# INF 553 FALL 2018 ASSIGNMENT 4

Name: Aditya Chavan
USC ID: 3416409224

**I am Running the Scala Code, using the following IDE:**
intellij IDEA

**As per instructions, the environment was set as follows:**
1) Java – 1.8.0_181
2) sbt – 1.2.3
3) Hadoop – 2.7
4) Scala – 2.11.12
5) Spark – 2.3.1

**Description:**

K-Means Algorithm is a clustering algorithm which clusters the data points into clusters. K is the number of clusters.
Parameters are Cluster Vector, number of clusters, number of iterations, the seed.

Problem 1: K-Means done from scratch. Input_file (Yelp Reviews) is read into an RDD. has two features: Word Count and Term Frequency- Inverse Document Frequency. Functions are made for Euclidean Distance calculation, Finding Centroids, Computing SSE and for K-Means. Clustering reviews data performed.
the output is a json file.

Problem 2: existing spark libraries (mllib libraries) used for K-Means and Bisecting K-Means clustering. Bisecting K-Means Algorithm perform better than K-Means & produces different clustering. kind of Hierarchical Clustering with divisive algorithm. Parameters are number of clusters, maximum number of iterations, seed. Clustering reviews data performed. output is written json file.

## Steps to Run the code using command line

**1.** Change to directory from the folder(Aditya_Chavan_hw4) that contains *Aditya_Chavan_Clustering.jar*

**2. To run .jar file for TASK 1, execute the command:**

$SPARK_HOME/bin/spark-submit --driver-memory 6g --class Task1 Aditya_Chavan_Clustering.jar
<input_filepath> <feature> <num_clusters> <num_iterations>

**Example:**
```
spark-submit --driver-memory 6g --class Task1
/Users/mahima/Desktop/Aditya_Chavan_Clustering.jar
/Users/mahima/Desktop/yelp_reviews_clustering_small.txt W 5 20
```

**Example:**
```
spark-submit --driver-memory 6g --class Task1
/Users/mahima/Desktop/Aditya_Chavan_Clustering.jar
/Users/mahima/Desktop/yelp_reviews_clustering_small.txt T 5 20
```

**3. To run .jar file for TASK 2, execute the command:**

```
$SPARK_HOME/bin/spark-submit --driver-memory 6g --class Task2
Aditya_Chavan_Clustering.jar
<input_filepath> <Algorithm> <num_clusters> <num_iterations>
```

**Example:**
```
spark-submit --driver-memory 6g --class Task2
/Users/mahima/Desktop/Aditya_Chavan_Clustering.jar
/Users/mahima/Desktop/yelp_reviews_clustering_small.txt K 8 20
```

**Example:**
```
spark-submit --driver-memory 6g --class Task2
/Users/mahima/Desktop/Aditya_Chavan_Clustering.jar
/Users/mahima/Desktop/yelp_reviews_clustering_small.txt B 8 20
```