UNIVERSITY OF TORONTO

MIE 1624 Group Project Report

Course Instructor: Prof. Oleksandr Romanko

Course Title: Introduction to Data Science and Analytics

Group Member: Jiayang Song, Jiani Jia, Liang Cai, Shuhang Li, Shuya Wang, Xiaoxi Yu

DATASET COLLECTION

In the first part of the project, we are expected to redesign the course curriculum for MIE 1624. We have to find technical, also business, and soft skills for students they should obtain. Our data are retrieved by two kinds of methods:

1. Fetching existing data source from Kaggle

2. Real-time data scraping from websites (Indeed and Coursera)

The reason we used two methods to acquire data is that for existing data from Kaggle, there is a high volume of duplicated data which may reduce our evaluation accuracy. Also, those data have not been updated by time and technical terms might be out of date. The drawback of our scrapped data is we were limited by the size of our data source. We may not get an insight view of those in-demand skills due to the lack of data.

By scraping data from Coursera, we extract the course name, course introduction, significant skills, score, course rating, and course URL of each course. We use two sets of course names to query the course content. In terms of technical courses, we fetch courses related to “data science”, ‘machine learning’ and ‘artificial intelligence’. When dealing with business skills, we query the courses with ‘business analytics’, ‘financial modelling’, and ‘financial analytics’.

On the other hand, we scrap thousands of job postings from Indeed by different position names. We notice that 'Data Scientist', 'Data Engineer', 'Data Analyst' are three common job roles in the world of data science. We want to find out those high frequent and in-demand skills required by most companies. Thus we extract the job title, company name, location, job descriptions, and job URLs.

We fetch and store the URLs for data from Coursera and Indeed in order to remove the duplicates data during the data preprocessing steps. Those duplicated data may add a significant amount of noise.

TEXT PREPROCESSING

The text preprocessing includes 8 parts: remove HTML tags and attributes, remove URLs, remove the non-letter characters and white space, remove stop words, lemmatization and tokenization.

Since our data was partially scraped from the Indeed website, so a considerable number of HTML characters and URL links had been enclosed in our data. We need to drop the redundant information to obtain a cleaned dataset and it is beneficial for skill extraction in later steps.

TECHNICAL AND BUSINESS SKILL EXTRACTION

The skill extraction had been designed into 2 parts. The first one is extracting skills from the Coursera dataset which was scrapped from the Coursera website. This dataset covered the syllabuses and contents from the related courses on Coursera. The skills from this dataset are accurate and specific, which means there were fewer redundant or meaningless words and phrases in these data. Considering the concise and precise features of the Coursera website, we took these skills as a base for projection, which means any skills from elsewhere should merge with the skills from Coursera data otherwise they would be judged as invalid skills. (Fig 2)

The techniques used to extract skills from Coursera are the most frequent words and Noun phrases. We extracted the top 3000 frequent words and Noun phrases as skills from the Coursera dataset. The grammar pattern used for Noun phrases is Noun-Noun combinations. We do not prefer any Verbs, Adverbs or Adjectives showing up in our skill list.

The second part is extracting skills from job descriptions. These data were collected and scraped from Kaggle, Glassdoor and Indeed. Same text preprocess had been applied to these data. The purpose is to find out the skills based on job market demands. We want to get skills that are required by hiring organizations and companies, which are more industry-oriented.

There are three steps in skill extracting in the second part: TF-IDF/ N-grams cross-filtering, Coursera projection filtering and manual filtering. Step one is to extract skills from job descriptions with TF-IDF and N-grams. TF-IDF was used since some words showed up frequently across all the job descriptions such as team, job, communication, etc. They were meaningless and might have over shallow other important skills, so we used TF-IDF to prevent this situation. The N-grams were applied because some skills were combined from two even three words, like big data, machine learning, neural networks. These combination-type skills were important and some of them were core-skills on the technical side. So, we need to extract these combinations out from the job description text.

After the initial extraction, we already had many words and phrases, but many of them are not skills, like team, people, work, etc. We demanded to get the unique and valuable skills specific to the technical and business side and we noticed these redundant words are similar in datasets from both technical and business jobs. Thus, we decided to implement a cross-filtering between two job skills that we remove the words and phrases which had high TF-IDF scores in both datasets. Also, we added some exceptions to avoid removing some important skills in both areas, like Python, SQL, statistics, etc. After this procedure, we had skill lists that were clearer and more concise.

The second step was to project skills from job descriptions to Coursera skills. We filtered skills based on Coursera skills that the skills and grams not in the Coursera skill list had been removed. This step can greatly narrow down the skill list size to 100. Most redundant words and grams had been eliminated, and the skill list was easy to have an overview of.

The last is to manually filter out some nonsense skills based on observation such as analysis machine, privacy, science data, etc. These are words and grams also frequently showed up in the datasets and contained some

information but combined in the wrong way. After the manual filtering, the number of extracted skills was limited to 50. And most of them were meaningful and important skills. (Fig 3)

MAIN SKILLS ANALYSIS

Since we are tasked with redesigning the curriculum for MIE 1624 and we should focus on technical perspectives. To start, we first explore the scraped data from Coursera in terms of technical skills and business skills (Fig 2). Those technical skills extracted from Coursera can be used as a base for skills analysis. Then we analyze the skills extracted from Indeed Job postings.

