# CS 429 – Information Retrieval

## Assignment 5 – K-Means Clustering

Mayank Bansal – mbansal5@hawk.iit.edu

April 28, 2018

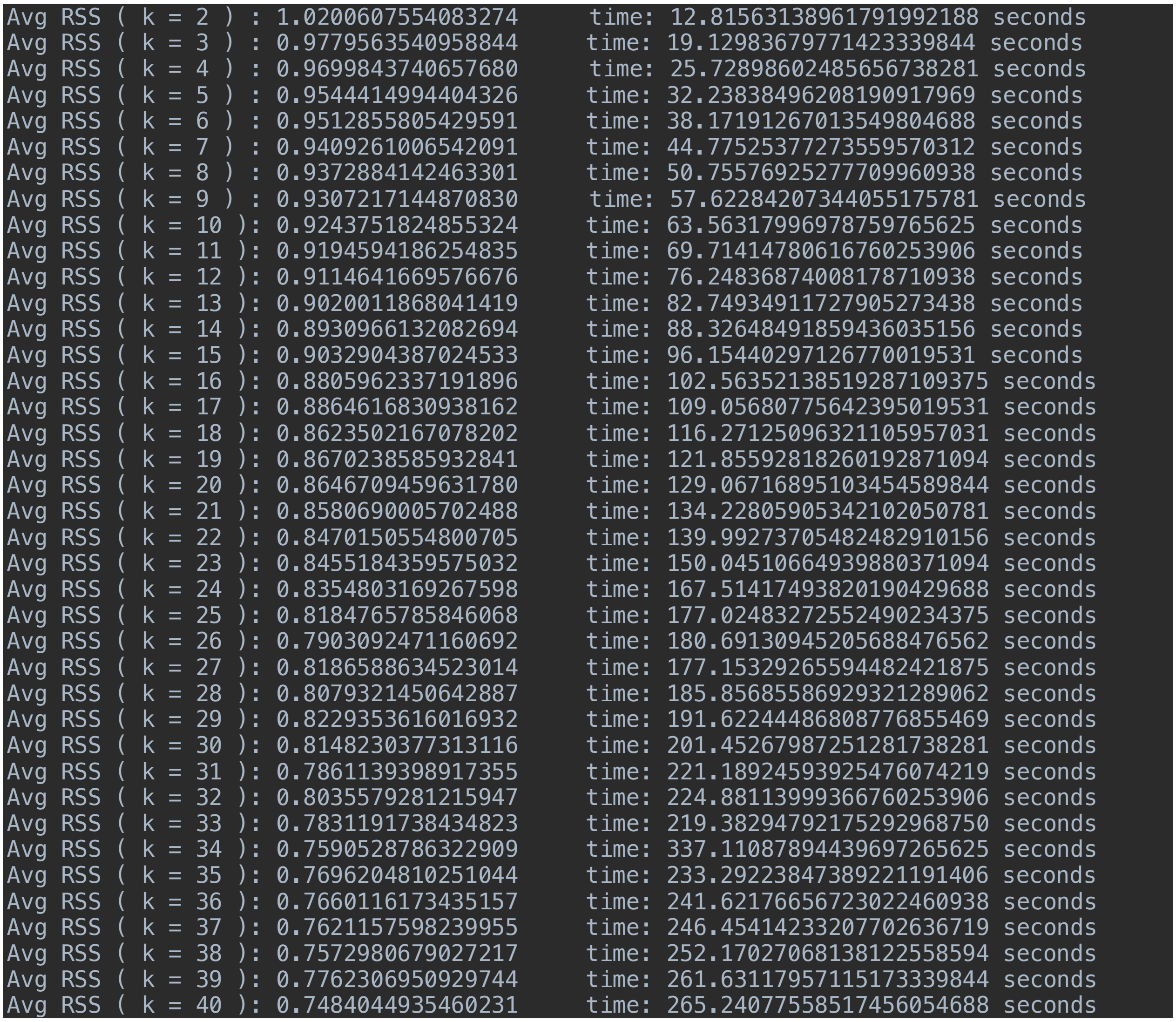
# Part-A: K-Means

| 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.

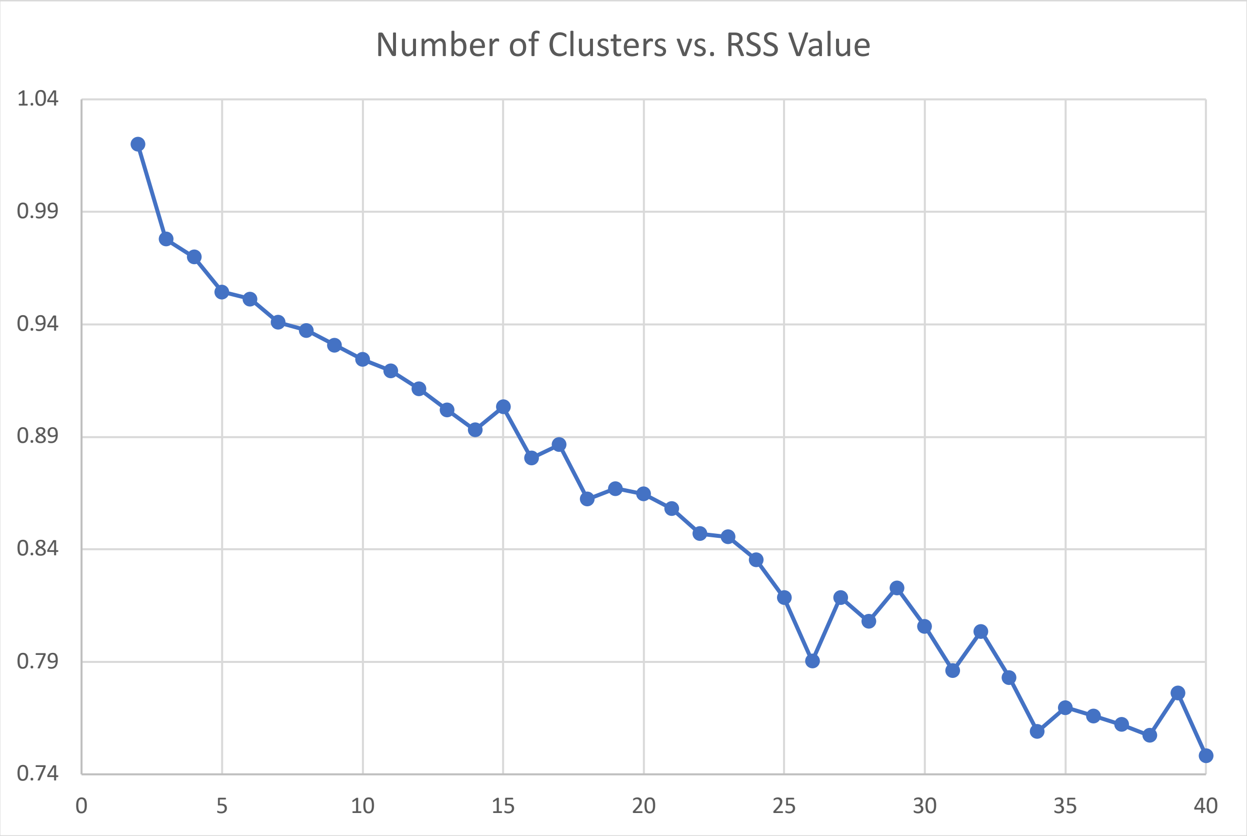
# Part-B: Experimental Study

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
```

| 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