CS 429 - Information Retrieval

Bonus Assignment - k-means clustering
Patrick Connolly

Part A:

Average RSS (k = 2): 1.0116225841473434
Clustering Time: 15.97928476333618164062 seconds

Average RSS (k = 3): 0.9787325044616905
Clustering Time: 28.40104937553405761719 seconds

Average RSS (k = 4): 0.9670643569278109
Clustering Time: 47.04922628402709960938 seconds

K random documents are chosen as the initial centroids. The cosine between the centroid and central document is checked and the distance is averaged. I find the nearest document and treat it as the new centroid for the cluster.

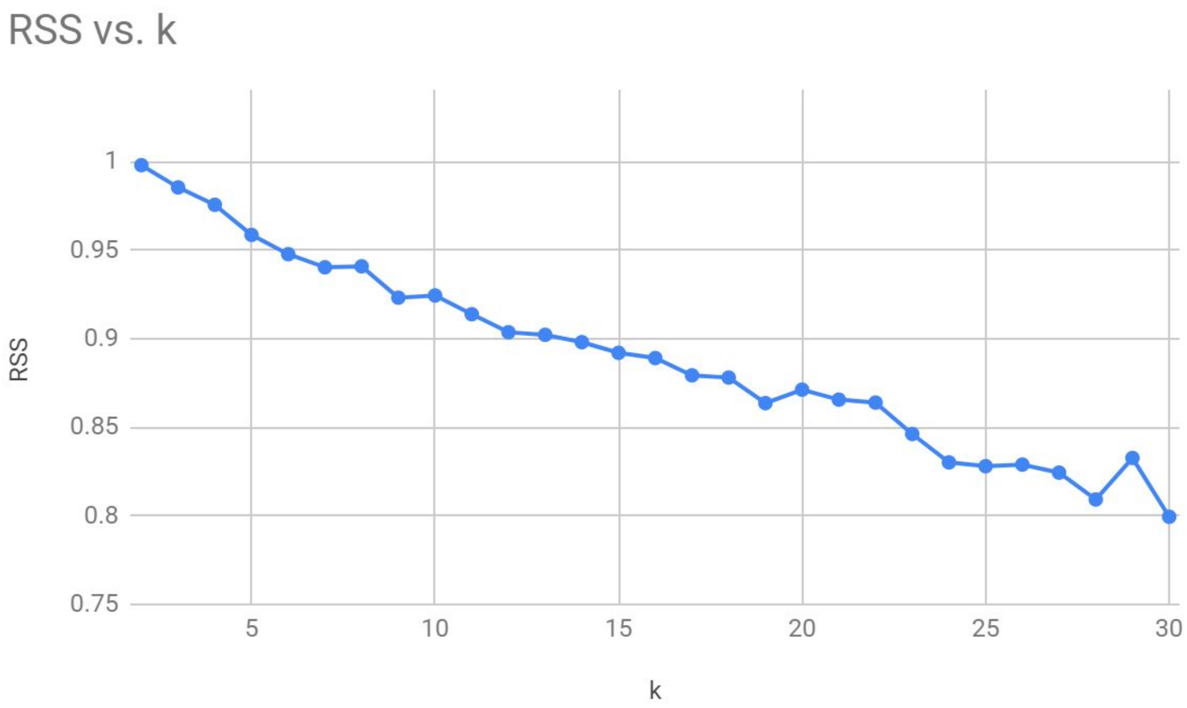
Part B:

Running tests k=2 to k=30

k	RSS
2	0.9979213067232229
3	0.9853638066169012
4	0.9755442483226051
5	0.9585556115961383

6	0.9476834528130137
7	0.9402683332974016
8	0.9408203464638032
9	0.9231139542206026
10	0.9243530578607349
11	0.9137735196834811
12	0.9036194109719726
13	0.9021544752045362
14	0.8980042958709324
15	0.8919853759260665
16	0.8890333567777093
17	0.8792880421481482
18	0.8780220686582076
19	0.8636052638890346
20	0.8711912957851788
21	0.8655699176354519
22	0.8638972775320644
23	0.846172956165254
24	0.8301300938730127
25	0.827988763482362
26	0.828909341917987
27	0.8243667347615921
28	0.8092121504728056

29	0.8326294342012179
30	0.7994430908239405



Good trade offs: 9,19,28
Bad trade offs: 8,20,29