We are going to start our analysis by looking at the most popular technical and business skills on Coursera. Most top key skills have been covered in our current curriculum such as machine learning and data analysis. Also, many important machine learning skills appear frequently such as regression, network, and deep learning. However, other important data science-related topics have not been covered in this course: Cloud and SQL. Usually, a data Scientist analyses various forms of data stored in the Cloud. Organizations are rapidly storing vast sets of data on the cloud with the advent of big data. It can be found that Cloud computing and SQL are widely used in data science courses and they should be covered in our class in the future. If we look at the most frequent skills extracted from business courses. We can find that Microsoft excels is another high frequent key skill that is not included in our current curriculum. Results state that Excel and Microsoft Excel are popular skills/tools covered in those business-related courses on Coursera.

Now, if we look at the technical skills extracted from job postings of Indeed (Fig 3). We would like to assess different skills from the career opportunity perspective. If a technical skill appears more frequently than others from postings, it can indicate that students who acquire those skills are more likely to find a job. Figure 3 clearly shows the distribution of key skills that a candidate should have from a company's perspective.

The most in-demand skill is SQL. As we discuss in the last part of Coursera analysis, SQL is also a high demand skill that is widely used on Coursera. There may be data from several sources. Data scientists may need to build their own database, and they may then store or erase data from it. For extracting data from databases, we need SQL. After that a certain phase of data cleaning takes place. Based on this, they can implement machine learning models. Moreover, many machine learning related topics have been shown hundreds of times thus we should focus on machine learning skills as well. In this case, Azure is the seventh most frequent skill that is required by data science positions, along with cloud computing. We should definitely introduce cloud concepts in the class in order to build a close connection to the industry.

From the Indeed skills analysis, the top skills required for data science students are: SQL, machine learning, statistical data, algorithms, azure, apache-spark and AI.

Based on the analysis of different data sources, more technical skills and tools should be included in this course. For example, we find cloud computing is broadly used in the data science field. We should introduce some useful cloud computing terms and assign a few handy cloud assignments for students. We find that Azure and AWS are the most popular cloud computing platforms in the industry. Moreover, SQL is a popular programming language that is highly in-demand in the industry. We notice that we don't have any database related materials covered in the class. We should introduce those basic concepts of database, particularly relational database concepts and SQL queries. Other concepts such as basic statistical studies and machine learning concepts we can keep as what we have for now.

Our group re-design the curriculum of MIE1624. We create a new schedule and list those significant topics that should be covered in our class. Every week certain material should be delivered to help students develop required skills

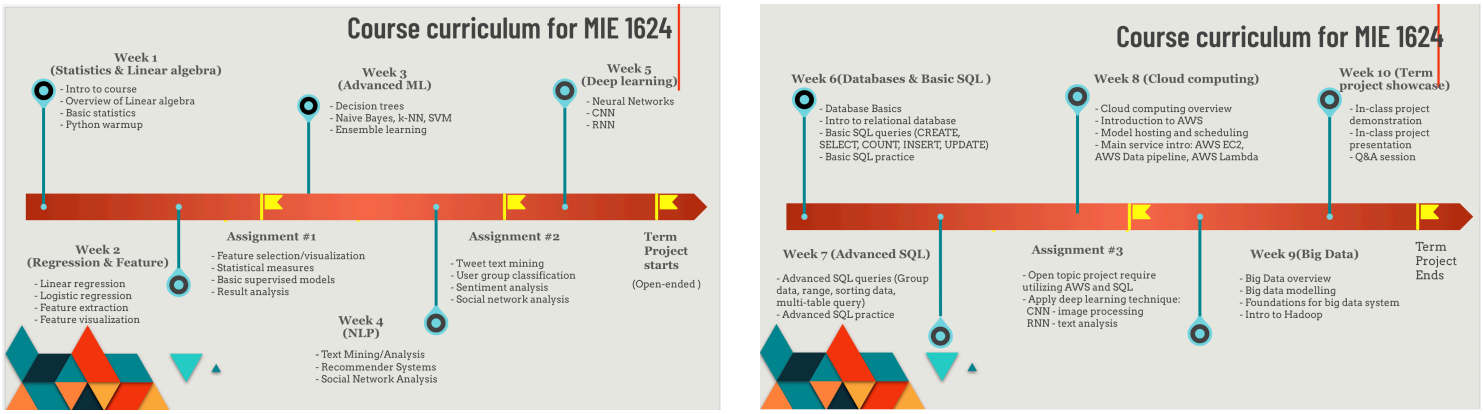

Figure 1: Course Curriculum for MIE 1624

MASTER OF DATA SCIENCE AND ARTIFICIAL INTELLIGENCE PROGRAM CURRICULUM DESIGN

In nowadays, data science and artificial intelligence have become increasingly important and play a dominant role in the technology and business industry. However, the talent gap in the industry is still very large. Therefore, our group has decided to design a new program named "Master of data science and artificial intelligence" which is both technically and business oriented. From this program, students could gain academic knowledge and hands-on experience at the same time and will be ready for future challenges.

