MIE 1624: Introduction to Data Science and Analytics

Homework 3

By: Alexis Bruneau: 1008704270

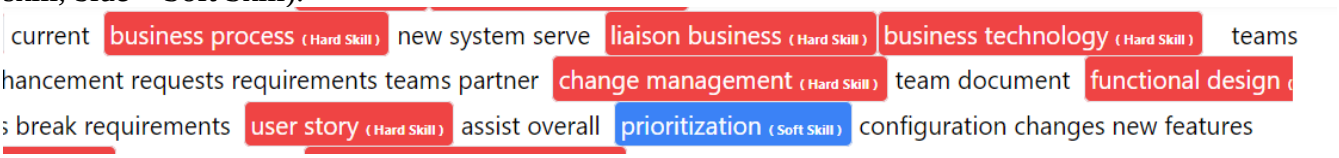
Presented to Dr.Oleksandr Romanko

November 30th, 2022

Q1 Data collection and cleaning) In part 1, the provided code was modified to search for ten job positions in the United States. The following positions were searched: Data Scientist, Data Analyst, Machine Learning Engineer, AI Engineer, AI system designer, Manager of Analytics, Director of Analytics, Data Science Manager, Business Analyst, and Business Analyst. The code was improved by reducing the wait time between searches and implementing try blocks to have a more robust code. The results were saved and used in the Jupyter notebook. (Code added in submission)

Q2 Exploratory data analysis and feature engineering)

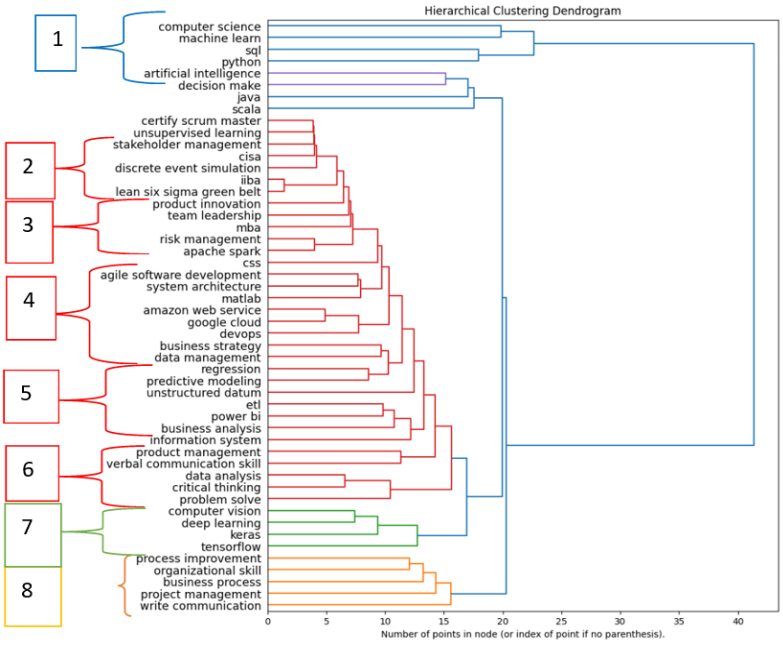
a – extracting skills) In this section, data exploration/cleaning was performed. Any descriptions that had an issue were removed. Afterward, the job title was renamed to one of the original searches using fuzz from fuzzy. Afterward, the job description was cleaned to have better results when using the NLP algorithm. The following was applied to the job descriptions: removed HTML, Unicode, rt @, punctuation, numbers, and used English stop words. An NLP package called SkillNer was used to extract Soft Skills, Hard Skills, and Certification from the job description. It is to note that method was time-consuming (2 hours). For this reason, the results were saved as a CSV file and used further in the code. It is also to note that upon inspection, the package did not extract some technical skills that are common in our field, and some were added later in the code. Also, when using their package, they return the word and a key value. To know whether that word is considered a soft/hard skill or a certification, the script needed to go read one of their JSON file indicating the type of skill (JSON file attached with submission). The figure underneath illustrates an example of how the NLP categorizes some words (red = hard skill, blue = Soft Skill).



b-c-d) After using the NLP, I chose to visualize the keywords that were extracted. The following word clouds were generated (All skills, soft skills, Hard skills, Certification). The Word cloud can be found in the appendix. As mentioned, some expected words were note extracted. Hence, some hard skills were added such as python and SQL. Once the skills/certifications were scrapped.

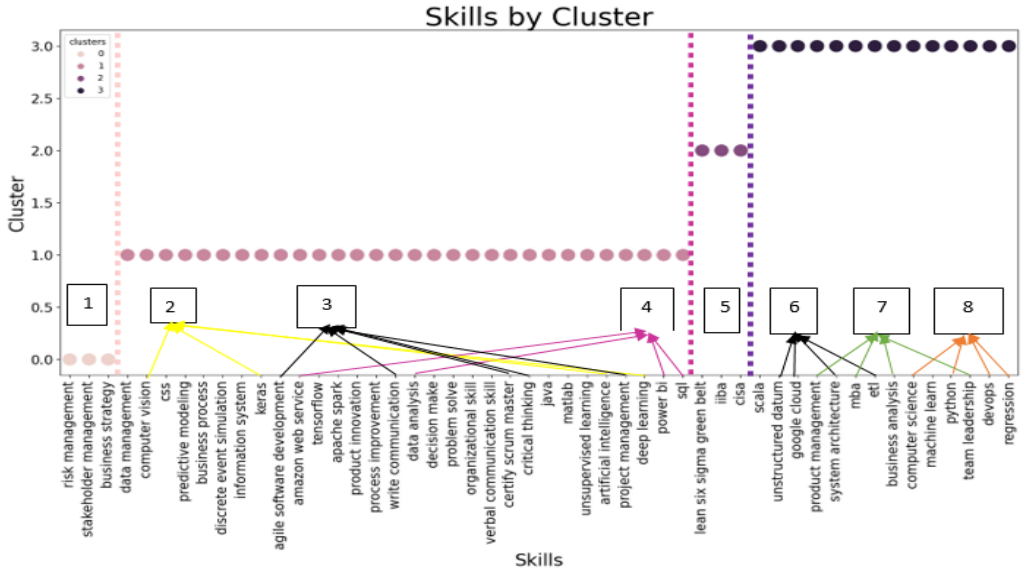
The number of times each skill appeared was plotted (max 1 per row, available in Appendix). Once the skills were chosen, a table was constructed for part 3. The table consisted of the following, one extracted job per row and the desired skills for each column. The columns were marked as 1 if they appeared in the description and 0 else wise. Another table was created for Part 4 (K-means). The skills were the rows, and 10 features were used. The following were used as features: (the average yearly salary, Rating avg, avg count average count for bachelor's and master's degrees required, whether it is a soft, hard, or certification, job title, City, and State. It is to note that the data was numerical and categorical. Hence, the categorical data was one hot encoded (table found in the appendix).

