Bashar Alhafni
USC ID: 9065856238

INF-553
Fall 2018 Assignment 4

Versions used for this assignment:

Spark: 2.3.1
Scala: 2.11.0

NOTE: both of task1 and task2 in a single jar file.

Task1:
To execute the code for task1 using word count as features, the following command must be executed:

```
spark-submit --class Task1 Bashar_Alhafni_Clustering.jar
/Users/alhafni/Desktop/repos/inf553_data_mining/hw4/Data/yelp_reviews_
clustering_small.txt W 5 20
```

To execute the code for task1 using tf-idf count as features, the following command must be executed:

```
spark-submit --class Task1 Bashar_Alhafni_Clustering.jar
/Users/alhafni/Desktop/repos/inf553_data_mining/hw4/Data/yelp_reviews_
clustering_small.txt T 5 20
```

Task2:
To execute the code for task2 using the Kmeans algorithm, the following command must be executed:

```
spark-submit --class Task2 Bashar_Alhafni_Clustering.jar
/Users/alhafni/Desktop/repos/inf553_data_mining/hw4/Data/yelp_reviews_
clustering_small.txt K 8 20
```

To execute the code for task2 using the Bisecting-Kmeans algorithm, the following command must be executed:

I had an OutofMemory error so I had to increase the size of the memory used.

```
spark-submit --driver-memory 4g --class Task2
Bashar_Alhafni_Clustering.jar
/Users/alhafni/Desktop/repos/inf553_data_mining/hw4/Data/yelp_reviews_
clustering_small.txt B 8 20
```

NOTE: I used the **json4s** library to format the outputs of both task1 and task2 as json files.

References:

https://spark.apache.org/docs/2.2.0/mllib-clustering.html

https://spark.apache.org/docs/latest/mllib-feature-extraction.html