Ορφανουδάκης Φίλιππος 03113140

Θέμα 3ο: Machine Learning - Ομαδοποίηση δεδομένων με εκτέλεση του k-means αλγόριθμου

1.

```
tripdata=sc.textFile("hdfs://master:9000/yellow_tripdata_1m.csv").map(lambda line: (float(line.split(',')[3]),
float(line.split(',')[4]))).filter(lambda line: (-75.0<=line[0] and
line[0]<=-73.0) and (40.0<=line[1] and line[1]<=41.0))
```

```
MAP(key,value):
    for each line:
        cell=line.split('',''')
        long=cell[3]
        lat=cell[4]
        if( -75.0≤long≤-73.0   and  40.0 ≤ lat ≤ 41.0)
            emit(long,lat)
```

2.

```
closer = tripdata.map(lambda p: (closerto(p,centroids),(p,1)))
```

```
MAP(key,value):
    calculate closer_index of (p,centroids)
    emit(closer_index,((long,lat),1))
```

3.

```
tmpcloser=closer.reduceByKey(lambda first,sec:
((first[0][0]+sec[0][0],first[0][1]+sec[0][1]),first[1]+sec[1]))
```

```
REDUCE(key,list(values)):
    longit=0
    latit=0
    count=0
    for each i in values:
        longit=longit+long
        latit=latit+lat
```

```
           count=count+1
      emit(index,((longit,latit),count))
```

4.

```
centroids=tmpcloser.mapValues(lambda calc:
(calc[0][0]/calc[1],calc[0][1]/calc[1]))
```

```
MAP(key,value):
      emit(index,(longit/count,latit/count)) //mean values
```

5.

```
centroids1=centroids1.map(lambda c: (c[0]+1,(c[1][0],c[1][1])))
```

```
Map(key,value):
      emit(index+1,(mean_long,mean_lat))
```