Q3) Exploratory data analysis and feature engineering) a-b-c)



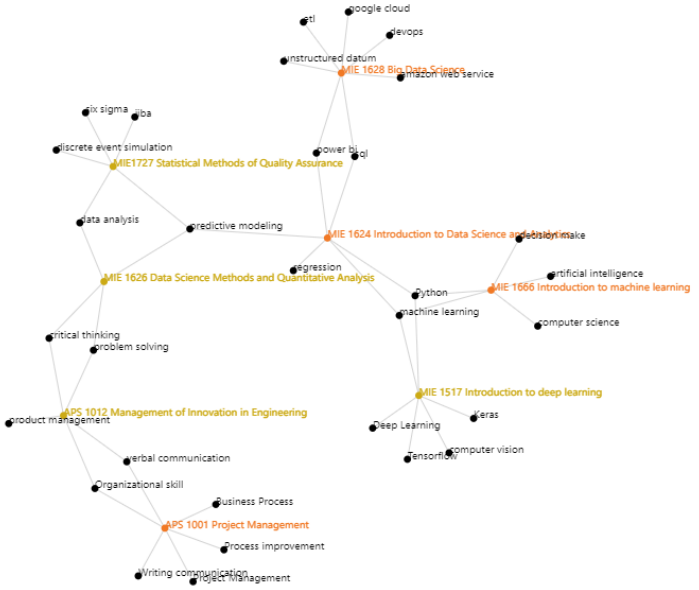
In this section, I used some of the code that sklearn provides for hierarchical clustering dendrogram [1]. In the dendrogram, 8 clusters were chosen, and each has between 4-8 skills. The approach was that each cluster would represent one course. It is to note that some courses would teach some skills outside of some clusters, but the main skills in each cluster would be the main topics. It was important to capture the soft and hard skills. It was also important to group skills that would make sense to teach together. For example, cluster 1 is mostly related to machine learning, and cluster 8 can be seen as skills that can be taught in a management course. Upon inspection of the clusters, most skills belonging to one cluster can be taught together.

Q4) K-means clustering implementation) In this section, a K-mean clustering algorithm was used. The setup of the table was discussed in question 2. The elbow method was used to decide the number of clusters (see appendix). From the graph, the optimal number of clusters can be interpreted as 2 or 4. Knowing I want to create a course sequence of a minimum of 8 courses, I opted to use 4 clusters. The figure underneath illustrates the results of the algorithm (skill belonging to one of four different clusters). The approach for clustering skills was the following. A minimum of 3 courses would be required for each chosen cluster. Also, I decided that no grouping could contain skills coming from other clusters (0-3). Hence, clusters 1 and 5 were automatically defined. Afterward, since within a defined cluster from the algorithm the skills are weighted the same, I tried to group skills that had similarities between them based on my knowledge.



Q5) Interpretation of results, discussion, and final course curriculum) When creating the course sequence, I tried to include the skills that appeared the most in the job scrapes. Looking at both algorithms, it seems that both algorithms performed well as they were able to cluster skills in a matter that makes sense. Between both algorithms, the hierarchical approach seemed to cluster skills better. There were the results I expected since it makes more sense to assume skills are related if they appear on multiple job postings compared to the K-mean method which relied on some features that might not provide some information on the skills such as Location. For both algorithms a course sequence of 8 courses is created, 4 courses per semester. Real courses from graduate studies at UOFT were chosen. The chosen skills try to match the class syllabus as best as possible. The final course curriculum for both algorithms is presented in the table underneath. It was important to include soft/hard skills, and skills oriented toward IT and management. Also, when creating the course sequence, it was important to not offer a course with many prerequisites in the first semester such as deep learning. For both algorithms, MIE 1624 and 1628 teach multiple skills that are important for courses in the second semester. The following combines **both curriculums**. would include MIE 1624, MIE 1666, MIE 1626, and APS 1001 in the first semester. It was important to offer MIE 1624 in the first semester as it gives a good base for multiple courses. MIE 1626 will offer more details on how to perform some data analysis. Finally, from our analysis, we saw the importance of some soft skills and management skills. APS 1001 will teach some fundamental skills of Project Management. In the second semester, MIE 1517 will be offered and will be a succession of MIE 1666 and dive into deep learning. APS 1012 will offer more management skills and focus on different soft skills. MIE 1727 will use some of the content seen in 1626 and 1624 to apply quality control techniques. Finally, we saw cloud was an important aspect in some job descriptions. Hence, the semester will end with MIE 1628. The Neural Network displays the courses (orange = 1st semester, yellow = 2nd semester and how courses overlap in some skills. (* Note, both figures are available in the Appendix bigger).

Hierchical				K-mean			
Course	Cluster	Semester	Skills	Course	Cluster	Semester	Skills
MIE 1666 Introduction to Machine Learning	1	1	Machine learning, python, sql, artifical intelligence, computer science	APS 1001 Project Management	1	1	six sigma, iiba, cisa
APS 1012 Management of Innovation in Engineering	3	1	Product Innovation, team leadership, mba, risk management	0974 Business Strategy	3	1	risk management, stakeholder management, business strategy
MIE 1628 Big Data Science	4	1	Agile Softw are Development, System Architecture, Matlab, AWS, google cloud, devops, business strategy, data management	MIE 1628 Big Data Science	8	1	AWS, data analysis, power bi, sql
MIE 1624 Introduction to Data Science and Analytics	5	1	Regression, predictive modeling, unstructured datum, etl, powerbi, business analysis	MIE 1624 Introduction to Data Science and Analytics	7	1	Computer science, python, devops, regression
2565 – Tools & Techniques of Business Process Management	2	2	stakeholder management, cisa, discrete event simulation, iiba, six sigma	3523 Agile Essentials in Project Management	5	2	Agile softw are development, writing communication, scrum master, critical thinking, project management
TEP 1502 Leadership in Product Design	6	2	Product Management, Verbal communication, data analysis, critical thinking, problem solving	MIE 1517 Introduction to deep learning	4	2	computer vision, keras, artificial intelligence, deep learning
MIE 1517 Introduction to deep learning	7	2	Computer vision, deep learning, keras, tensorflow	INF 1343	2	2	unstructured datum, google cloud, system architecture, etl
APS 1001 Project Management	8	2	Process improvement, organizational skill, business process, project management, writing comunication	APS 1012 Management of Innovation in Engineering	6	2	product management, mba, business analysis, team leadership



