In part 1, I choose the position of Data Analyst and location in Vancouver, BC. After website scraping, there are 1289 positions in the data frame.

In part 2, I build a list of hard skills and soft skills to explore data analysis and do feature engineering. Then I create a bag of words from the columns “Descriptions” in the data frame of jobs. I select all the columns in my list of skills to create a new data frame called “dataframe\_new”, then combine “dataframe\_new” and term frequency data frame as “data”.

背景图案

描述已自动生成

The data frame with hard skills is from the 47th column in “data”, the data frame with soft skills is from the 5th to the 46th column, and the data frame with all skills is from the 5th column in “data”. Then I use the method wordcloud to show the importance of skills, which is explained by the frequency of a word in the data frame “tf\_new”. Plots of word could of hard skills, soft skills and all skills are shown below:

文本

描述已自动生成文本

描述已自动生成

文本

描述已自动生成

In part 3, I try to implement a hierarchical clustering algorithm. Based on the codes provided in tutorials, I show the plot of dendrograms and the vertical line = 3000. After locating the vertical line, I count how many times horizontal lines are crossed by it :10 times. So, 10 seems a good indication of the number of clusters that have the most distance between them, which is between 8 to 12 courses.

In part 4, I implement k-means clustering algorithm. I use elbow method to determine the optimal number of clusters. X axis stands for how many clusters we choose. Y axis stands for sum of squared distances. Based on the plot, as the number of clusters is 11, the sum squared distances is around 19000. This is small enough. Plot is shown below:

图表, 折线图

描述已自动生成

Thus, I decide to choose 11 clusters in part 4. Next, I add a new column of cluster number to the end of the data frame “data\_all”. Every row corresponding to each job is distributed to different clusters.

电脑屏幕的照片

中度可信度描述已自动生成

And the clustering results are visualized as below:

图表, 散点图

描述已自动生成

In part 5, I design courses’ names and topics for methods in part 3 and part 4. There are 10 courses in part 3 and 11 courses in part 4.图片包含 文本

描述已自动生成

文本

描述已自动生成

I use word cloud to visualize course with its name and topics in part 3 and part 4. The word cloud plots of course 1 are shown below:

图形用户界面, 文本, 应用程序

描述已自动生成图形用户界面, 文本, 应用程序

描述已自动生成

According to the dendrograms in part 3, I found many courses contain one topic only. And some courses contain lots of unrelated topics. I redesign the curriculum by myself, distribute 3 to 4 related topics to each course. Based on KNN, there are 11 clusters, which are designed by model directly. This method is more accurate and smarter than that in part 3. I found most topics are related with each other in each course in part 4, but there are some similar topics between courses. Overall, I choose outputs from part 4 as my final curriculum design.