MIE 1624 Introduction to Data Science and Analytics

Assignment 3: Course Curriculum Design

Yunjie Xu 1008601951

Department of Mechanical and Industrial Engineering, University of Toronto

December 9, 2022

Introduction

For this assignment, the aim is to design an applicable course curriculum in line with the needs of the industry for a new "Master of Business and Management in Data Science and Artificial Intelligence" program at the University of Toronto. The dataset includes more than 1400 job descriptions from Indeed's web scraping. These jobs are all from the North American job market, and the job positions are all Data Analyst or Data Scientist.

Part 1

Firstly, I do two web scraping for Data Analyst and Data Scientist separately. Then I combine two sheets, renaming the new sheet as 'webscraping_results_assignmnet3.csv'. I only keep the 'Descriptions' and 'Title' columns because data for other columns are mainly missing or difficult to classify, so I drop these columns. Lastly, I drop duplicate jobs, and there are 1452 jobs left.

Part 2

a/b

I separately extract technical/hard skills and business/soft skills by searching for keywords. Then I create a column for each skill. If the description includes this skill, I use the number 1 to represent that the job needs this skill; otherwise, I use the number 0. After finishing this step, the data transform into a logically formatted data structure for clustering analysis

c/d

In general, 75% of all the requirements of the candidate for the enterprise are about hard skills, and the remaining 25% are about soft skills (Figure 1). For Data Analysts, the proportion of soft skills has reached 31.5% (DS is 24.1%), which means that the standards for measuring Data Analysts are more diverse (Figure 2).

from the perspective of skills, the soft skills that companies value most are communication, leadership, research, passion, and innovation (Figure 3). The essential hard skills can be divided into two different types. One is theories, and the other is tools. The critical tools are R, Python, SQL, Visio, Excel, etc. And the curtail theories are about AI, machine learning, statistics, computer science, etc (Figure 4).

On the side of different types of jobs, the job requirements for da are more straightforward; usually, a qualified DA needs to be good at communication, curious, and proficient in using R, Excel, SQL, Visio, etc (Figure 6). However, the requirements of DS are much more complicated. From the perspective of the cloud map, different DS positions have additional requirements, which leads to

the fact that most of the words in the cloud map are relatively small and dense. Generally, a good DS should handle R, AI, Python, Statistics, machine learning, and research. In contrast, Excel and SQL are no longer essential tools for DS (Figure 7).The DS position has stricter and more varied technical requirements for candidates, and different positions have different needs in detail. The requirements of the DA position are more straightforward, and recruiters care more about candidates' soft power.

Part 3

a/b

The most challenging thing in this part is to find the distance matrix. I redefine some features by myself. For example, I combine SQL, NoSQL, MySQL, and pgAdmin to the new topic SQL. The advantage of this is that the number of features is significantly reduced. And more importantly, the distance between each feature is not limited to 1 or 0. After integrating features, I deleted the technical skills with frequencies less than 100 and other soft skills that cannot be combined (these are mostly character traits, and it is challenging to design courses to improve). Then I used a method similar to the tutorial to calculate the distance matrix and draw the dendrogram (Figure 8).

c

Based on figure 8, hierarchical clustering can divide the skills into a cluster clearly. I make bullet point to these cluster.

- Cluster 1: Python, R, Deep Learning
- Cluster 2: Statistics, Research, Modeling Technology
- Cluster 3: Hadoop, Spark, Big data, Java
- Cluster 4: Power BI, SAS
- Cluster 5: Excel, Visio, Communication

Based on the above results and combined with reality, I designed ten courses for this project. And the result shows below:

1. Course 1 : Introduction to Statistical Programming based on R/Python
2. Introduction to Deep Learning in R/Python
3. Course 3 : Advanced Statistics for Data Science
4. Course 4 : Introduction to Statistical Modeling
5. Course 5 : Introduction to Big Data with Spark and Hadoop
6. Course 6 : Advanced Big Data with Java
7. Course 7 : Business Intelligence in Power BI/SAS
8. Course 8 : Business Practice in Excel/Visio
9. Course 9 : SQL for Data Science
10. Course 10 : Building High-Performing Teams and Running Business