The structure of the master program is composed of two parts: compulsory part and optional part. The compulsory part is made up of 5 core courses including aspects of business communication, machine learning,

data science, database and statistical analysis. All the core courses are designed based on the skills selected from our previous data preparation. Some top skills like data visualization, database, python, cloud, ML and so on are all covered here. These courses can give students a solid foundation of data science and AI and be ready for more advanced topics in the future.

For the optional part, students are free to choose a business stream or technical stream. In the business stream, it emphasizes real-world management analytics and the usage of predictive tools in machine learning and artificial intelligence for the innovation and tech-driven economy. Some business courses include team projects which can enhance the public speaking and presentation for students . In the technical stream, it provides students with advanced statistical tools in order to properly analyze complex and large data and how to prepare and interpret visual representation of complex and large data. Most technical courses have the team projects part that could give students practical working knowledge and team-working environment.

Overall, students need to take five compulsory courses and choose 5 courses from the two optional streams in order to get the master degree diploma. After graduating from the program, students should master the skills required to succeed in the field of management analytics, with cutting-edge coursework that teaches you the newest techniques in machine learning and artificial intelligence, enabling them to tackle the business challenges of today and tomorrow.

The next three tables include the course names and descriptions for our thirteen courses. The five courses in compulsory part include both technical and business basic knowledges in the field of data analysis. The courses cover the basic statistical knowledges and various predictive modelling using R language; technologies for big data; introduction for machine learning and Python implementation; data visualization and business communication skills; database management and database design.

Compulsory Course	Description
Statistical Predictive Modeling For Analytics	The course introduced to the basic concepts in predictive analytics, it is popular in data mining. The process of analytics involves specifying a question, problem, or decision, and finding the right answers using data, Our course will teach appropriate models, tools, and methods for analysis. Allow students to develop and use advanced predictive analytics method, develop the skills to use popular tools and software for predictive analytics and apply these methods to answer the respective questions.
Visualizations And Business Communications	Communicating with data is an essential skill for professionals in any industries. The course designed to developing the visualization concepts skills and gaining a working knowledge of data visualization programs. We will introduce the fundamental tools for presenting data analysis using visuals, tables, information graphs, and reports. Learn the verbal and visual presentation approaches, and techniques to better communicate with and business stakeholders. How to best communicate with these business decision-makers using data visualization tools, such as Tableau, Matlab, Python, excel. Students should be able to participate collaboratively and responsibly in teams and to reflect upon their own contribution to the team.

UNIVERSITY OF TORONTO

Compulsory Course	Description
Big Data	The emphasis of this course is on mastering two most important big data technologies: Spark and Deep Learning with TensorFlow. The course begins with a basic introduction to big data, loading data and handling files in Hadoop, get data from Hadoop and querying big data with Hive. Combine big data with machine learning, learning machine learning tools such as Spark, Azure. It also provides a first hands-on experience in handling and analyzing large, complex data structures.
Introduction to Machine Learning	Machine learning is wide using in modern data analysis, this is an introduction to machine learning. We will first provide an overview to machine learning, and review interdisciplinary techniques such as linear algebra, probability theory and distribution, information theory, decision theory. Learning the concepts behind several machine learning algorithms without going deeply into the mathematics and gain practical experience applying them. Course will also teach students how to extract and identify useful features that best represent the data, some of the most important machine learning algorithms, and how to evaluate the performance of your machine learning algorithms.
Introduction to File and Database Management	The primary goal of this class is to learn principles and practices of database management and database design. Explain the relational model, normalization, and how to transform an entity-relationship data diagram into a relational model. The standard navigation language for relational databases SQL will be introduced, learning the two-tier and three-tier architectures, and the internet database environment, explain data warehouse architectures. The roles of data administration and database administration, their function, and their importance to an information resource will be discussed.

Four optimal technical courses involve more mathematics and programming, students will learn different data structures and Implement a variety of algorithms, providing an overview to recent neural network algorithms, solving the problem using the knowledge of neural networks and deep learning, using and effective current techniques and toolkits for natural language processing, and for the data science capstone, students should combine different techniques, build and test model to solve the practical data problem.

Optional Technical Course	Description
Data structure and Algorithm	Achieve an understanding of fundamental data structures and algorithms, the focus of this course is on solving computational problems that involve collections of data. Describe abstract data types including stacks, queues, lists, sets, maps and graphs. Implement a variety of algorithms for searching and sorting, including linear search, binary search, insertion sort, selection sort, merge sort, quick-sort, and heap sort. Analyze the time and space efficiency of data structures and algorithms.

Optional Technical Course	Description
Neural network & Deep learning	Deep learning is inspired by a simplified model of how the human brain works by building effective hierarchical representations of complex data. This course gives an overview of both the foundational ideas and the recent advances in neural network algorithms, build from a one node neural network to a multiple features and multiple output neural networks, explore the parameters for neural networks. All the steps are explained using working code to solve problems to have a working knowledge of neural networks and deep learning. This course provides the necessary required background to understand Time Series Analysis and Natural Language Processing courses.
Natural language processing (NLP)	Introduction to Natural Language Processing, the study of computing systems that can process, understand, or communicate in human language. This course is intended as a theoretical and methodological introduction to the most widely used and effective current techniques, strategies and toolkits for natural language processing, with a primary focus on those available in the Python programming language. Students will also learn how to employ literary-historical NLP-based analytic techniques. No prior knowledge of digital technologies or computer programming is required for this course.
Data Science Capstone	This is a capstone course, students should work in a 5 people group to complete a project for the whole semester. The project require student to integration and application of knowledge and skills gained during the program through hands-on projects supported by our industry partners to build a full data science pipeline from preparing, analyzing and visualizing data to building and testing models. Communication and presentation of insights and recommendations derived from data analysis using visualization and storytelling techniques.

