Report

Firstly, discuss the whole configuration of my program.

The environment of my program is using Python3.7 and using Spyder IDE and the operation system is Windows 8

In the program, we have the following function:

main() function:

It is defined for running the program. In this function, firstly, it finds the directory where the 'kmeans.py' is located by using the os.getcwd() method. And it then calls the buildIndex() to build index. And then it gives the three k values to test the function.

buildIndex() function:

In this function, it reads documents from time, tokenize and build the index with tokens. And we need to put 'time' file in the same directory as the 'kmeans.py'. And then it will generate tis document vectors.

clustering() function:

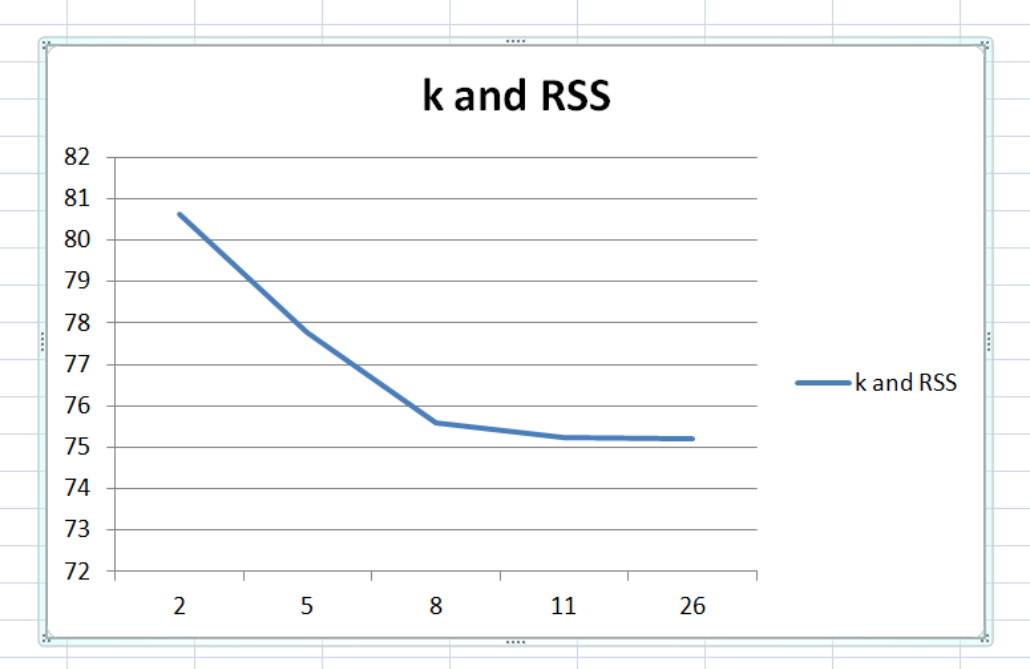
In this function, it implements kmeans clustering algorithm. And then call printTo() function to output the result in “Output.txt”, which is located in the same directory as the “kmean.py” located.

printTo () function

In this function, it output the below: for each cluster, its RSS values and the document ID of the document closest to its centroid, the average RSS value and time taken for computation.

Secondly discuss the experimental study:

Below are pictures about the relation between RSS value and k.



The way to choose best k is to choose from the picture, where the change tendency is becoming flat. As seeing in above picture, we will choose eight as the best k. The procedure for selecting the initial set of centroids is that I randomly choose different document ID from the document ID, as an example, if the k value is three, I randomly choose three numbers as the k values from the document ID. The stopping condition in my implementation is that I circulation the function as the certain times, after that times, the result will be converge.