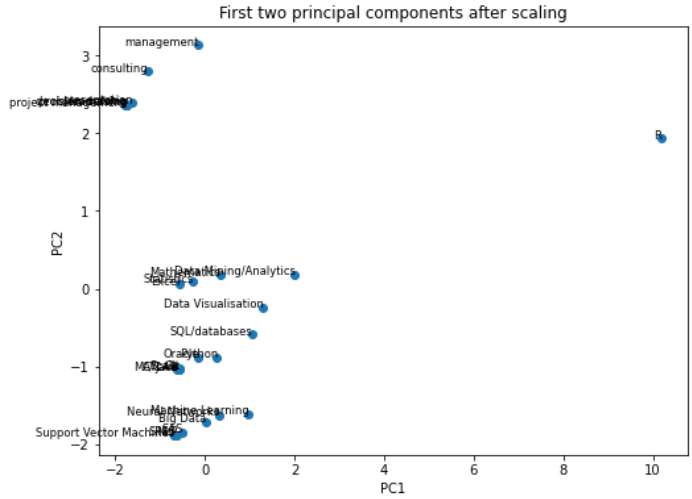

Figure 3: Scatterplot from k-means clustering algorithm

5. Interpretation of results, discussion, and final course curriculum:

The course curricula were visualized for both clustering algorithms. From hierarchal clustering the following curriculum was determined optimum. It was decided that the soft skills should be taught before the hard skills to give students to bridge the gap in their knowledge for the upcoming technical courses. After covering business and mathematical courses, students are introduced to the different programming languages and then to more advanced topics such as deep learning and NLP. The curriculum obtained through k-means clustering was then visualized[Figure 4].

	Course	Topics
0	Foundations of Project Administration	consulting, leadership, project management, excel
1	Introduction to Business Management	communication, presentation, management, decis...
2	Mathematical modelling and Statistics	statistics, probability, statistical modeling,...
3	Foundations of Machine Learning Languages	Scala, Java, R, Python
4	Advanced Analytics in Data Science	SAS, MATLAB, C++
5	Introduction to Deep Learning	Neural Network, SVM
6	Big Data Science	Big Data, Data Mining, Cloud Computing
7	NLP and Machine Learning	NLP, Machine Learning, SPSS, Data Visualization

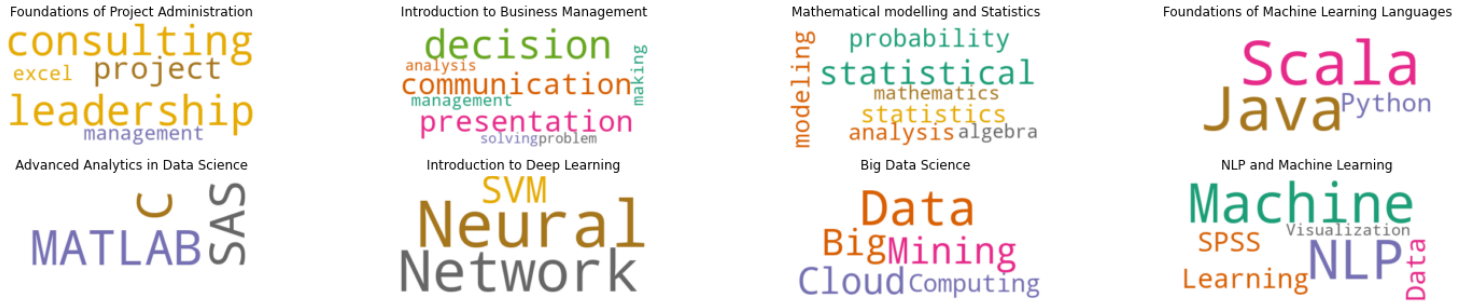

Figure4: a) Table of hierarchal clustering curriculum b) Worcloud of topics taught in each course

Similarly for K-means, the cluster labels were used to determine the following curriculum[Figure 5].

	Course	Topics
0	Introduction to Business Administration	presentation,leadership, decision making,proje...
1	Foundations of Programming Languages	Scala, Java, R, Python, Git, MATLAB, Oracle, S...
2	Mathematical Modelling	Statistics, Mathematics, Excel
3	Advanced Analytics in Data Science	SAS, SPSS, NLP, Neural Network, Big Data, SVM
4	Statistical Computing Language	R Language
5	Foundations od Data Analytics and Visualization	Data Visualization, Data Mining/Analytics
6	Business Management	management, consulting
7	Introduction to Machine Learning	Machine Learning

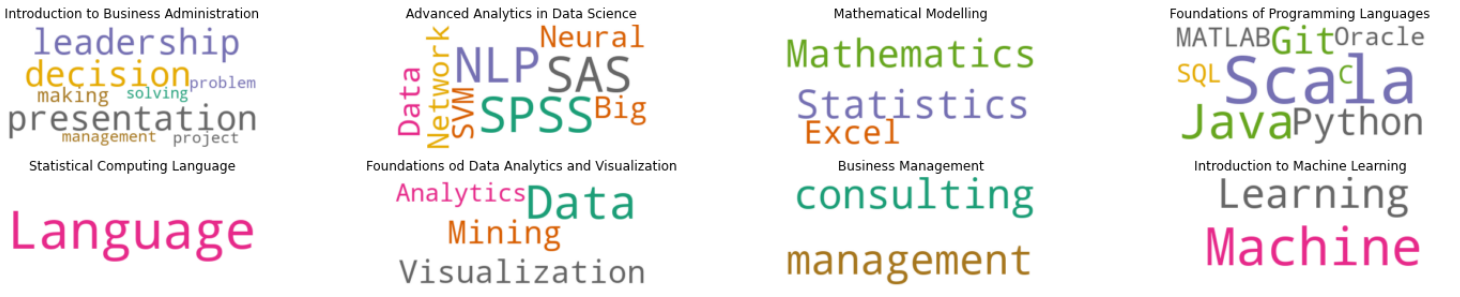

Figure4: a) Table of k-means curriculum b) Worcloud of topics taught in each course for k-means

Although the curricula derived from both clustering methods are very similar, k-means ended up being a more streamlined method of determining the separation of courses as it did not rely on interpretation of a plot to obtain the clustering. Using a pre-trained NLP algorithms like SkillNer that automatically extract skills/features from your dataset might have been a more accurate way to derive these results.