Reference

[1] https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html

Appendix

Question 2

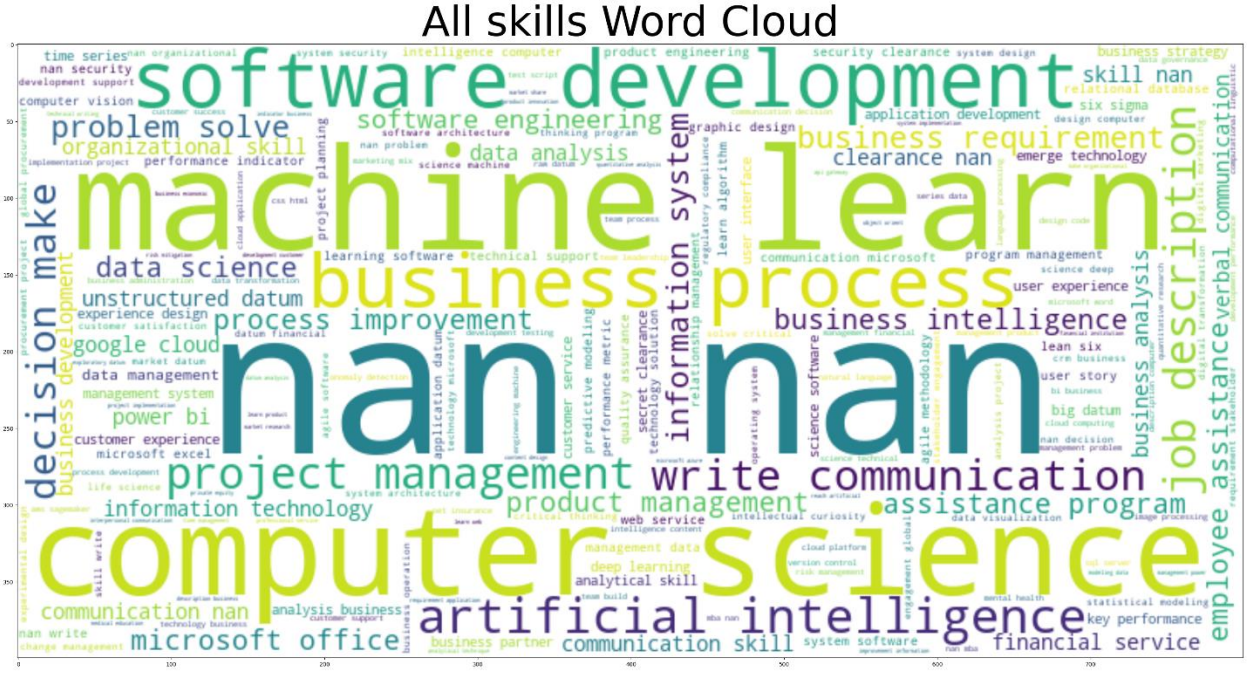


Figure 1 All Skills Word Cloud

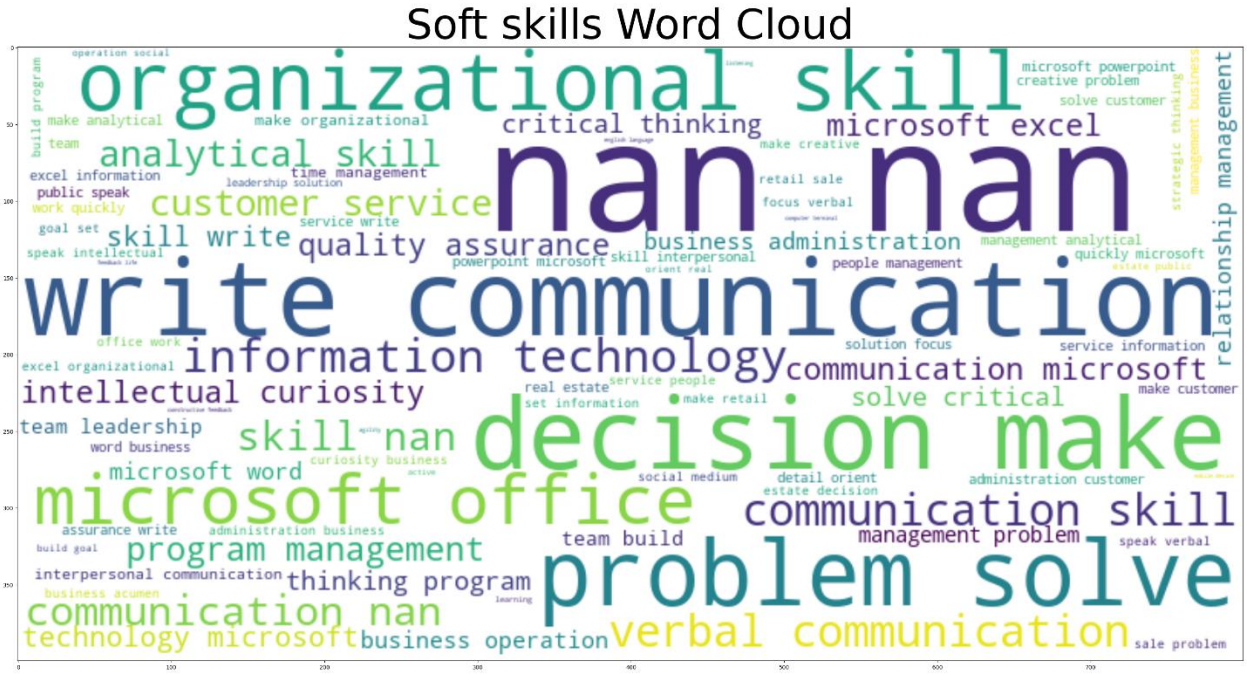


Figure 2 Soft Skills Word Cloud

Hard skills Word Cloud

Figure 3 Hard Skills Word Cloud

Certification Word Cloud