Part 4

Compared with hierarchical clustering, K-means clustering is not flexible enough because I need to pre-determine how many clusters I want to divide. Thus, I use the elbow method to decided the best number of clusters.The elbow method is a heuristic used in cluster analysis to estimate the number of clusters present in a data set. Plotting the explained variation as a function of the number of clusters, the procedure entails choosing the elbow of the curve as the appropriate number of clusters. From the figure 9, the elbow of curve is 13. However, the maximum number of classes is 12, thus I take 12 as the best number of clusters. And the result shows below:

- **Cluster 1**: Optimization, SAS, Java, Spark, Hadoop, Power BI, Big data
  Course Name: Big Data Science with Mutiple Tools
- **Cluster 2**: R
  Course Name: Introduction to Statistical Porgramming: R
- **Cluster 3**: Communication, Visio, Excel
  Course Name: Business with Advanced Excel,SQL,and Visio
- **Cluster 4**: Statistical Learning
  Course Name: Introduction to Statistical Learning
- **Cluster 5**: Cloud Computing
  Course Name: Cloud Practitioner Essentials
- **Cluster 6**: Research, Database, Statistics, Visualization, ERP, Modeling Technology
  Course Name: Advanced Statistics and Modeling Technology in Data Science
- **Cluster 7**: Team Sprit
  Course Name: Building High-Performing Teams
- **Cluster 8**: Deep Learning
  Course Name: Introduction to Deep Learning
- **Cluster 9**: Python
  Course Name: Introduction to Statistical Porgramming: Python
- **Cluster 10**: Data Structure and Algorithm
  Course Name: Introduction to Data Structure and Algorithm
- **Cluster 11**: Commerce Skills
  Course Name: Foundations of Commerce
- **Cluster 12**: SQL
  Course Name: SQL for Data Science

Since the number of clusters in k-means clustering is pre-determined, we can see that the number of topics in each cluster is very inconsistent. Some clusters contain only one topic, while others contain many, which is very different from hierarchical clustering. K-means clustering tends to group less relevant topics, which is why some clusters have many topics.

Part 5

In the K-means clustering, the project has too many skills, which could make the project design more challenging, and it is also hard for students to properly absorb the abilities they should obtain in one or two semesters. Besides, the course curricula built from hierarchical clustering may better identify business and technical skills and give a curriculum containing both business and technical stream courses. Those two streams of courses supplied can better agree with the program's need to focus on technical and soft skills.

Course curriculums will be shown in the appendix (Figure 10 and Figure 11).

Appendix

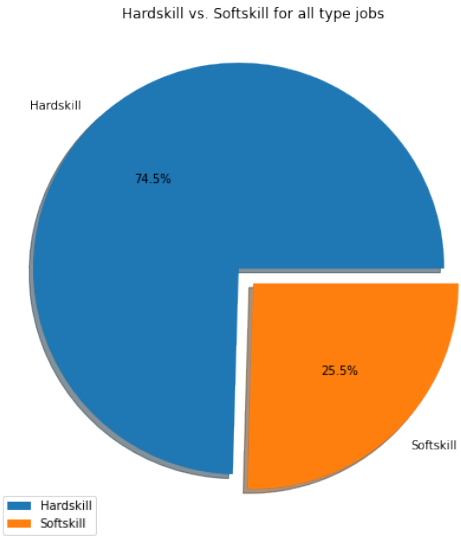

Figure 1: 'Hardskill vs. Softskill for all type jobs'

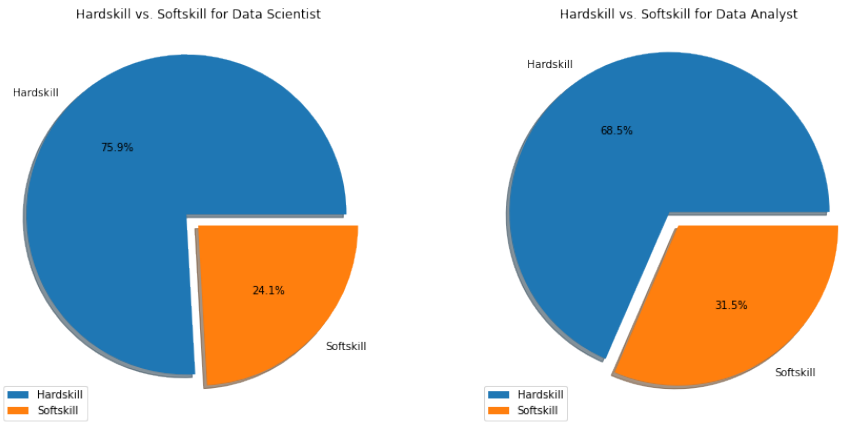

Figure 2: 'Hardskill vs. Softskill'

