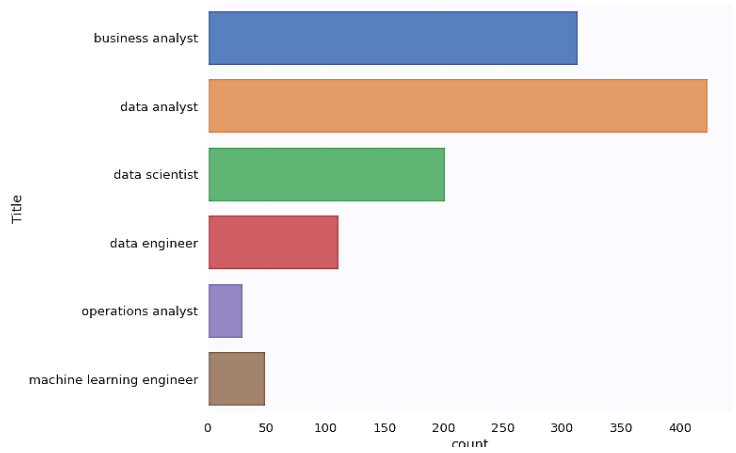
Exploratory Data Analysis and Feature Extraction

(i) A master list of skills to be sought out in the job descriptions was prepared based on domain knowledge and by taking help from gpt-3.5-turbo API.

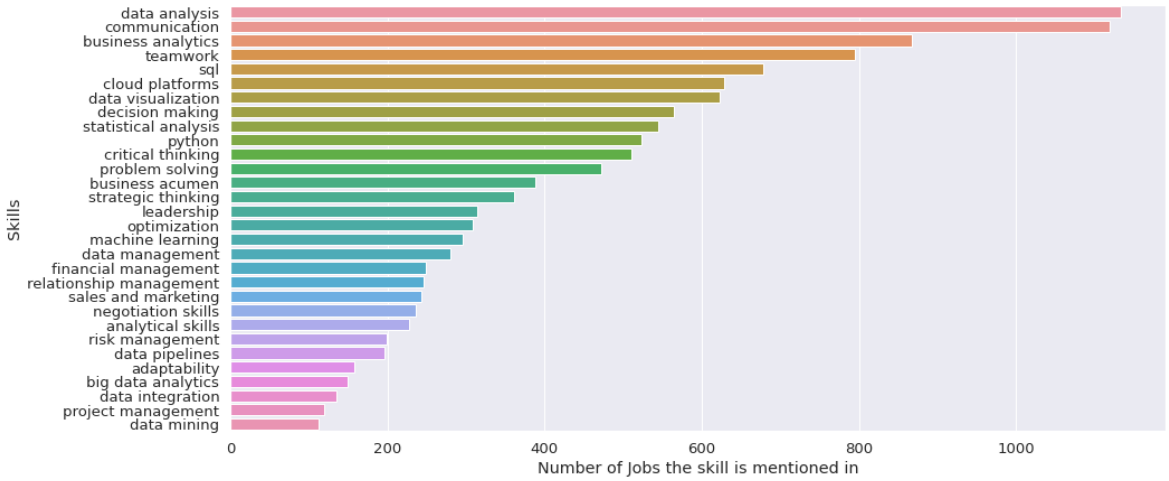
(ii) An array of patterns from the above-mentioned list was created by stemming the individual words. Phrase matcher function from spacy was used to find locations of matching patterns in the entire job descriptions. These matches along with their neighbouring words were picked and snippets of these job descriptions were created which only mention the relevant skills that we desire to extract.

(iii) Skills comprising of one word were directly found from these snippets by string search. However, for two-word skill, string search cannot be used directly. Additionally, we also wanted to find skill mentions which can be represented by different word combinations such as 'cloud services' and 'cloud environments' can both match with 'cloud platforms'.
In order to do this, we converted the above-mentioned snippets to bi-grams, and then filtered those bi-grams which have at least one word common with a skill pattern. We then fed these filtered bigrams as a list corresponding to one job description, to Chat GPT and asked it to match them with the reference list and return an array representing the presence or absence of each skill in the reference list.