For business courses are focused on business analysis skills, computer simulation model in business situation, how to structure, analyze, and solve business decision problems, building their own decision-making styles. Identify the opportunities for revenue optimization in different business cases and understand the concept of pricing, learning various popular forecasting techniques to improve the forecasting quality.

Optional Business Course	Description
Business Analysis	This course focus on developing analytical skills that are consistently in high demand for employment market. A series of powerful quantitative methods that will be taught to help you making more effective business decisions. Topics include probability, and use of probability in decision-making, assessment utility for different decisions. Explore different difficulty types of models, approaches to model development: reductionist and holistic and how models are developed within a business organization. Discuss cost estimating and various types of business costs. Create and use computer simulation model in business situation risk analysis.

Optional Business Course	Description
Business Decision-Making Through Advanced Analytics	This course focus on how to structure, analyze, and solve business decision problems, aim to improve decision-making skills. Covering the basic decision-making techniques and analysis modelling approaches, examples in finance, marketing, and operations will be illustrated. The course will focus on the how to apply the model rather than the mathematical details. By the end of the course, students are able to use decision analysis techniques, understand their own decision styles and have increased confidence in their decision-making.
Revenue Management and Pricing	How a firm should set and update pricing and product availability decisions in order to maximize its profitability is also called revenue management. In this course, you will learn to identify and exploit opportunities for revenue optimization in different business contexts, understand the relationship between price, demand and supply. Introduce basic concepts about dynamic pricing, to provide information to students about promotions, markdowns and customized pricing.
Economic Forecasting	The forecasting process and the techniques are designed to cover in this course. Students will learn how to examine the data and how to prepare data for forecasting and various popular forecasting techniques will be introduced to improve the quality of forecasting. The concept includes basic statistics and econometrics knowledge, forecasting with classical regression method, time-series modelings, time-series decomposition, time-series forecasting such as ARMA models.

EDTECH STARTUP

EdTech is a program that changes the education material from books to digital content. It enhances the old education system by improving the pedagogy and learning process that ultimately results in enhancing the education system. There's rapid exponential growth in the EdTech field nowadays. The popularity of EdTech has increased after the COVID-19 Pandemic hit the world since the majority of the universities shifted their classes from traditional classrooms to online courses.

EdTech's target audience is not limited to the education sector only, but it is also used for corporate learning in large corporations. These professionals have surged the demand for reskilling and online certificate categories. Thus, our EdTech program, Data Science for Education, aims at helping students approach internships of Canadian companies in data science. It will connect the students with the courses in universities as well as the program in large corporations that are in huge demand of help. While attending this program, students need to take three courses as well as one capstone project. Students must finish their courses first, in order to do well in their capstone project. The courses will be divided into four emphases as well, each emphasis has one compulsory course and two optional courses and is focused on different programming language learning. Since all the courses in this program are composed of several small projects, students will have a lot of coding work to do which could improve their programming skill for certain types of companies. This capstone project is a steppingstone for job seekers in data science to reach out to their desired companies and professors. There will

be four main directions for the capstone project, which major in financial analysis, information technique for IT services, academic research and A.I in big data.

From the job collections we analyzed before, there is a huge demand for data analysts with competencies related to advanced fields in Data Science and Artificial Intelligence. Since different companies differ in professions, sometimes it can be unnecessary and tedious to go through all the course descriptions for students to choose their directions. So, in this program, we would build a recommendation system to help students select courses based on their demands. The demand could come from the job requirement or their own interest. Students may come from different departments who are interested in data science and excited to find a career related to data science. They may not have sufficient technical skill, and management skill to fit the requirement of data scientist jobs. Therefore, this recommender system brings an insight of helping students make up for their missing skills, and gain competitiveness in the data science field.

In order to do that, students need to fill in a questionnaire including several course-related questions such as: 'What kind of job do you want to find?', 'What kind of programming language do you want to learn?', 'What kind of management ability do you possess?', etc. This questionnaire data will rule out the courses with skills you already possessed, find the job required skills and student's interested skills then provide three relevant courses in our previous designed curriculum to this student.

For example, if a student wants to find a job as a data engineer. This job requires python, data mining skill, visualization data, sql and Hadoop. The course selection system will rule out courses provided with MATLAB skill and provide students with courses that improve R skills, python skills, etc.

Since this program also provides students with different opportunities to collaborate with professors or companies, students also need to fill out some project related questions in the questionnaire as well, such as: 'What programming language you master best?', 'What kind of job are you seeking?', 'Which corporation is your dream?', etc. We use machine learning skill to predict these data along with the course-related question data, classify the students into four capstone project directions.

