MIE1624 Assignment 3 report

Ziqingqing Ye

1009167544

Introduction

The goal of this assignment is designing a course curriculum for a new “Master of Business and Management in Data Science and Artificial Intelligence” program at University of Toronto with focus not only on technical but also on business and soft skills. I designed 8 courses for students to obtain necessary technical and business skills to pursue a successful career as data scientist, analytics and data manager, data analyst, business analyst, AI system designer, etc.

Data collection and cleaning

Web-scrapping

I extract skills that are in demand at the job market from job vacancies posted on *http://indeed.com* web-portal and apply clustering algorithms to group/segment skills into courses. I selected the posts of 'Data Science' in Vancouver, BC, Canada, and Chicago, state of Illinois, USA, separately and gather them together as webscraping_results_assignmnet3.csv and this data set contain 1100 unique job postings. Web scraping are done with provided sample Python code.

Data cleaning

Since the dataset has been collected directly from *http://indeed.com*, it is particularly important to clean the data by removing or modifying data that are messy. Before start to cleaning data, I use *.value_counts()* method to check the frequency of words and listed them in descending order. The first five words are ‘and’, ‘to’, ‘the’, ’of’ and ‘in’ which are adverbs unrelated with course designing.

Detailed steps for cleaning data are listed below:

- Convert all words to lowercase.
- Remove irrelevant information, including username starts with @, hash tag #, links starts with https and http, stop words package download from *nltk* library, non-alpha characters, numeric and extra space in the text.
- Use *.split()* method to convert to list from string.
- Use *PorterStemmer* to reduce inflected words to their word stem, base or root form—generally a written word form
- Lemmatization. Restore the attributes of words to the most normal form: such as removing the tense of verbs, changing plural nouns to singular nouns.

Exploratory data analysis and feature engineering

First, i extract data for technical/hard skills, business/soft skills, position title, company, and any other relevant information from job postings texts. Define key words of skills as ‘communication’, ‘presentation’, ‘problem solving’, ‘project management’, ‘teamwork’, ‘python’, ‘R’, ‘Java’, ‘C’, ‘sql’, and so on. Separating them into hard skills and soft skills and iterating those skills from the ‘Descriptions’ and inserting them to new columns ‘Skills’, ‘Soft_Skills’ and ‘Hard_Skills’.

Then I used the TF-IDF method as feature engineering technique. TF-IDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus and transform text into a meaningful representation of numbers which is used to fit machine algorithms for prediction. I used TF_IDF on ‘Title’, ‘Location’, ‘Company’ and ‘Descriptions’. After feature engineering, we can say that in this dataset, job titles usually include data, analyst, senior analyst and programmer. And most of the jobs are located in Chicago and within Illinois. The companies that provide the most vacancies are Parexel and Procom since they have the first two highest *tfidf* value.

Then I visualize key word with Word Cloud of hard skills, Wordcloud of soft skills, Wordcloud for all skills, Wordcloud of locations, Wordcloud of titles and Wordcloud of companies. I also plot the percentage of occurrence of hard/soft skills in job descriptions. We can tell from the plots that communication, leadership and passionate are particularly important in soft skills. Hard skills such as sas, sql, python and tableau are highly likely to appear in job descriptions, which indicates that the main skills required for applying for data analysis are communication, leadership, python and sas.

Hierarchical clustering implementation

Hierarchical clustering is achieved by using Euclidean distance between single observations in data set and linkage criterion. First, calculate the Euclidean distance between each skills. Then using the dendrogram to plot the clustering result with Ward linkage. In order to design 8-12 courses, i set the threshold line at index 13 which gives 8 intersection points with the dendrogram, which indicates 8 clusters.

K-means or DBSCAN clustering implementation

Another method that can help us with unsupervised clustering is k-means clustering. We would assign each observation to the cluster with the closest mean and recompute the means for the observations assigned to each cluster, repeating the steps until they converge. First, to decide how many groups we should use in the algorithm K means. I use the elbow method to find a compromise between inertia and number of groups. We may find that 7 or 8 groups might be appropriate. Then we apply the K-Means algorithm with 8 groups and plot the scatterplot to visualize the results.

Interpretation of results, discussion and final course curriculum

- Hierarchical clustering, I would like to design 8 courses based on the results of hierarchical clustering. The skills "Presentation" and "Analytical" are bundled in one group. These are soft skills and can relate to verbal and social skills. This allows us to design a course that can teach speaking and presentation skills. 'SAS', 'Excel', 'Tableau' are tools that can analyze the data and present the statistical result. So, they can be taught in the Statistical Analysis Tool course. 'Scala', 'Hadoop', 'Matlab', 'javascript', 'Django' can be used for algorithm development and refer to open-source Big Data Framework. Besides, 'Mongodb', ‘Spark’,’Java’,’snowflake’, ‘api’, ‘nosql’, ‘scala’, 'SPSS', ‘excel’, ‘tableau’, 'pandas', these are tools that can help us analyze and manage data and can be taught in Machine Learning, which can also cover cloud and deep learning. SQL, Python are two very necessary skills at work, we can have two separate SQL and Python courses. "Communication" and "leadership" are classified together and can be taught in leadership training. “Communication skills” are also two very important soft skills can be taught separately in communication training course and leadership raining course as those skills mandatory required in the job description, which should form the basis of the position, so that they can be taught in a specific course.

- K-means clustering The K means that the grouping doesn't give us very clear grouping results since we set the grouping number to 8. There are many abilities in a group, so we cannot get much information from the group result. We can design the courses based on both the clustering result table and the scatterplot. Communication skills are very different from other skills and are often mentioned in job descriptions, so there should be a course to improve communication skills. For the abilities grouped in Group 1. We can combine the results of the hierarchical grouping and the scatterplot of the K-means grouping to make the decision. Python, SQL can be taught in the statistics course as these tools can be used for statistical visualization. SPSS, Teradata, Pandas can be taught in data analysis course because these data analysis tools are simple and common. We can design a computer science course for people to take Programming design and algorithm development can teach Java and Matlab. "Leadership", "Organization" are close together on the scatterplot, so we can open a course called "Leadership Training." Python and SQL are the most needed skills, as mentioned earlier. We can design a Python course and a SQL course separately. Finally, three soft skills: analytics, presentation and passion are close to each other in the scatterplot, we can have a course of speaking and presentation skills as mentioned in Hierarchical clustering.