briefly explain how you implemented k-meansclustering in a document (txt / doc / pdf) with addressing following questions in the document: What is your r  
value (the number of random restarts)? Why do you select this number (r value)? What are your stopping criteria  
to terminate k-means clustering? Why did you select it?

After calculating tf-idf vectors on each document’s content, I randomly choose k vectors as initial centers.

For each document, I calculate the Euclidean distance between this document and each center, then assign the document to the cluster with nearest distance. Next, I recalculate each center by meaning all the document’s vectors which belong to this cluster. Repeat above two steps till the center do not change. I use abs(newCenters-centers).sum() as stopping criteria of k-means clustering. The reason is when this criteria is very small, all the centers and each document’s cluster do not change obviously anymore.

My r value is 30. In each restart, I calculate RSS and record the current best RSS. I select this number because when I restart 30 times, the current best RSS do not decrease obviously anymore.