For example, for interns in banks such as RBC and Scotiabank, they require that students not only get a good command of data analysis but also know fundamental financial principles. As for interns in Yelp, this app is based on A.I in big data, students need to do projects using machine learning methodology to analyze realistic problems with big data. Universities like Queen's University, McGill University and University of Toronto are also big chances for students since our professors often collaborate with many companies and do great jobs in academic research. If students want to spend their intern inside the campus, this suits them best.

APPENDIX: GRAPH

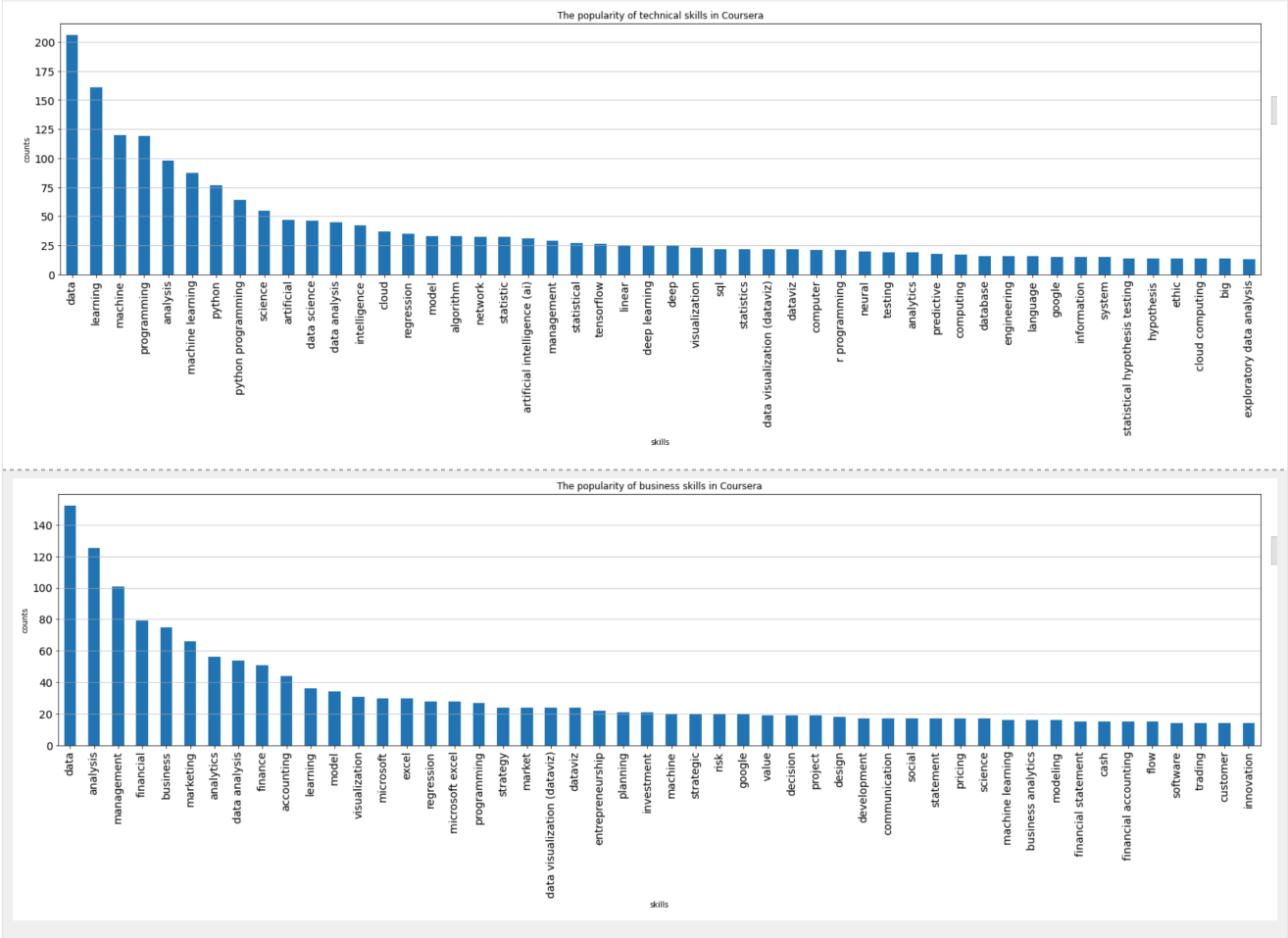

Figure 2: Popularity of Technical and Business Skill in Coursers

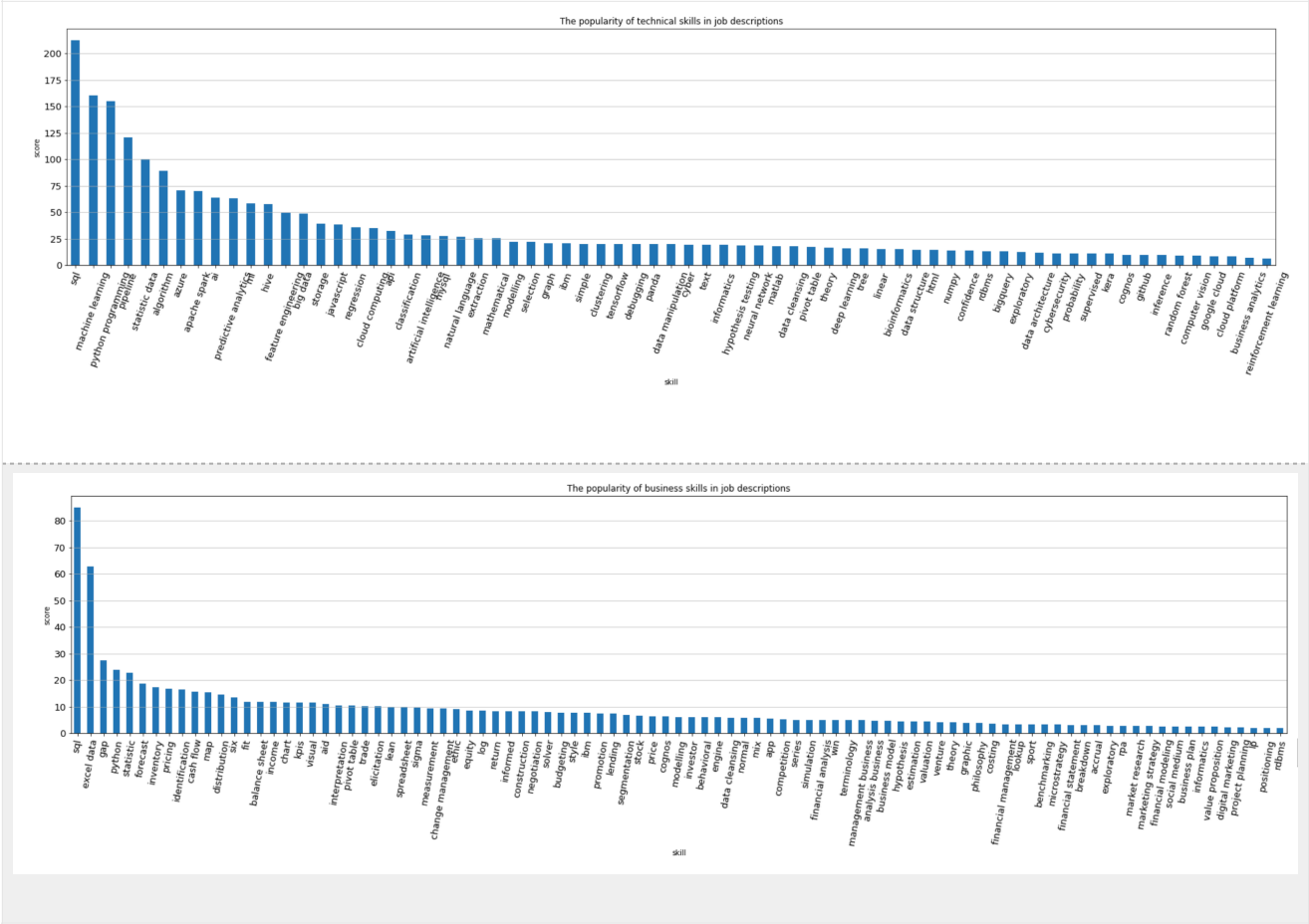

Figure 3: Popularity of Technical and Business Skill in Job Descriptions

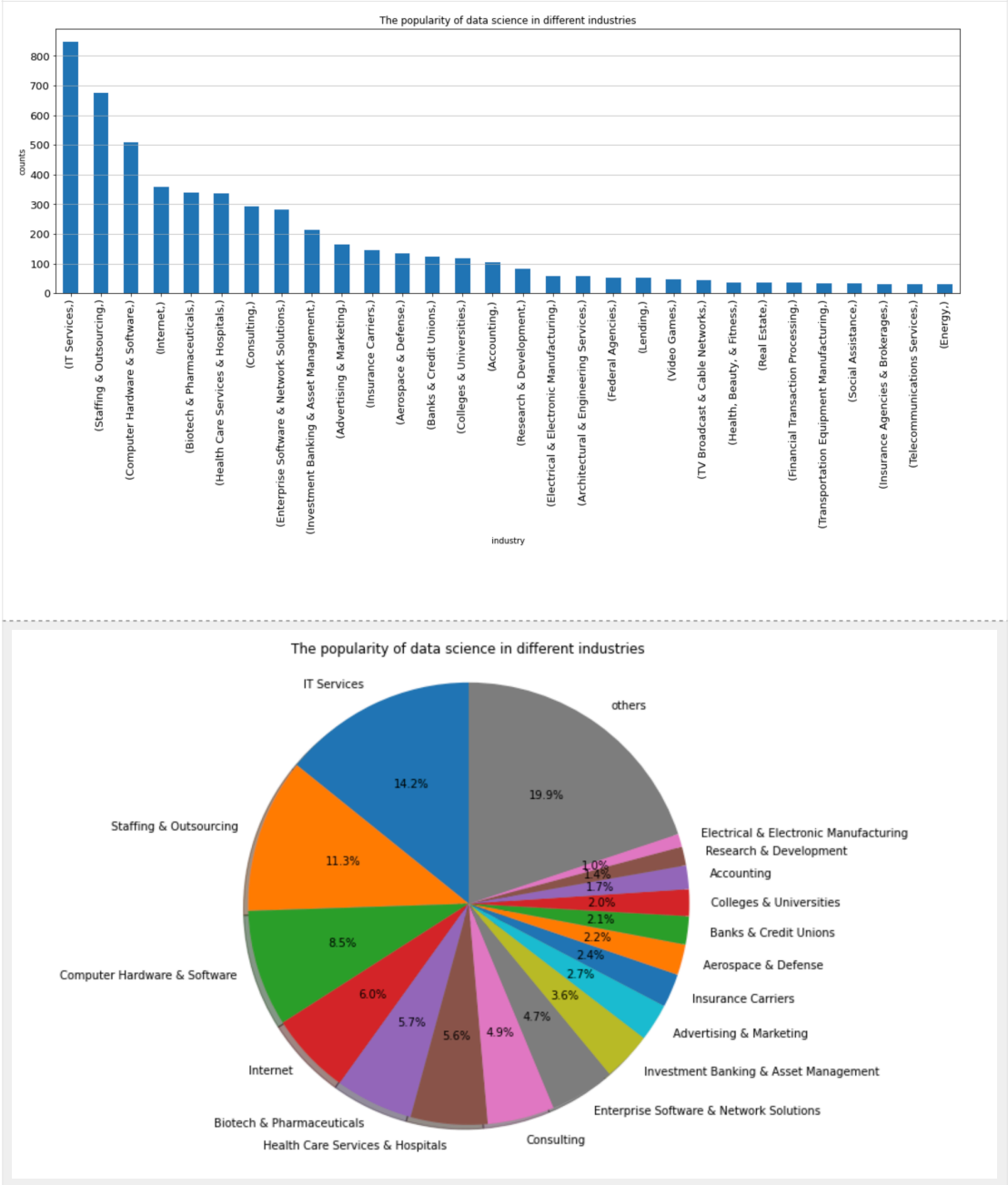

