Jiani Gao
206844103

Clustering Process:
Step 1: Build up the index of all the terms in the collection. (Used when calculating Td-idf)
Step 2: Calculate the Td-idf vectors for each document.
Step 3: Randomly choose k centroids and assign each document to a cluster
Step 4: Recalculate the centroids
Step 5: Calculate the RSS to determine whether continue the loop or break
Step 6: Repeat from step 3 for several times and at last choose the best result with smallest RSS

Q1. What is your r value (the number of random restarts)?
$r = \log n$ where n is the number of document in the collection.

Q2. Why do you select this number (r value)?
I think the r value should be decided according to the number of documents in the collection. For collections that have a large number of documents, the r value should be larger. Thus, I used $\log n$ as the value of r where n is the number of document in the collection.

Q3. What are your stopping criteria to terminate k-means clustering? Why did you select it?
I set up two stopping criteria to terminate k-means clustering. Firstly, I set a maximum clustering round value to prevent costing too much time, here the value is 10. Then, within each round, I checked whether the current RSS is equal to last round's RSS. If true, then the clustering will break in advance. I did this because once the RSS converged to stable, there is no need to continue the loop.