Hw5

Cr statement: Linbin Sun
trainingDataFileName:corresponds to a subset of the data that should be used as the training set for your algorithm

K: the value of k to use when clustering

Clustering option: takes one of the following five values,
1 (use the four original attributes for clustering, which corresponds to Q3.1)
2(apply a log transform to reviewCount and checkins, which corresponds to Q3.2)
3(use the standardized four attributes for clustering, which corresponds to Q3.3)
4(use the four original attributes and Manhattan distance for clustering, which corresponds to Q3.4)
5(use 3% random sample of data for clustering, which corresponds to 3.5)

Your code should read in the training sets from the csv file, cluster the training set using the specified value of k, and output the within-cluster sum of squared error and cluster centroids. For the centroid of each cluster root the values for each of the four attributes in the following order

Attitude, longtitude, reviewCount, checking

The expected output is given below. Note that this yelp.csv, k=4, cluster option 1

```
$ python kmeans.py yelp.csv 4 1
WC-SSE=15.2179
Centroid1=[49.00895,8.39655,12,3]
…
CentroidK=[33.33548605,-11.7714182,9,97]
```

2 Means
2.1 theory: what are the benefits of the k-means clustering algorithm?