MIE 1624 Assignment 3 Report

Jiayang Lai

Part 1: Data Collection and Cleaning

For this part, I scraped data containing several on-site and remote jobs, most of which were US postings. There is a total of 4 datasets merged into one final dataset. For the cleaning process, all elements in the data frame “Description” column are converted to the lower letter and generated a new “Description-Lower” column for part 2 expository data analysis.

Part 2: Exploratory data analysis and Feature engineering

After part 1 data cleaning, The NLTK stop words library was implemented for stop words, and several customized stop words were also added to process the job description better. And finally, the “Description-Lower” column was used toward generates several words cloud. The following words clouds are generated as key information:

-Words cloud for all skills

-Words cloud for hard skills

-Words cloud for soft skills

Part 3: Hierarchical clustering implementation

After part 2, the two lists of soft and hard(technical) skills are created to find essential skills employers require from candidates. For soft and hard skills, the program scraps the “Description-Lower” column, counts the total number of appearances for each word, puts it into one data frame, and normalizes it. After having two data frames,

Justification: Hierarchical clustering was implemented based on the count of skills keywords frequency. In that way, we can observe the most wanted skills and filter the skills demand by different clusters, which is more precise than visually inspecting the skill frequency bar plot. Clusters will be used to design electives and mandatory courses or create a new course that fits those skill clusters if those skills are relevant. Based on that thought, two dendrograms were generated for technical and business skills.

The skill dendrograms show that many skills clustered together are closely related. From the hard skill dendrograms, we choose the following clusters (rank by demand descending order):

|  |  |
| --- | --- |
| Analytic Skills | Analytic, Machine Learning |
| Artificial Intelligence Skills | Deep learning, System, Data Mining |
| Data Visualization and Algorithm Design Skills | Algorithm, Data visualization |
| Database Skills | SQL, Python |
| Advanced Data Processing and Visualization Skills | Advanced analytics, Big Data, Language processing, Tableau |

图片包含 图形用户界面

描述已自动生成

The soft skills clusters are as follows.

|  |  |
| --- | --- |
| Management Skills | Business, Manage |
| Research and Creativity Skills | Design, Research |
| Fundamental Social Skills | Plan, Help, Understanding |
| Design Skills | Creating, Responsible |
| Teamwork Skills | Serve, Training, Collaborate |
| Financial Skills | Financial, Reporting |
| Communication Skills | Communication, Presentation |

图片包含 图表

描述已自动生成

From the above clusters, we may consider their frequencies to determine the core, major and elective courses. The pro of this approach is that we can find top-demanded skills separately, and the curriculum can reflect technical and business skills. The cons are that soft and hard skills courses are designed independently, and we need newly designed courses that combine soft and hard skills.

Part 4: K-mean clustering

For this part, the elbow method using distortion was implemented to determine the best k for k mean clustering. From the plot, we observed that when k=11, the decreasing distortion rate tends to flatter. Therefore, the k value will be set to 11. The clustering method will combine hard and soft skills to see if the curriculum can recommend courses without separating technical and business courses. Therefore, the two data frames obtained from part 2 were combined into one for k means clustering to see if high-demand hard and soft skills can be integrated into a new course.

Clustering Result:

|  |  |
| --- | --- |
| Skills | Cluster |
| Business, Analytics | 1 |
| System, Deep Learning, Research, Lead | 2 |
| PowerBI, Hadoop, AWS, Azure, Reporting, Financial, Training, Presentation. | 3 |
| Modelling, Java, Pyspark, Communication, Business acumen, Decision marking. | 4 |
| Algorithm, help | 5 |
| Tableau, Language Processing, Collaborate, Serve | 6 |
| Python, SQL, Plan, Understanding | 7 |
| Advanced Analytics, Creating | 8 |
| Data Mining, Managing | 9 |
| Big Data, Responsible | 10 |
| Machine Learning, Learning | 11 |

The above clusters show that top-demanded soft and hard skills are viewed as one cluster, an advantage that helps design new courses that can target and combine those skills. However, it may also cause that all courses are business oriented, but some technical courses only sometimes need business skills.