Figure 4 Certification Word Cloud

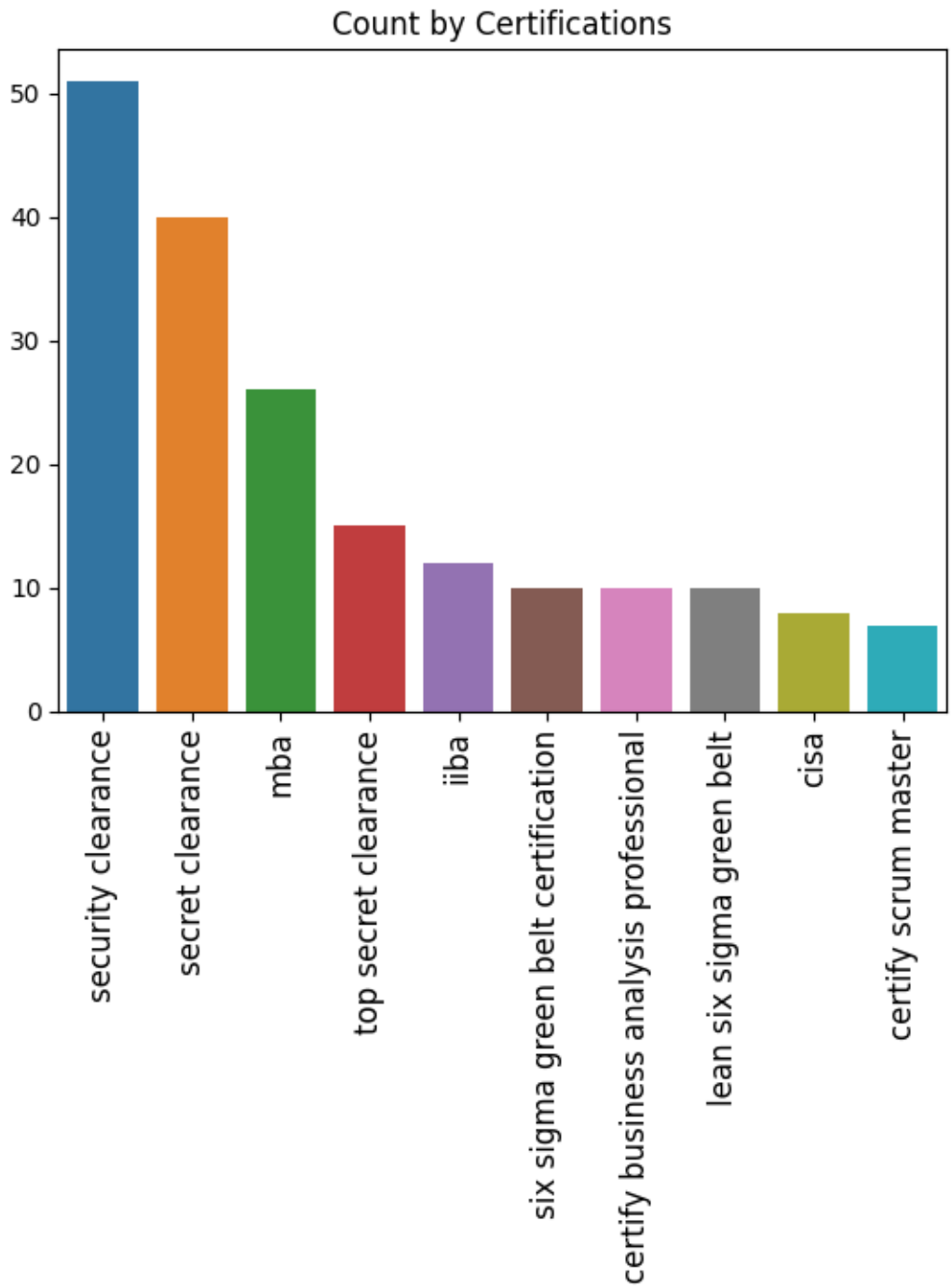


Figure 5 Count by Certification

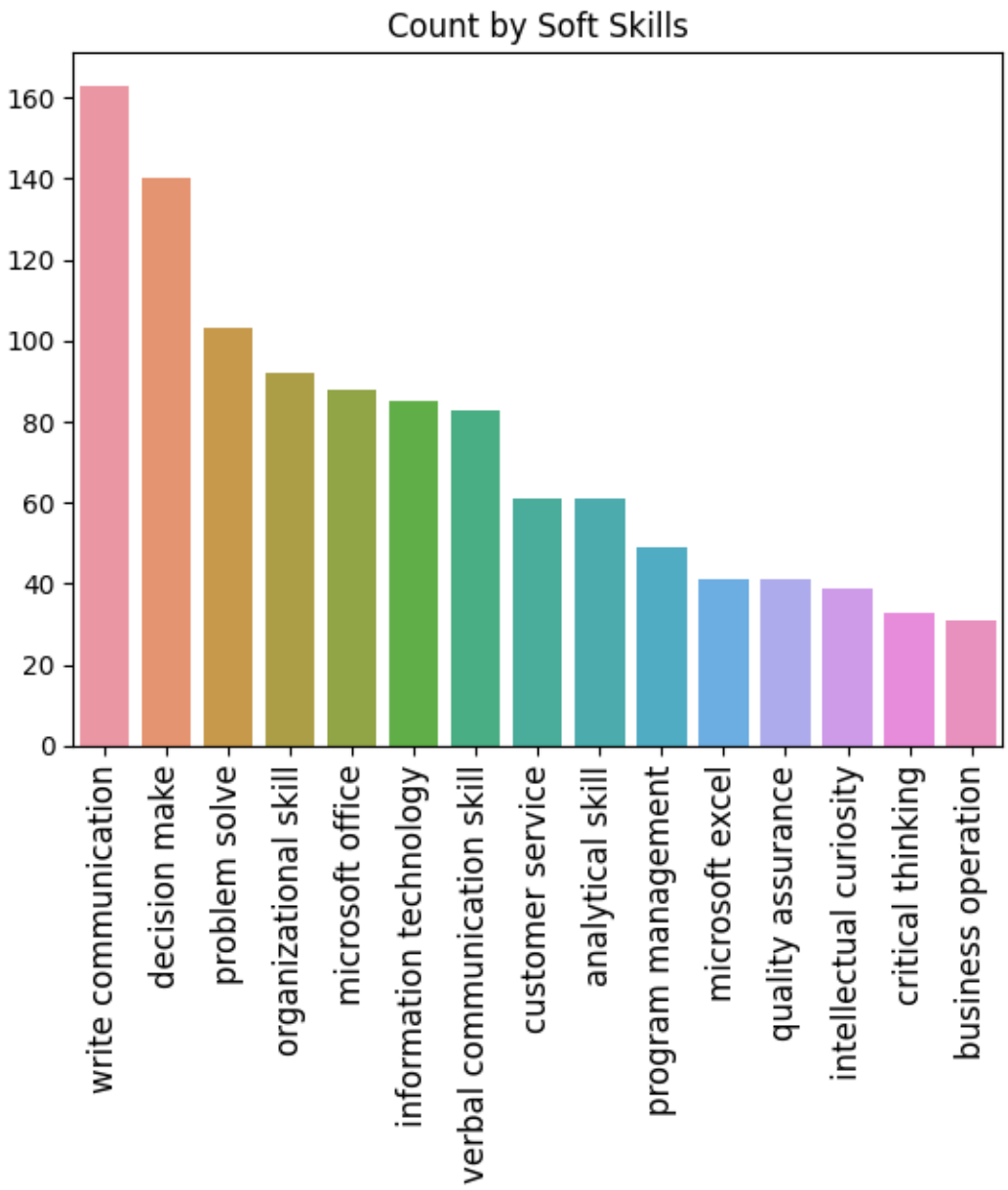


Figure 6 Count by Soft Skills

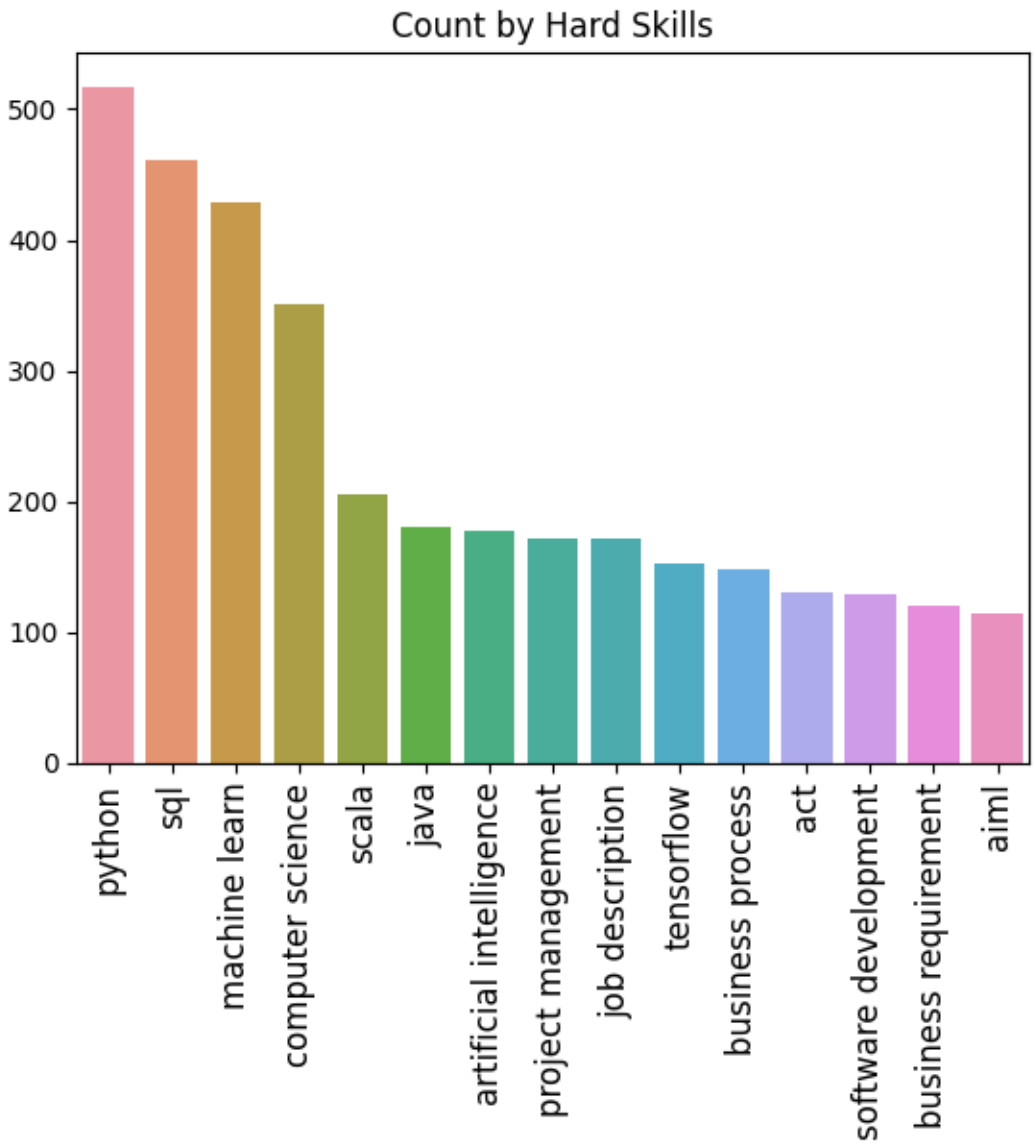


Figure 7 Count by Hard Skills

	decision make	problem solve	organizational skill	verbal communication skill	python	sql	machine learn	computer science	artificial intelligence	project management	...	devops	scala	scikit-learn	tensorflow	keras	lean six sigma green belt	iiba	certify scrum master	mba	cisa
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	1	1	0	0	0
2	0	0	0	0	0	1	0	0	0	1	...	0	1	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Figure 8 Datatable Structure