Figure 3: 'The Worldcloud of Soft Skills'

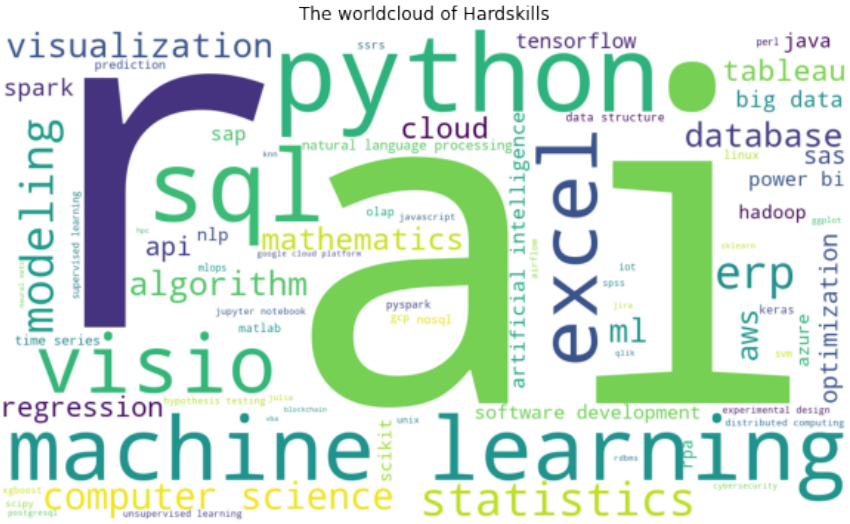

Figure 4: 'The Worldcloud of Hard Skills'

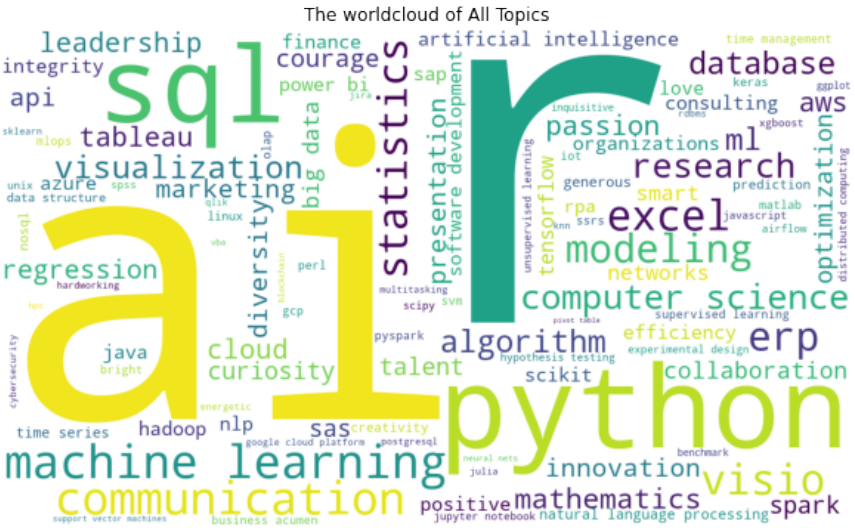

Figure 5: 'The Worldcloud for All Topics'

Figure 6: 'The Worldcloud for All Topics to DA'

