Read Me Document

1) The versions of python , spark and scala are as below
   Python Version : 2.7 ,
   Spark version : 2.3.1 ,
   Scala : 2.11

2) K means Algorithm has been implemented in this assignment.
3) For task 1 - First initial centroids are generated and then the Euclidian distance of all documents with this initial centroid is calculated and the one with minimum distance is allotted to the corresponding cluster.
4) After the new cluster is formed new centroids are calculated for each cluster and the process is repeated until clusters in consecutive iterations are same or until maximum number of iterations is reached.
5) For finding the maximum number of words in each cluster, word count for all documents is done and the top 10 frequent words are returned in a list.
6) For task 2 – mllib is used and clusters are generated.
7) For finding the maximum number of words in each cluster, word count for all documents is done and the top 10 frequent words are returned in a list.
8) The data structure used for TF and TFIDF is sparse vector for both implementation.
9) The command to execute task 1 is as below

   spark-submit parth_vaghani_Kmeans.py "/input/file/path/yelp_reviews_small.txt" <feature(W/T)> <N> <Max iterations>

   example:

```
spark-submit parth_vaghani_Kmeans.py
"/Users/parth/Desktop/USC/Data
Mining/Assignment4/Data/yelp_reviews_clustering_small.txt" W 5 20

spark-submit parth_vaghani_Kmeans.py
"/Users/parth/Desktop/USC/Data
Mining/Assignment4/Data/yelp_reviews_clustering_small.txt" T 5 20
```

10) The command to execute task 2 is as below

    spark-submit parth_vaghani_Cluster.py "/input/file/path/yelp_reviews_small.txt" <algorithm(K/B)> <N> <Max iterations>

    example:

```
spark-submit parth_vaghani_Cluster.py
"/Users/parth/Desktop/USC/Data
Mining/Assignment4/Data/yelp_reviews_clustering_small.txt" K 8 20

spark-submit parth_vaghani_Cluster.py
"/Users/parth/Desktop/USC/Data
Mining/Assignment4/Data/yelp_reviews_clustering_small.txt" B 8 20
```