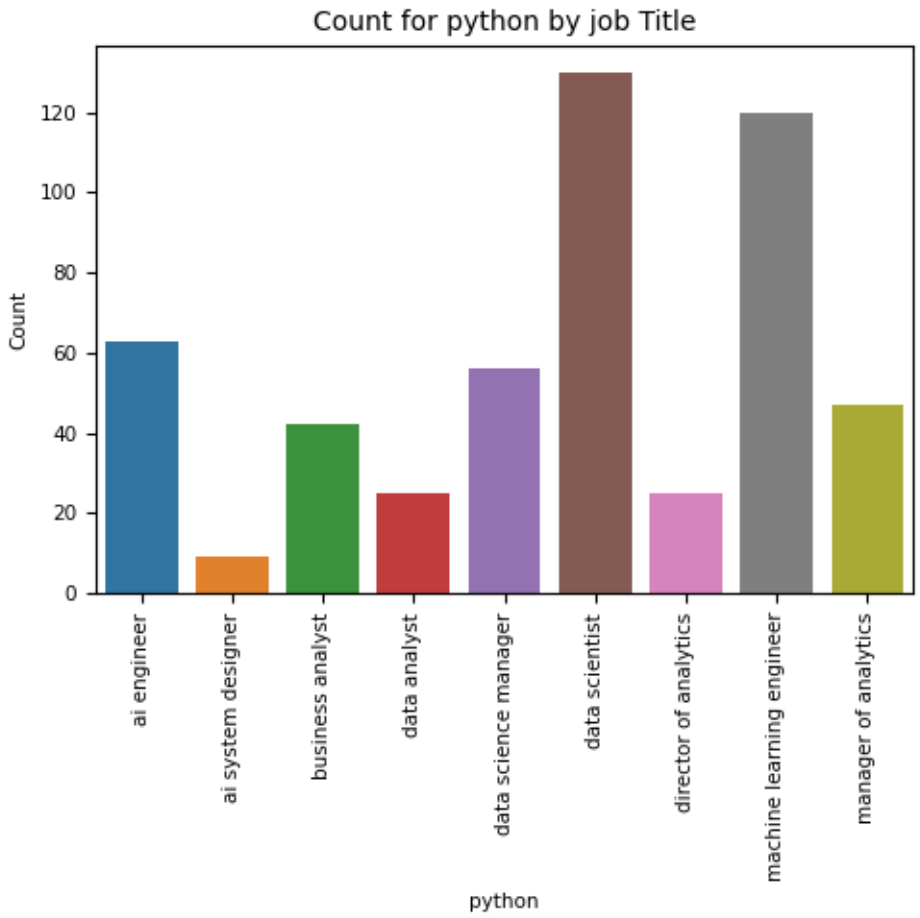


Figure 9 Skill Count by Job Title

	Skills	Salary avg	Rating avg	bachelor avg	master avg	Title	City	State	Soft Skill	Hard Skill	Certification
0	write communication	120000	3.9	0.3	0.2	machine learning engineer	remote	remote	1	0	0
1	decision make	120000	3.7	0.4	0.2	manager of analytics	remote	remote	1	0	0
2	problem solve	110000	3.8	0.2	0.1	business analyst	remote	remote	1	0	0
3	organizational skill	110000	3.6	0.5	0.1	business analyst	Iselin	NJ	1	0	0
4	verbal communication skill	120000	3.6	0.4	0.0	data science manager	remote	remote	1	0	0