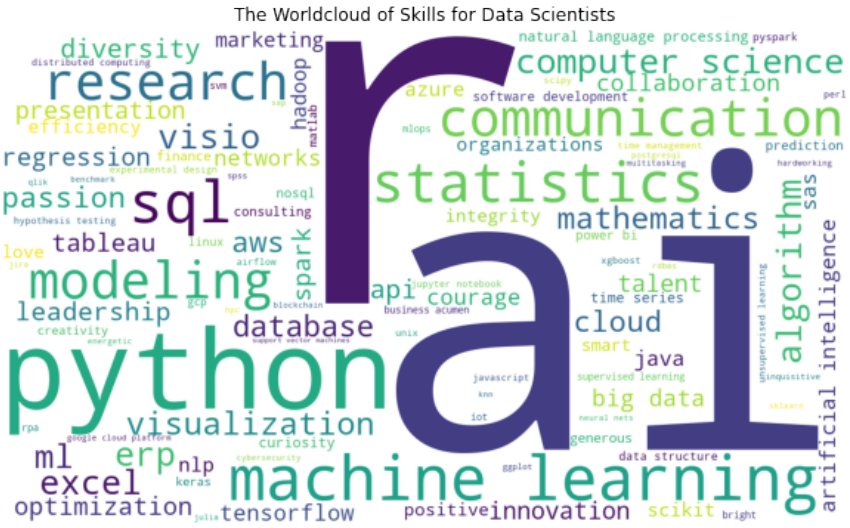

Figure 7: 'The Worldcloud for All Topics to DS'

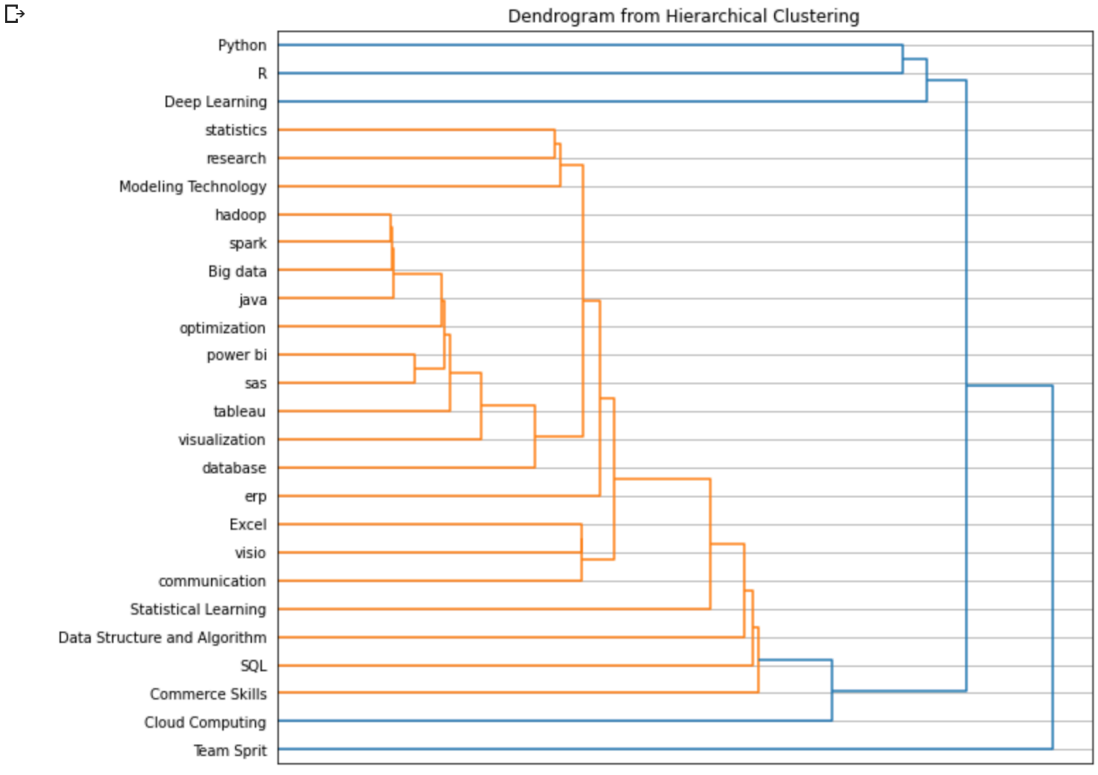

Figure 8: 'Dendrogram from Hierarchical Clustering'

