Spark Version: 2.3.1
Scala Version: 2.11.0

Run the code w/ same instructions as described in the assignment document:

KMeans was implemented for task 1. K Means and Bisecting K Means functionality from ml library was used for task 2.

Explanation of implementation K Means:
Began by loading text-file into RDD[Array[String]] by splitting reviews into words.
Created a dict mapping vocab from the entire corpus to number of documents containing.
Created 2 dicts to encode and decode words into positions in an array. Created RDD[(cluster #, Array) from each data point and initially assigned every data point to cluster 0. The array contained either the tf-idf values or word count values for every word in the corpus. **Choose n random centroids by selecting n random points from data points.** Repeatedly assigned cluster and then grouped points and took avg (For n iterations).

Task 1

```
$SPARK_HOME/bin/spark-submit --class Task1 Firstname_Lastname_KMeans.jar
<input_file> <feature> <N> <Iteration>
```

Task 2

```
$SPARK_HOME/bin/spark-submit --class Task2 Firstname_Lastname_Cluster.jar
<input_file> <algorithm> <N> <Iteration>
```

Note: As per recent change in requirements, GMM and compatibility for large datasets has not been implemented.