Figure 10 K-mean feature engineer table

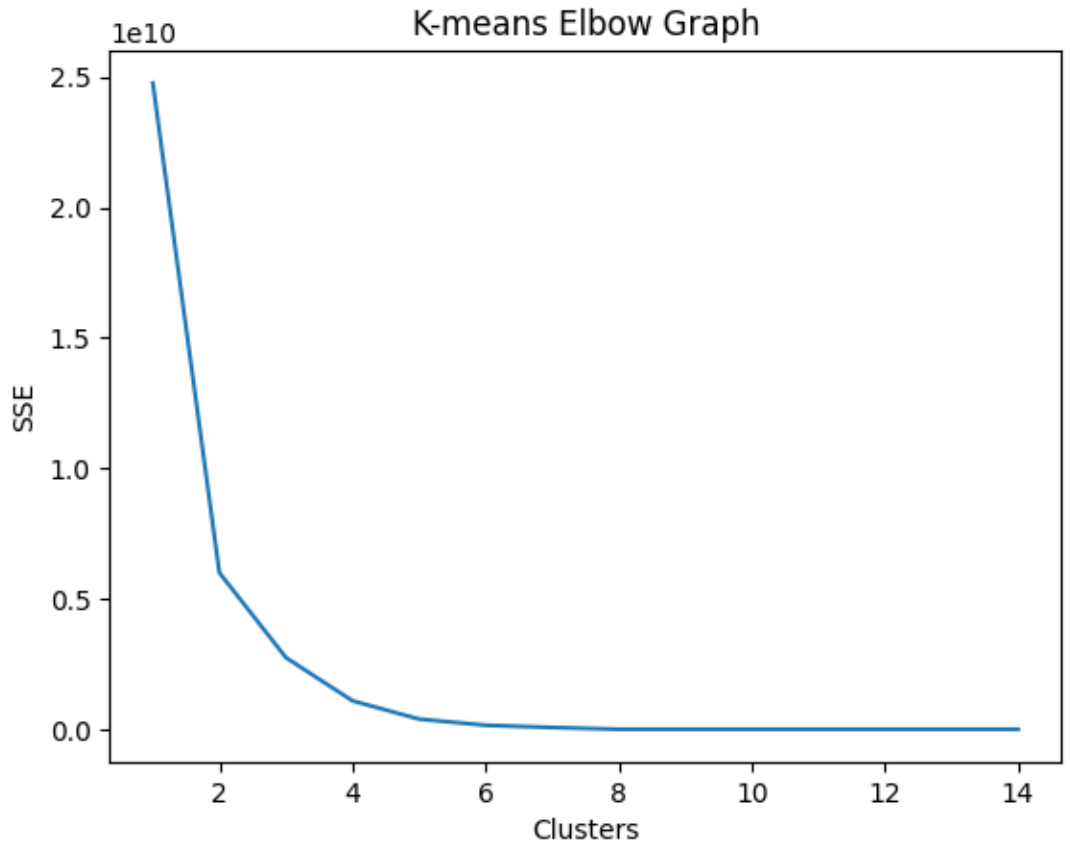


Figure 11 K-means Elbow Method

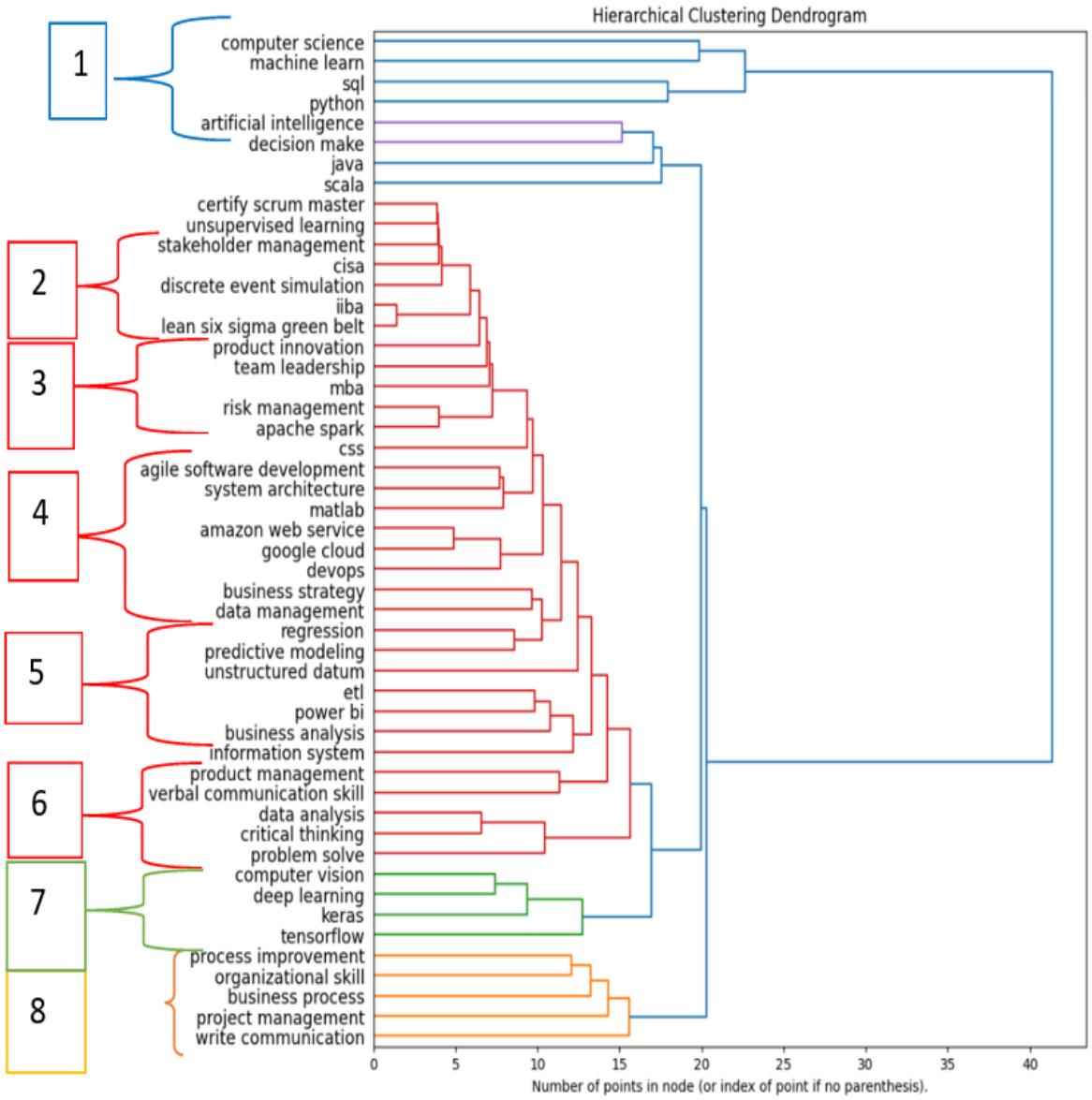


Figure 12 Dendogram

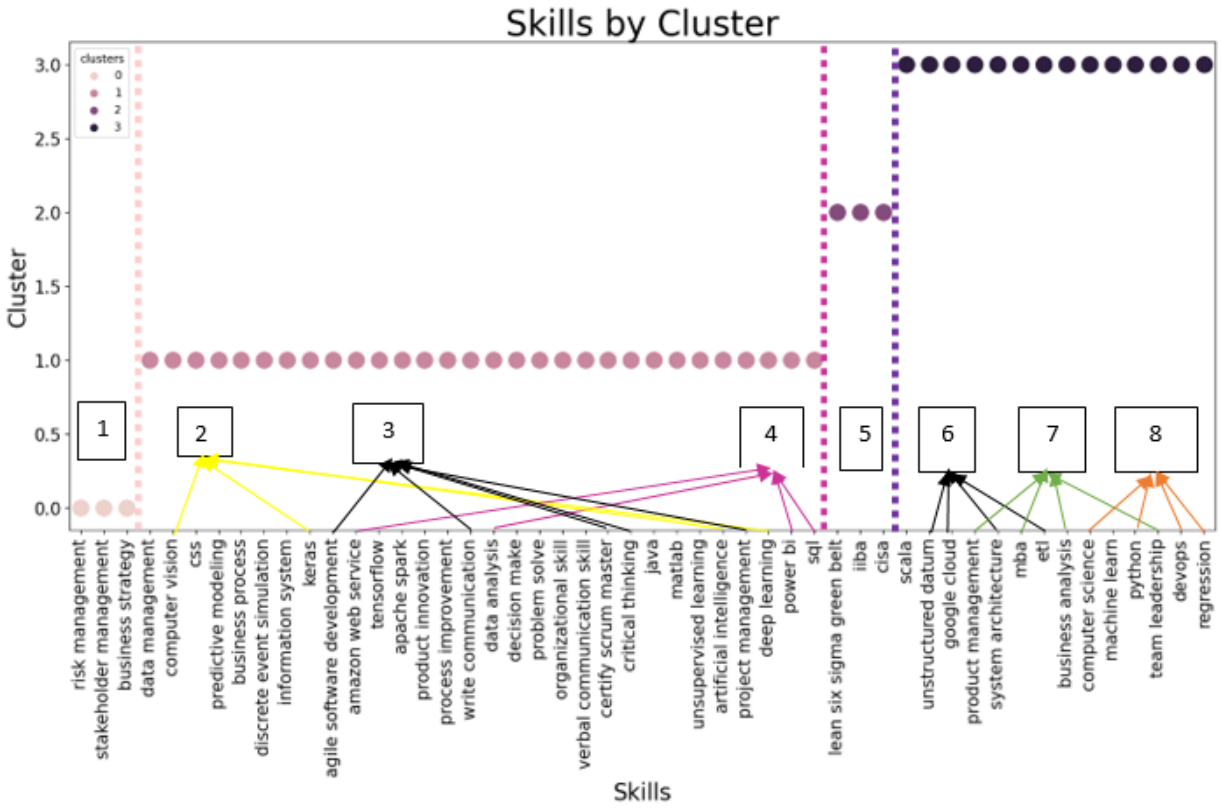


Figure 13 Skills by Clusters

Hierchical				K-mean			
Course	Cluster	Semester	Skills	Course	Cluster	Semester	Skills
MIE 1666 Introduction to Machine Learning	1	1	Machine learning, python, sql, artifical intelligence, computer science	APS 1001 Project Management	1	1	six sigma, iiba, cisa
APS 1012 Management of Innovation in Engineering	3	1	Product Innovation, team leadership, mba, risk management	0974 Business Strategy	3	1	risk management, stakeholder management, business strategy
MIE 1628 Big Data Science	4	1	Agile Software Development, System Architecture, Matlab, AWS, google cloud, devops, business strategy, data management	MIE 1628 Big Data Science	8	1	AWS, data analysis, power bi, sql
MIE 1624 Introduction to Data Science and Analytics	5	1	Regression, predictive modeling, unstructured datum, etl, powerbi, business analysis	MIE 1624 Introduction to Data Science and Analytics	7	1	Computer science, python, devops, regression
