**MIE 1624 Introduction to Data Science and Analytics – Fall 2022**

**Assignment 3**

**Objective:**

(1) To web scrape data and extract skills from job descriptions (indeed)

(2) Use hierarchical clustering and K means to cluster the extracted skills.

(3) design a course curriculum for a new “Master of Business and Management in Data Science and Artificial Intelligence” program

1. **Data Collection and cleaning**

Data is collected from indeed, I have improved the code provided to scrape data (used the concept of parallelization) to reduce the run time from ~3.5hrs to ~17mins for 1000 job scraping. As for cleaning the data, not much is done as I have used spacy and SkillNer NLP model.

1. **Exploratory data analysis and feature engineering**

* Used a NLP model called SkillNer to extract skills, SkillNer is an NLP module to automatically Extract skills and certifications from unstructured job postings, texts, and applicant's resumes.
* That said it was very new and far from complete, has non-exist docs, therefore after spending a while on their github figured out how to make use of it. (more information and examples in Python notebook)
* Once the skills are extracted they are our features, and if they exist in description, then it is 1 if not 0 value in given.
* Around ~4000 were extracted but ,uch cleaning was required as model was overfitting the data.
* Finally ended up with 127 skills in around 1137 descriptions.
* some visualisations are below.

Text

Description automatically generatedText

Description automatically generated

Figure 2 Hard skills WordCloud

Figure 1 Soft Skills WordCloud

**NOTE:** Other visualizations such as all skills, salaries are available In python notebook

1. **Hierarchical clustering Implementation**

Distances plotted:

* cosine
* Euclidean
* Jaccard
* Hamming

Chosen: **Jaccard**

While all of these resulted in very similar dendrogram I have personally felt **Jaccard** distance one did better job.

Cosine is not suited as it is sensitive to repetitions of word (which we do not have any), same with Euclidean.

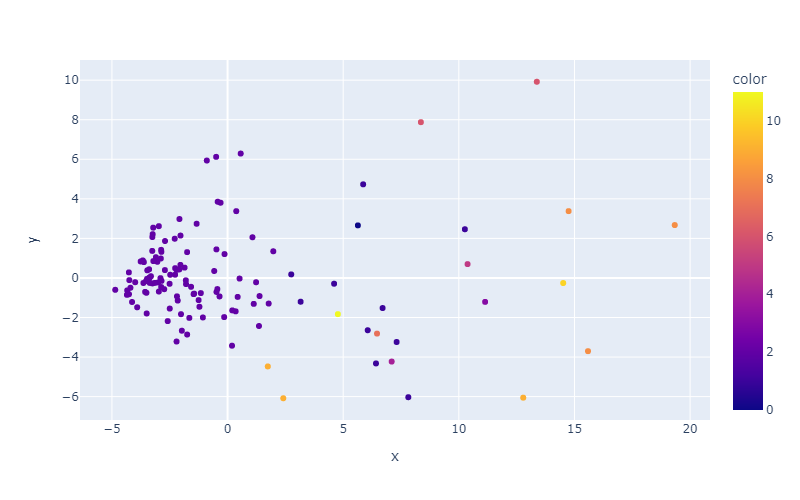
Hamming and **Jaccard** are suited for binary such as 1 and 0 in our case, Jaccard has advantage as it can work with any length whereas hamming expects the descriptions of be of same length.

Clusters chosen 8, as we can see some of the clusters cannot lead to course by itself so looking all the courses, I believe.

*All the graphs are available in Python Notebook. (they are very large to present here)*

1. **K-means clustering implementation:**

After using elbow method I have got 12 clusters as the elbow point thus k means is done for 15 clusters, do note, the results when compared to Clustering are really bad, so I have did the k means again separately for hardskills and softskills, which made it slightly better but not better compared to hierarchical clustering.

