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 restart number 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 best RSS can usually decrease to the point I get when I restart 100 times.

```
test 10 documents
current best RSS:  1344.06700944
current best RSS:  1344.06700944
current best RSS:  935.11019544
current best RSS:  935.11019544
current best RSS:  935.11019544
current best RSS:  935.11019544
current best RSS:  935.11019544
current best RSS:  895.142329542
current best RSS:  895.142329542
current best RSS:  895.142329542
current best RSS:  895.142329542
current best RSS:  895.142329542
current best RSS:  895.142329542
current best RSS:  895.142329542
current best RSS:  895.142329542
current best RSS:  895.142329542
current best RSS:  895.142329542
current best RSS:  895.142329542
current best RSS:  895.142329542
current best RSS:  895.142329542
current best RSS:  521.861459196
current best RSS:  521.861459196
current best RSS:  521.861459196
current best RSS:  521.861459196
current best RSS:  521.861459196
current best RSS:  521.861459196
current best RSS:  521.861459196
current best RSS:  521.861459196
current best RSS:  521.861459196
```

```
test 50 documents
current best RSS:  12381.5845562
current best RSS:  12074.7821413
current best RSS:  11866.7964841
current best RSS:  11866.7964841
current best RSS:  11866.7964841
current best RSS:  11866.7964841
current best RSS:  11475.0328702
current best RSS:  11475.0328702
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
current best RSS:  10976.4757267
```