2565 - Tools & Techniques of Business Process Management	2	2	stakeholder management, cisa, discrete event simulation, iiba, six sigma	3523 Agile Essentials in Project Management	5	2	Agile software development, writing communication, scrum master, critical thinking, project management
TEP 1502 Leadership in Product Design	6	2	Product Management, Verbal communication, data analysis, critical thinking, problem solving	MIE 1517 Introduction to deep learning	4	2	computer vision, keras, artificial intelligence, deep learning
MIE 1517 Introduction to deep learning	7	2	Computer vision, deep learning, keras, tensorflow	INF 1343	2	2	unstructured datum, google cloud, system architecture, etl
APS 1001 Project Management	8	2	Process improvement, organizational skill, business process, project management, writing comunication	APS 1012 Management of Innovation in Engineering	6	2	product management, mba, business analysis, team leadership

Figure 14 Hierarchical and K-mean schedule

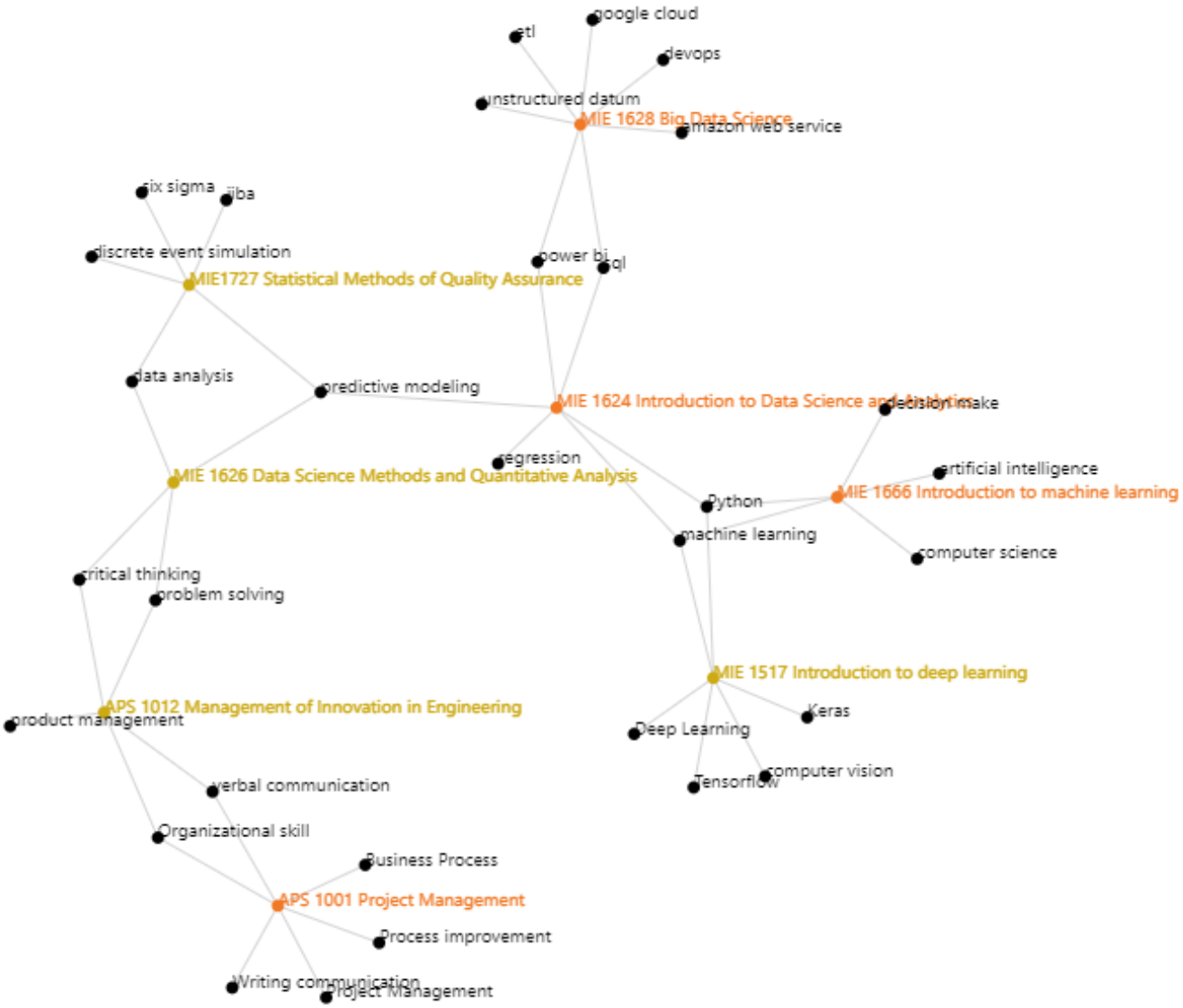


Figure 15 Combine courses curriculum