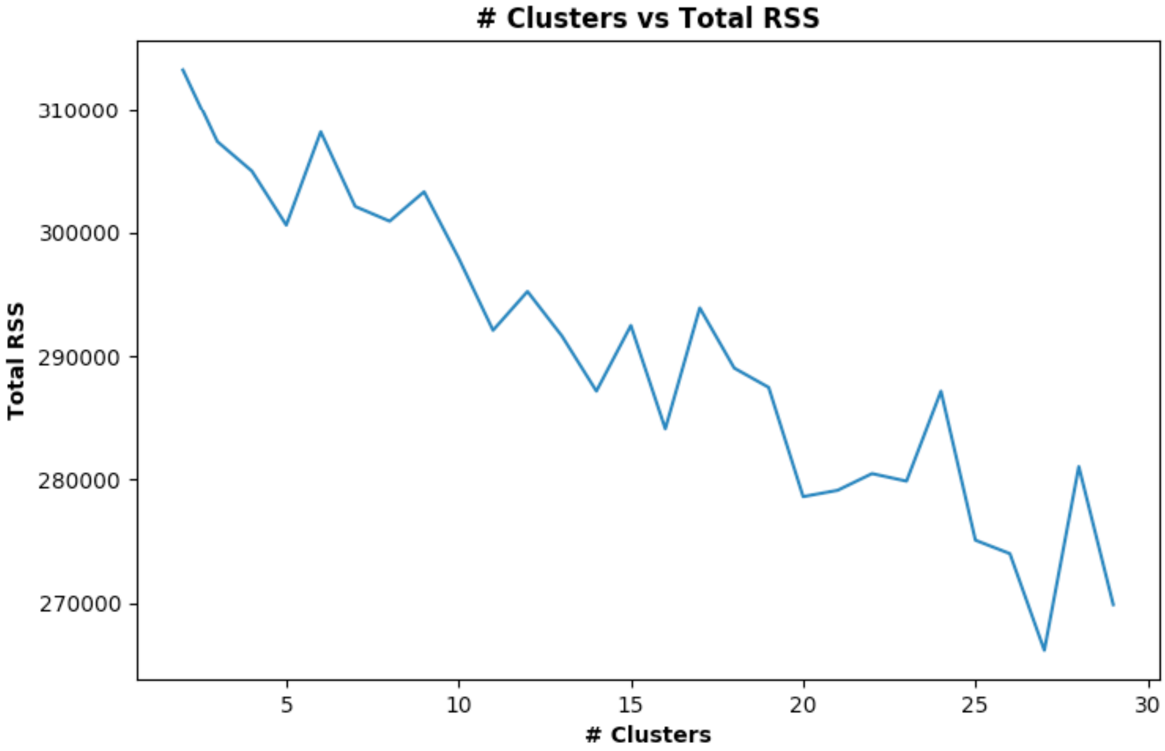
K-Means Clustering Experimental Study



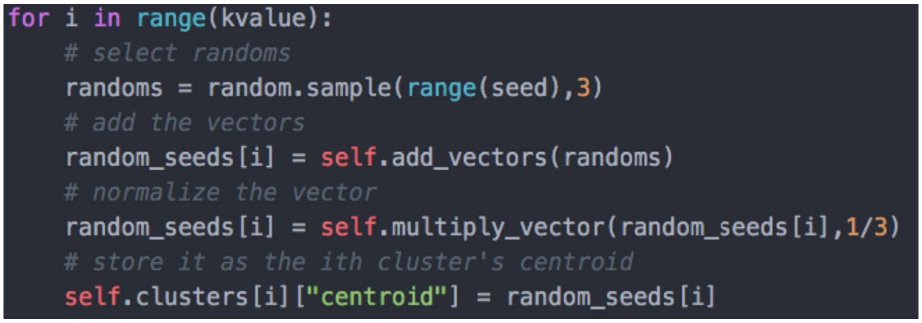
Optimal Number of Clusters:

The optimal number of clusters are the areas in the graph where the total sum RSS is relatively flat. From the above plot, the optimal number of clusters are **6**, and **20-23**

Procedure for Setting Initial Centroids

For each cluster, we select a set of 3 random documents and take the average of them and use that as the centroid. For the random sampling, we do not replace documents after they are already used as a centroid to ensure that we cover the entire document space effectively.

This code snippet below shows how the centroids are computed are stored



```
for i in range(kvalue):
    # select randoms
    randoms = random.sample(range(seed),3)
    # add the vectors
    random_seeds[i] = self.add_vectors(randoms)
    # normalize the vector
    random_seeds[i] = self.multiply_vector(random_seeds[i],1/3)
    # store it as the ith cluster's centroid
    self.clusters[i]["centroid"] = random_seeds[i]
```

Stopping Condition for Clustering

The clusters have converged when the change in RSS values is less than 1000. The change in RSS is computed as absolute_value(old_rss - new_rss) where new_rss is the total RSS after the centroids are recomputed and the documents are clustered around them.

From the plot, what is the value of 'k' that provides a good tradeoff with change in RSS?

The good K values are 6, 20, 21, 22, and 23