

CS 429 – Information Retrieval

Assignment 5 – K-Means Clustering

Mayank Bansal – mbansal5@hawk.iit.edu

April 28, 2018

Part-A: K-Means

| K | AVERAGE RSS | TIME TO COMPUTE |
|---|--------------------|---------------------------------|
| 2 | 1.0100071355169122 | 15.95866298675537109375 seconds |
| 3 | 0.9904275632312971 | 24.31481981277465820312 seconds |
| 4 | 0.9826224823459061 | 31.65370512008666992188 seconds |
| 5 | 0.9616689566184693 | 40.58723115921020507812 seconds |
| 6 | 0.9486874078624905 | 50.30588006973266601562 seconds |

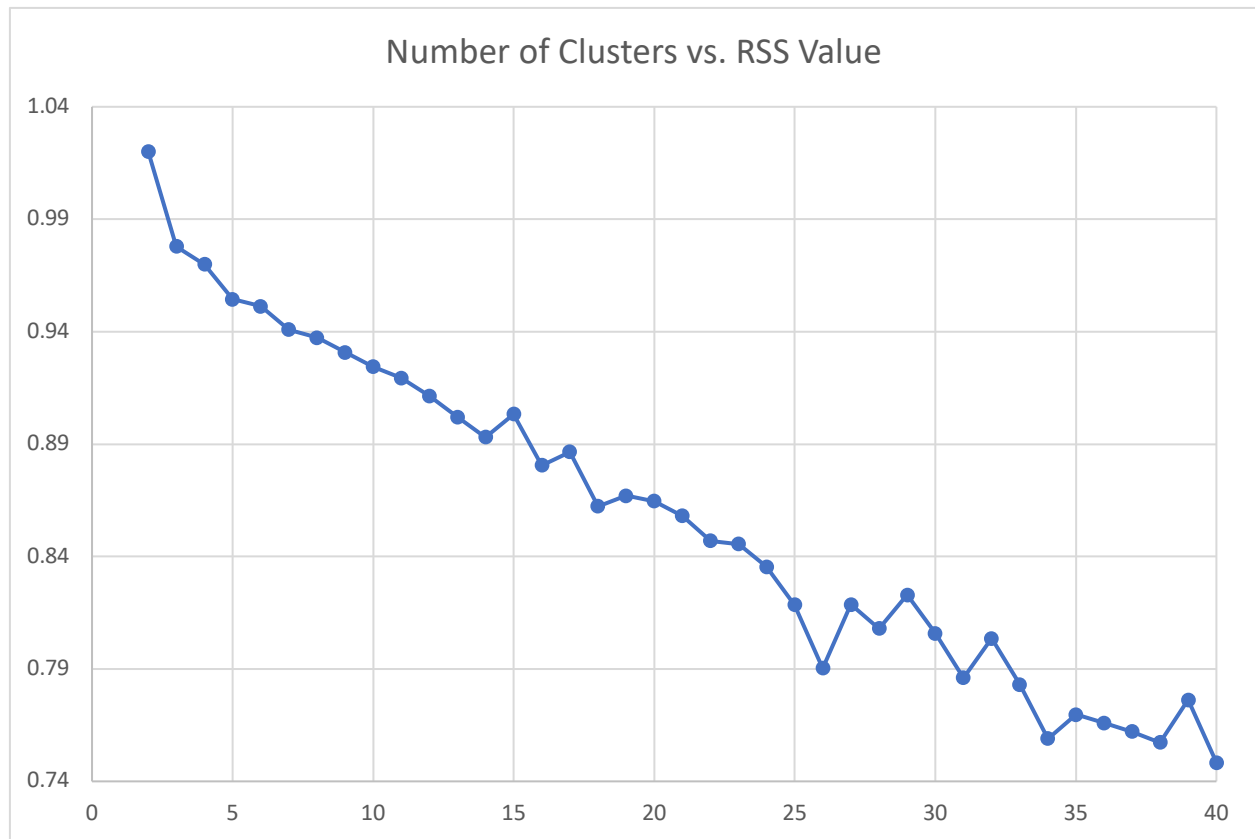
In this implementation, K random documents are chosen as the initial centroids. Once the clusters are formed, I check the cosine distance between the centroids and the cluster documents. After averaging the distances and finding the new centroid, I find the nearest document and treat that as the new centroid for that cluster.

Part-B: Experimental Study

Tests were run on a **32 vCPU server with 64 GB of RAM** in **6 mins** as a Mac/Windows PC will not give the performance shown below. Running tests from K=2 to K=40 on a Mac/Windows PC will take **over 2 hours 20 mins**