Part 5: Interpolation of result and Final Curriculum

Hierarchical Clustering course curriculum

After hierarchical algorithm clusters, many essential skills related to data science are determined. We want to design most skills clusters as courses. Fortunately, the University of Toronto already offers most courses for those skills. The new course will be implemented when designing the curriculum for some clusters that current courses cannot fulfil.

Master of Business and Management in Data Science and Artificial Intelligence

Technical Courses:

|  |  |
| --- | --- |
| Course Title | Topics |
| MIE1624: Introduction to Data Science and Analytics | Analytics, Machine Learning, Python |
| MIE1517: Introduction to Deep Learning | Deep learning |
| INF1343: Data Modelling and Database design | SQL |
| CSC2537: Information Visualization | Data Visualization |
| CHE1147: Data Mining in Engineering | Data Mining |

MIE 1624 can be replaced by APS 1070: Foundation of Data Analytics and Machine Learning

Business Skills:

|  |  |
| --- | --- |
| STA2453: Data Science Methods, Collaborations, and Communication | Communication, Collaboration |
| MIE1622: Computational Finance and Risk Management | Financial |

Electives (Both technical and business skill):

|  |  |
| --- | --- |
| RSM1282: Statistics for Management | Management |
| APS1070: Foundation of Data Analytics | Analytics, Machine Learning, Python |

K-means Clustering course curriculum.

Due to substantial irreverent skills, some soft and hard skills in one cluster cannot be combined into one course. But to fill this gap, the curriculum can be designed with more business courses or take courses from other university departments.

Master of Business and Management in Data Science and Artificial Intelligence

|  |  |
| --- | --- |
| Course Title | Topics Related Clusters from Part 4 |
| MIE1624: Introduction to Data Science and Analytics | 10 |
| MIE1517: Introduction to Deep Learning | 2 |
| INF1343: Data Modelling and Database design | 7 |
| MIE1622: Computational Finance and Risk Management | 3,11 |
| RSM1282: Statistics for Management | 9,10 |
| CSC2537: Information Visualization | 3 |
| New Course 1: Applied Data Science in Financial Sector | 3 |
| New Course 2: Business Programming Application | 4 |

Combining two curriculums as Final Curriculum

From those two curriculums, we can find that the curriculum from K-means is more business-oriented. Still, in general, both curriculums are similar, and it is unlikely to combine some soft and hard skills as one course. In conclusion, Curriculum from Hierechicial Clustering will be modified for students who have more interest in Business. We can create a specialization in Business and develop additional courses for those students to earn that specialization.

Final Course Curriculum combined from Part 4 and Part 5:

Master of Business and Management in Data Science and Artificial Intelligence

Technical Courses:

|  |  |
| --- | --- |
| Course Title | Topics |
| MIE1624: Introduction to Data Science and Analytics | Analytics, Machine Learning, Python |
| MIE1517: Introduction to Deep Learning | Deep learning |
| INF1343: Data Modelling and Database design | SQL |
| CSC2537: Information Visualization | Data Visualization |
| CHE1147: Data Mining in Engineering | Data Mining |

MIE 1624 can be replaced by APS 1070: Foundation of Data Analytics and Machine Learning

Business Skills:

|  |  |
| --- | --- |
| STA2453: Data Science Methods, Collaborations, and Communication | Communication, Collaboration |
| MIE1622: Computational Finance and Risk Management | Financial |

Electives (Both technical and business skill):

|  |  |
| --- | --- |
| RSM1282: Statistics for Management | Management |
| APS1070: Foundation of Data Analytics | Analytics, Machine Learning, Python |

Optional Business Specialization Courses (Choose two of three to earn specialization):

|  |  |
| --- | --- |
| New Course 1: Applied Data Science in Financial Sector | Big Data, Financial |
| New Course 2: Business Programming Application | Business, Java, Python, Pyspark |
| APS1088: Business Planning and Executions for Canadian Entrepreneurs | Business, Planning |