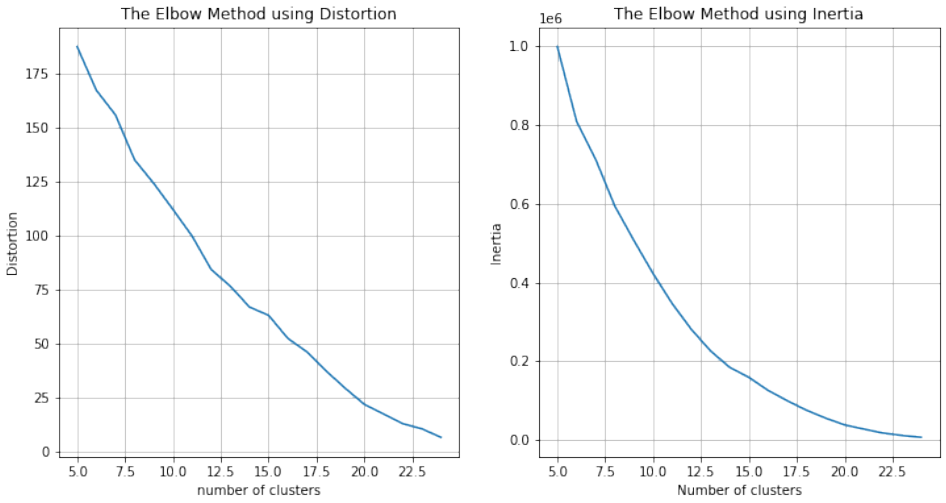

Figure 9: 'Elbow Method Visualization'

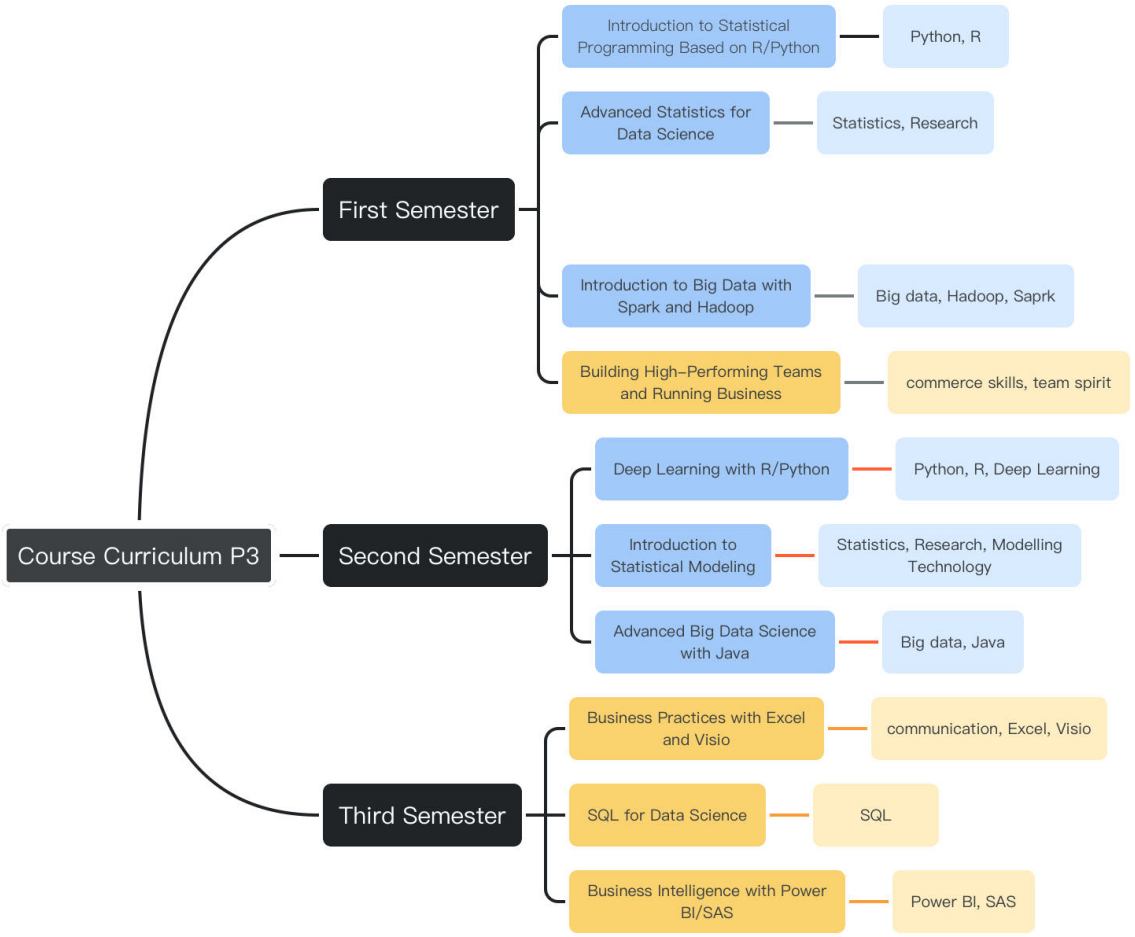

Figure 10: 'Course Curriculum P3'