| | | |
|---|--------------------------------|---------|
| Avg RSS (k = 2) : 1.0200607554083274 | time: 12.81563138961791992188 | seconds |
| Avg RSS (k = 3) : 0.9779563540958844 | time: 19.12983679771423339844 | seconds |
| Avg RSS (k = 4) : 0.9699843740657680 | time: 25.72898602485656738281 | seconds |
| Avg RSS (k = 5) : 0.9544414994404326 | time: 32.23838496208190917969 | seconds |
| Avg RSS (k = 6) : 0.9512855805429591 | time: 38.17191267013549804688 | seconds |
| Avg RSS (k = 7) : 0.9409261006542091 | time: 44.77525377273559570312 | seconds |
| Avg RSS (k = 8) : 0.9372884142463301 | time: 50.75576925277709960938 | seconds |
| Avg RSS (k = 9) : 0.9307217144870830 | time: 57.62284207344055175781 | seconds |
| Avg RSS (k = 10) : 0.9243751824855324 | time: 63.56317996978759765625 | seconds |
| Avg RSS (k = 11) : 0.9194594186254835 | time: 69.71414780616760253906 | seconds |
| Avg RSS (k = 12) : 0.9114641669576676 | time: 76.24836874008178710938 | seconds |
| Avg RSS (k = 13) : 0.9020011868041419 | time: 82.74934911727905273438 | seconds |
| Avg RSS (k = 14) : 0.8930966132082694 | time: 88.32648491859436035156 | seconds |
| Avg RSS (k = 15) : 0.9032904387024533 | time: 96.15440297126770019531 | seconds |
| Avg RSS (k = 16) : 0.8805962337191896 | time: 102.56352138519287109375 | seconds |
| Avg RSS (k = 17) : 0.8864616830938162 | time: 109.05680775642395019531 | seconds |
| Avg RSS (k = 18) : 0.8623502167078202 | time: 116.27125096321105957031 | seconds |
| Avg RSS (k = 19) : 0.8670238585932841 | time: 121.85592818260192871094 | seconds |
| Avg RSS (k = 20) : 0.8646709459631780 | time: 129.06716895103454589844 | seconds |
| Avg RSS (k = 21) : 0.8580690005702488 | time: 134.22805905342102050781 | seconds |
| Avg RSS (k = 22) : 0.8470150554800705 | time: 139.99273705482482910156 | seconds |
| Avg RSS (k = 23) : 0.8455184359575032 | time: 150.04510664939880371094 | seconds |
| Avg RSS (k = 24) : 0.8354803169267598 | time: 167.51417493820190429688 | seconds |
| Avg RSS (k = 25) : 0.8184765785846068 | time: 177.02483272552490234375 | seconds |
| Avg RSS (k = 26) : 0.7903092471160692 | time: 180.69130945205688476562 | seconds |
| Avg RSS (k = 27) : 0.8186588634523014 | time: 177.15329265594482421875 | seconds |
| Avg RSS (k = 28) : 0.8079321450642887 | time: 185.85685586929321289062 | seconds |
| Avg RSS (k = 29) : 0.8229353616016932 | time: 191.62244486808776855469 | seconds |
| Avg RSS (k = 30) : 0.8148230377313116 | time: 201.45267987251281738281 | seconds |
| Avg RSS (k = 31) : 0.7861139398917355 | time: 221.18924593925476074219 | seconds |
| Avg RSS (k = 32) : 0.8035579281215947 | time: 224.88113999366760253906 | seconds |
| Avg RSS (k = 33) : 0.7831191738434823 | time: 219.38294792175292968750 | seconds |
| Avg RSS (k = 34) : 0.7590528786322909 | time: 337.11087894439697265625 | seconds |
| Avg RSS (k = 35) : 0.7696204810251044 | time: 233.29223847389221191406 | seconds |
| Avg RSS (k = 36) : 0.7660116173435157 | time: 241.62176656723022460938 | seconds |
| Avg RSS (k = 37) : 0.7621157598239955 | time: 246.45414233207702636719 | seconds |
| Avg RSS (k = 38) : 0.7572980679027217 | time: 252.17027068138122558594 | seconds |
| Avg RSS (k = 39) : 0.7762306950929744 | time: 261.63117957115173339844 | seconds |
| Avg RSS (k = 40) : 0.7484044935460231 | time: 265.24077558517456054688 | seconds |

| K | RSS VALUE |
|----|-------------|
| 2 | 1.020060755 |
| 3 | 0.977956354 |
| 4 | 0.969984374 |
| 5 | 0.954441499 |
| 6 | 0.951285581 |
| 7 | 0.940926101 |
| 8 | 0.937288414 |
| 9 | 0.930721714 |
| 10 | 0.924375182 |
| 11 | 0.919459419 |
| 12 | 0.911464167 |
| 13 | 0.902001187 |
| 14 | 0.893096613 |
| 15 | 0.903290439 |
| 16 | 0.880596234 |
| 17 | 0.886461683 |
| 18 | 0.862350217 |
| 19 | 0.867023859 |
| 20 | 0.864670946 |
| 21 | 0.858069001 |
| 22 | 0.847015055 |
| 23 | 0.845518436 |
| 24 | 0.835480317 |
| 25 | 0.818476579 |
| 26 | 0.790309247 |
| 27 | 0.818658863 |
| 28 | 0.807932145 |
| 29 | 0.822935362 |
| 30 | 0.805787491 |
| 31 | 0.78611394 |
| 32 | 0.803557928 |
| 33 | 0.783119174 |
| 34 | 0.759052879 |
| 35 | 0.769620481 |
| 36 | 0.766011617 |
| 37 | 0.76211576 |
| 38 | 0.757298068 |
| 39 | 0.776230695 |
| 40 | 0.748404494 |



Initial Centroids: Random
Stop Condition: 5 iterations

From the plot, 17, 26, 34 seem to give a good tradeoff for k vs. RSS value