K MEANS CLUSTERING

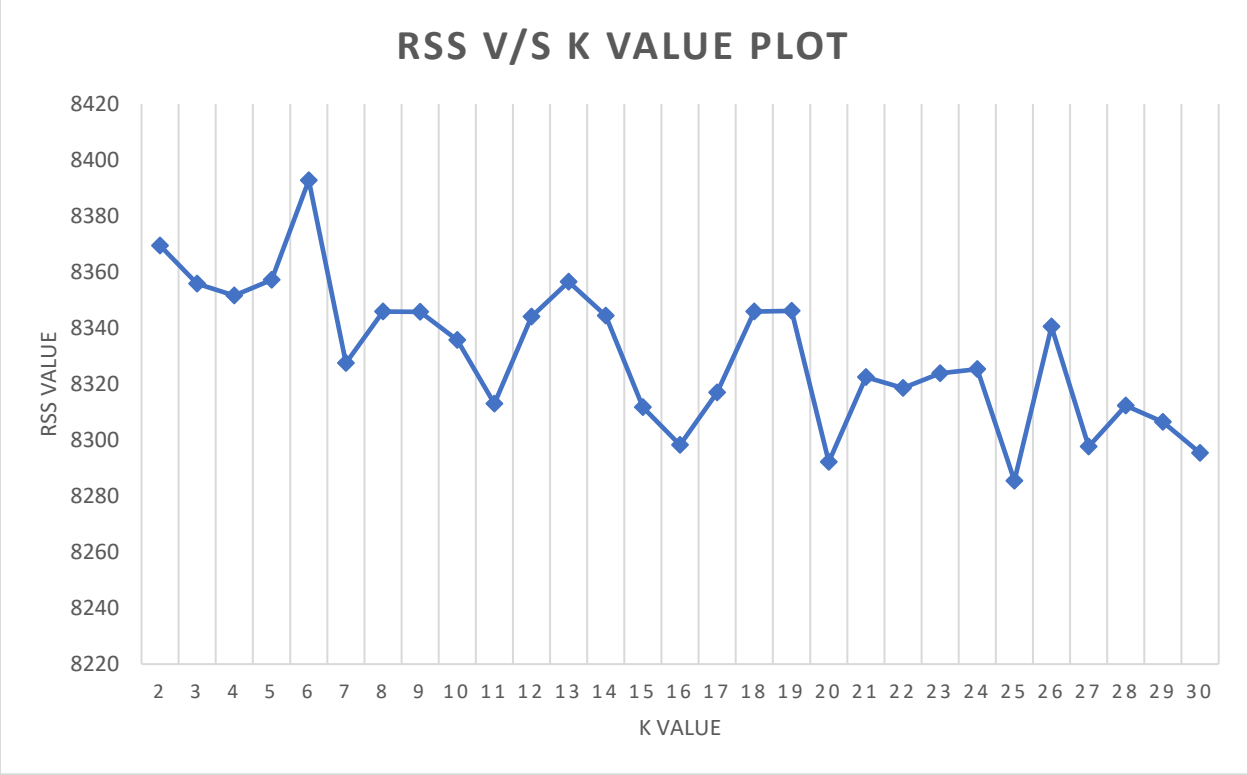
APRIL 30, 2018

ARKADITYA VERMA
A20414100

K means Clustering

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided.

The experiment is conducted of 350 set of documents from the collection set with k value ranging from 2 to 30. The initial centroids are selected randomly from the set of documents provided. The clustering iteration is ran 5 times before the cluster is computed at each step.



```
Total Clustering time :  1892.586 seconds
kvalue  RSS
========================
2       [8369.442132470842, 6.189]
3       [8355.84602814336, 7.12]
4       [8351.538824594298, 8.029]
5       [8357.122569035424, 8.504]
6       [8392.739062537217, 9.107]
7       [8327.447010877348, 9.857]
8       [8345.888860268236, 10.453]
9       [8345.748846046085, 10.859]
10      [8335.71347984255, 12.356]
11      [8312.922718056769, 12.511]
12      [8344.047462886107, 13.728]
13      [8356.449205641791, 14.307]
14      [8344.429901676083, 13.894]
15      [8311.722881225885, 124.702]
16      [8298.176209345762, 32.433]
17      [8316.960946027933, 17.0]
18      [8345.890576455477, 16.832]
19      [8346.113810085815, 17.603]
20      [8292.173640395607, 18.775]
21      [8322.477457817022, 19.253]
22      [8318.522543786174, 20.889]
23      [8323.784405838076, 21.727]
24      [8325.272776783511, 22.197]
25      [8285.48165251979, 1328.72]
26      [8340.460918975772, 21.198]
27      [8297.696866262617, 21.863]
28      [8312.31573741833, 22.951]
29      [8306.411021911359, 24.398]
30      [8295.408275465223, 25.036]
```

The above plot says that we need to normalize and smoothen it more using various algorithms for better results. As the plot says, we get the best results with k = 25