Chart, line chart

Description automatically generated

1. **Interpretation of results, discussion, and final course curriculum:**

I chose hierarchical clustering to provide better results thus course made from it was way better. Adobe Illustrator is

***K means***

Courses using k means are as follows: (clusters are printed in notebook)

Using plotly tree

Chart, treemap chart

Description automatically generated

The numbers correspond to course:

1) Operations Management, 2)AI and Big Data, 3)Statistics for Data Scientists, 4) *R* for Data science, 5) Introduction to computer science, 6) Introduction to machine learning, 7) Highlighting yourself as leader, 8) Python for Data science, 9) Data Visualization, 10) Collaborations for engineers, 11) Maintainability in Data Science, 0)Research techniques for engineers

**Note:** High quality interactive plot for above is available in notebook, cannot bring it here without losing significant quality or consuming complete page, numbers in plot are the groups and match with numbers above

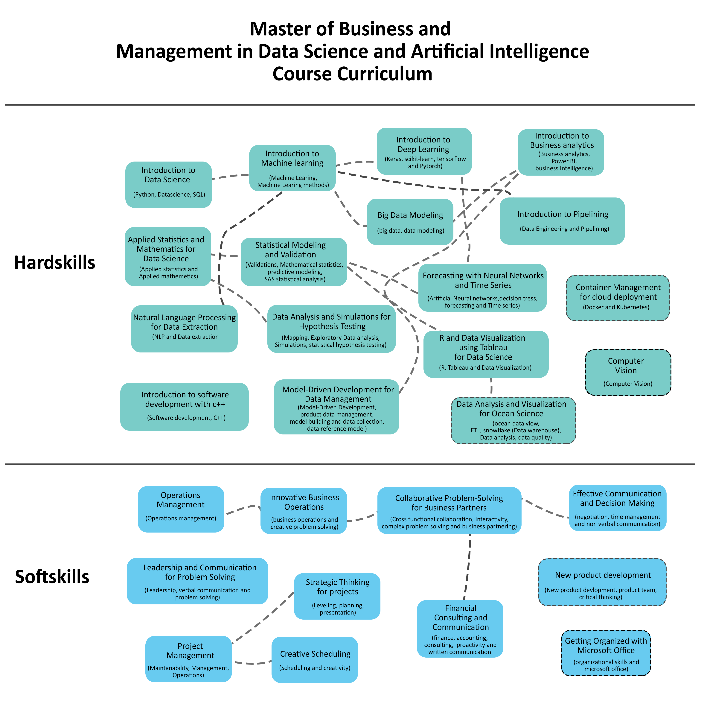
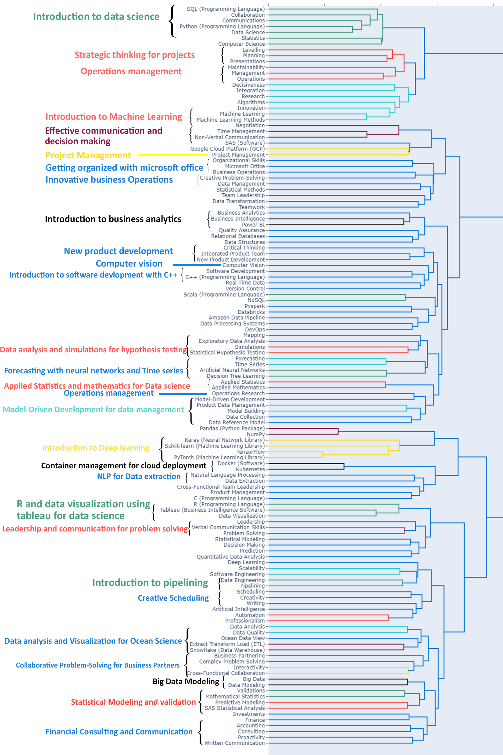
***Hierarchical clustering***

Figure 3 Hierarchial courses

Figure 4 For Bonus mark

**Note: was trying to maintain the 3 pages limit, I will also upload (High quality) images independently**