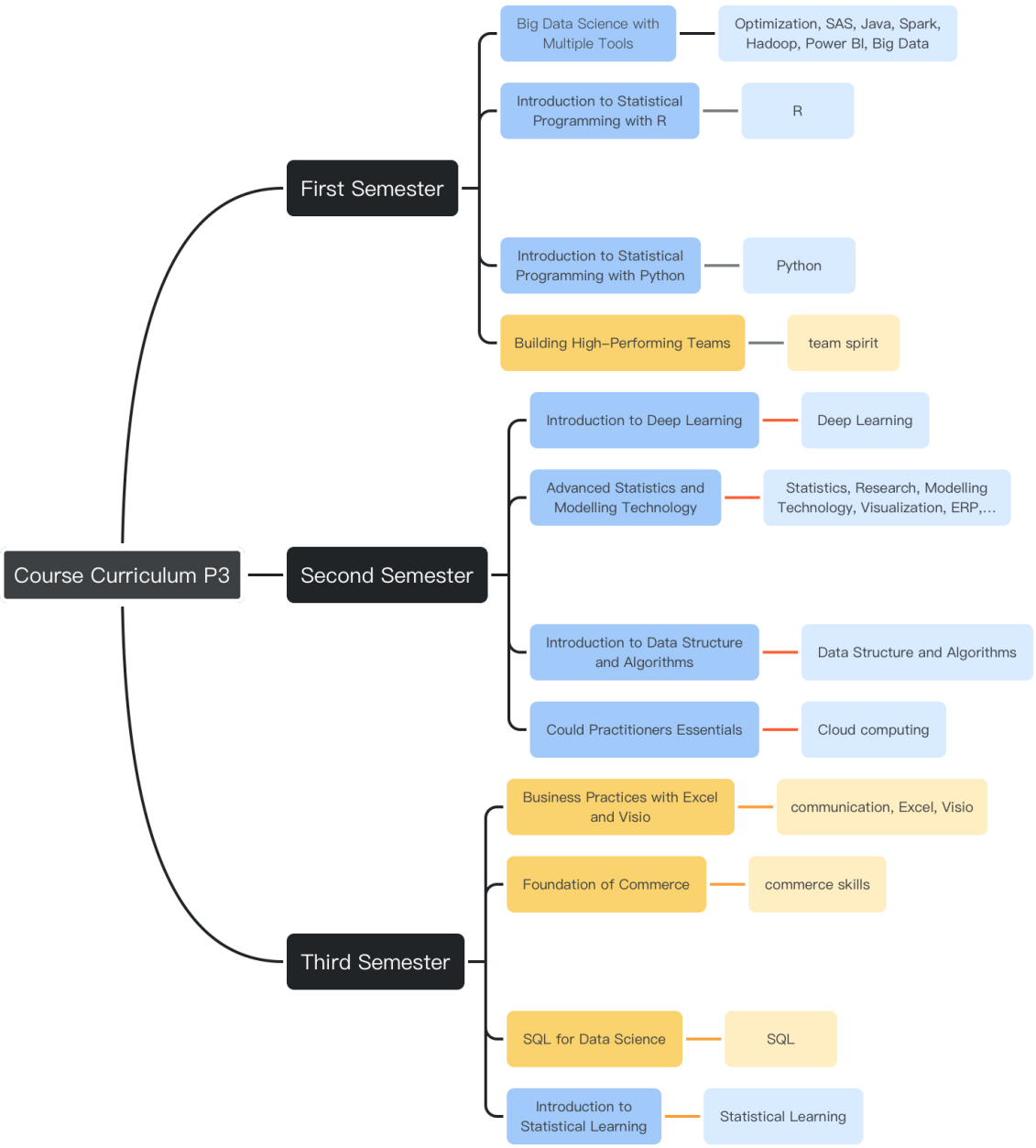

Figure 11: 'Course Curriculum P4'