Figure 4: Popularity of Data Science in Different Industries

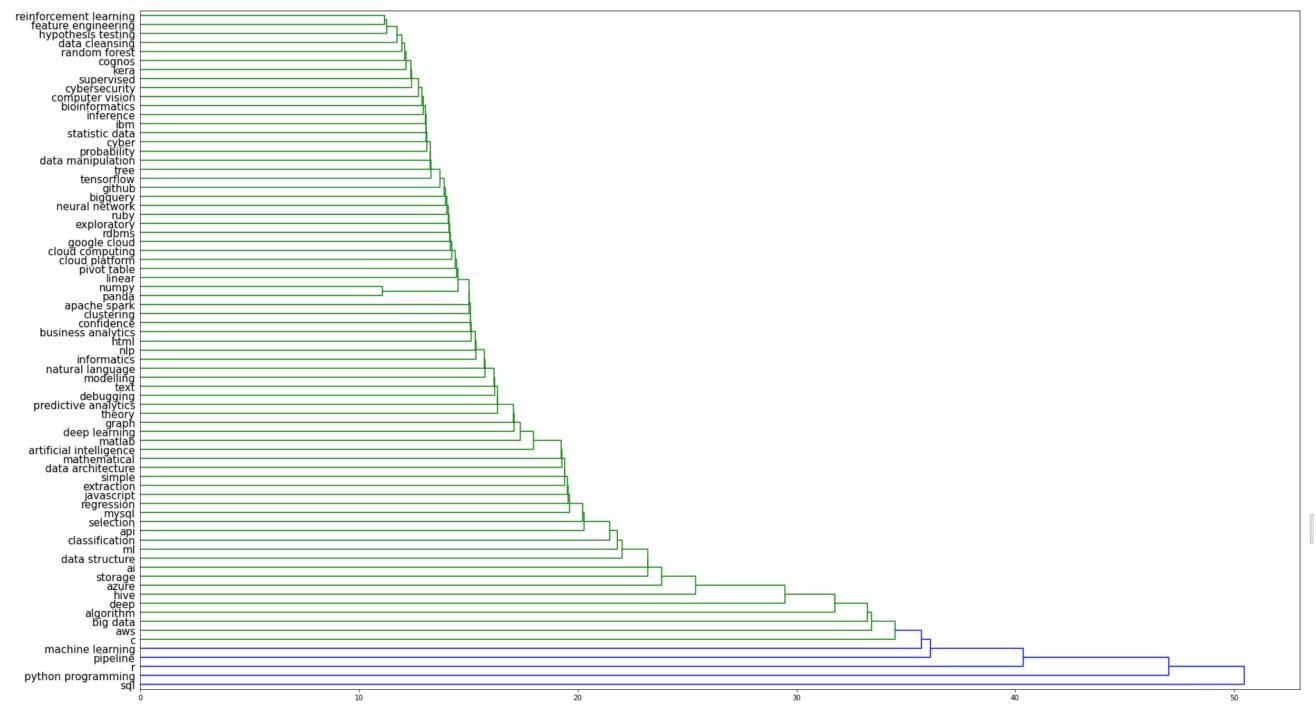

Figure 5: Hierarchical Clustering Analysis for Skills

Figure 6: Program Poster

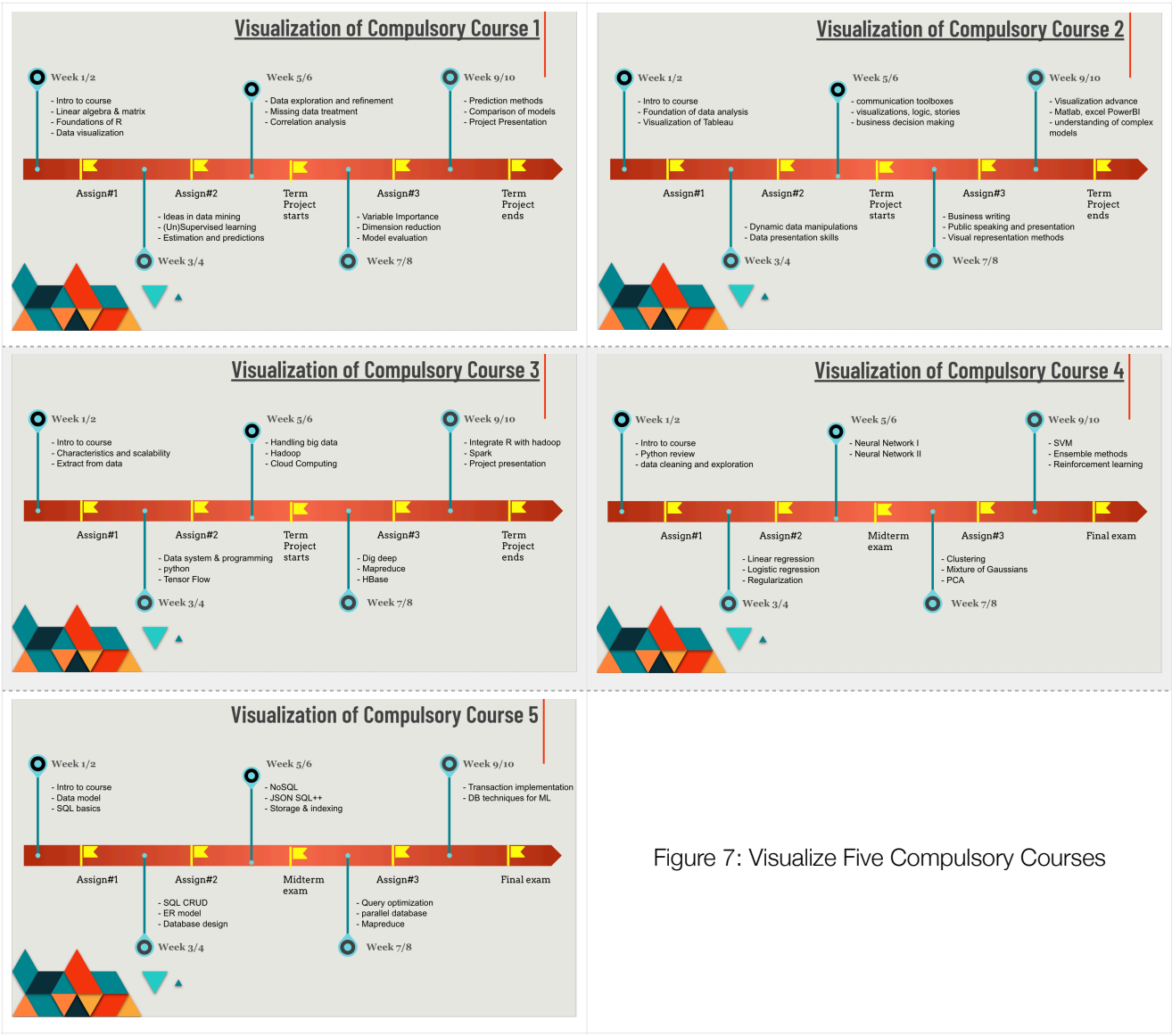

Figure 7: Visualize Five Compulsory Courses

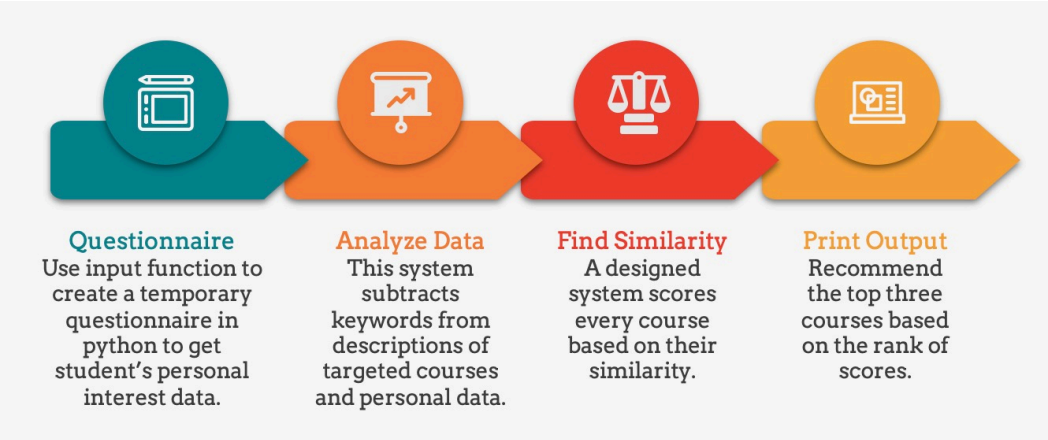

Figure 8: Course Selection System