(iv) Demand of popular roles in the job market was explored and below were the results,



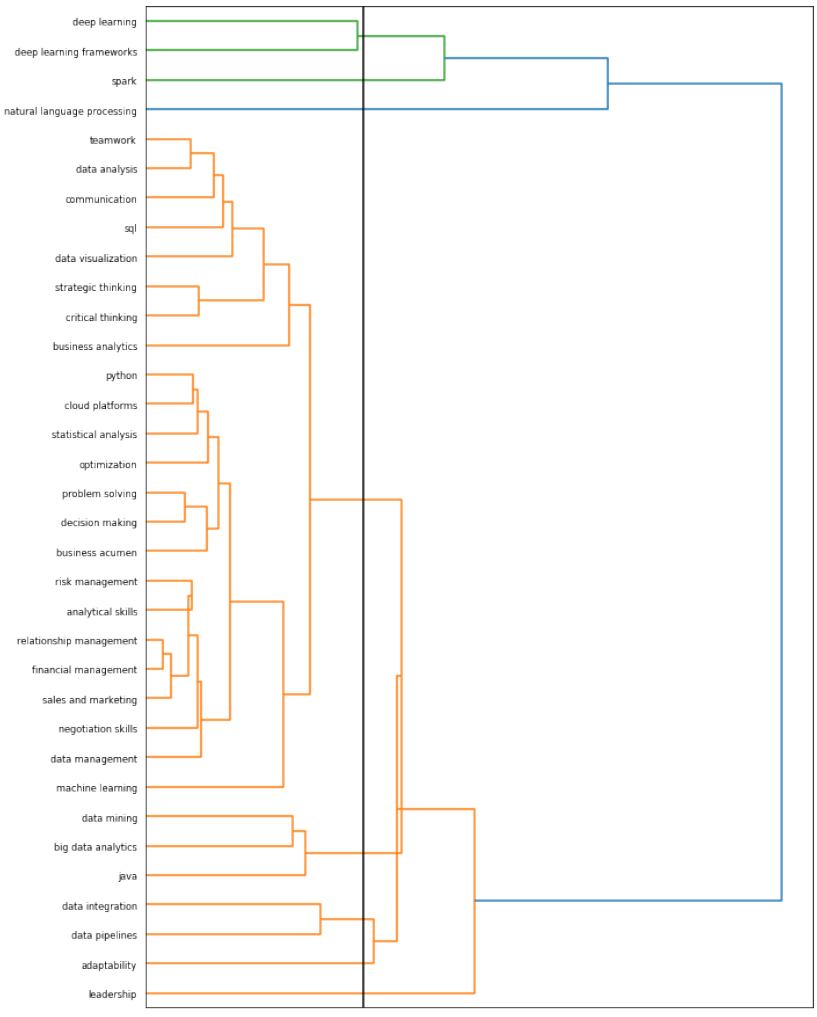
(v) Most sought after skills were also studied with following results,



Hierarchical Clustering Implementation

(i) A symmetric matrix containing the number of jobs common between pairs of skills was created from the original matrix obtained after skill extraction process.
From this matrix of common jobs, a proximity matrix was created by first dividing each element of the matrix by the corresponding element in the diagonal. By this we mean that the elements in the upper triangular half of the diagonal were divided by the diagonal element in their rows, and the elements in the lower triangular half were divided by the diagonal element in their columns. This was done to remove the bias of frequency of mentions, as we were interested in calculating proximity only among the jobs in which at least one of the skill in the pair was mentioned.
If this is not done, then pair of skill such as deep learning and pytorch which were mentioned very few times as compared to say, team-work and communication, would appear to be much far from each other, which is not true.

(ii) From this proximity matrix, hierarchical clustering was performed, and curriculum can be designed as follows using below dendrogram,



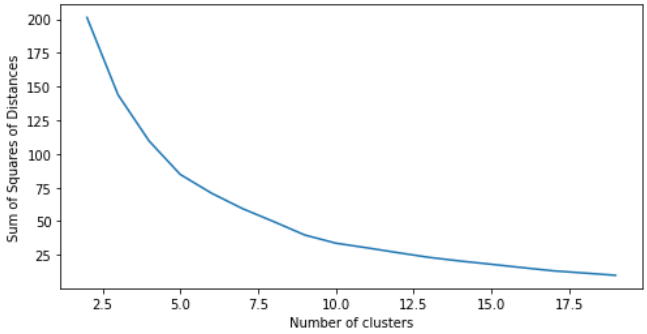
Proposed curriculum in logical sequence of courses:

1. Data Visualization, Data Analysis, Strategic Thinking, Business Analytics, SQL
2. Python, Cloud Platforms, Statistical Analysis, Optimization, Decision Making
3. Data Mining, Big Data Analytics, Java
4. Risk Management, Financial Management, Relationship Management, Negotiation Skills, Sales and Marketing
5. Distributed Computing, Spark, Hadoop
6. Data Integration, Data Pipelines, Data Engineering
7. Neural Networks, Deep Learning, Deep Learning Frameworks
8. Natural Language Processing, Sentiment Analysis in NLP, Clustering in NLP

K Means Clustering Implementation

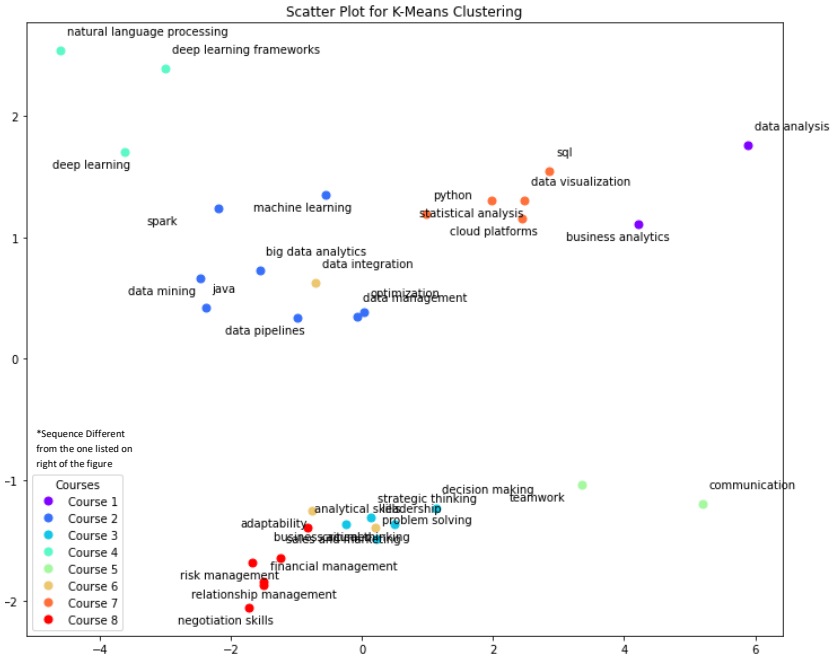
(i) 10 unique features were engineered for 34 skills to prepare a dataset for performing K-Means clustering. For instance, one feature was directly obtained by transforming the proximity matrix data onto a scale by combining information of distance between a pair of skills with reference to all other skills. Another feature was created by categorizing as technical or business skill, while one was created using the frequency of each skill.

(ii) Below is the elbow plot for K-Means clustering for different values of K.



8 or 9 number of clusters would be an appropriate choice for this elbow plot.

(iii) Therefore, using 8 number of clusters, scatter plot of clustered results and proposed curriculum is as follows,



Final Proposed curriculum in logical sequence of courses:

1. Data Analysis, Business Analytics
2. Communication, Teamwork
3. SQL, Cloud Platforms, Data Visualization, Statistical Analysis, Python
4. Leadership, Adaptability, Data Integration
5. Optimization, Machine learning, Data management, Data pipelines, Big Data analytics, Data Mining, Java, Spark
6. Decision Making, Critical Thinking, Problem solving, Business Acumen, Strategic Thinking
7. Deep Learning, Deep Learning Frameworks, Natural language processing
8. Financial Management, Relationship Management, Sales and Marketing, Negotiation Skills, Analytical Skills, Risk Management

The curriculum obtained seems to be a very reasonable one.:

Course 1 is the Data Analytics Course which will cover the fundamental aspects of data analytics

Course 2 is about essential skills in the work environment emphasising on team-work and communication

Course 3 is the Data Science course, which is resonating closely with MIE1624 at UofT

Course 4 Business Management equivalent course focussing on managerial aspects such as adaptability and leadership

Course 5 is a Big Data Course which will cover all aspects of Big Data Architecture, tools for Big Data processing such as Spark, and Data Management and Data Movement techniques. This is very similar to MIE 1628 at UofT

Course 6 is the Business Strategy Equivalent course which is covering aspects such as Decision Making and Strategic Thinking

Course 7 is about Neural Networks and their Applications such as NLP

Course 8 is the core Business Skills course, which can intuitively be interpreted as a mini-MBA