K-means for Geo-location Clustering in Spark

Option: Project3

Course: CSE427S Cloud Computing for Big Data Application

Instructor: Marion Neumann

Student:

Start: 11/20/2018

End: 12/14/2018

**Milestone3: Final Report**

1. motivation (5%)

K-mean as one of unsupervised learning algorithms is very powerful in data

visualization and data interpretation. It accomplishes the idea of auto-clustering

or auto-categorizing. We can use k-mean in marketing data, document data and

geographic data. It occupies different meanings in classification, behavior

clustering and location evaluation. In this project, we will apply k-mean

algorithm in SPARK environment where we will extract different sizes of big

datasets and fit them into k-mean algorithm. We will visualize results from

clustering and show our experimental results with our comprehensions.

2. Documentation of Approach

Choices: k-mean, k-median, EM algorithm.

Model: k-mean

Training Parameters: k centroids, distance measurement.

Visualization: matplot

visualization

clusters

K-mean

Data

# of clusters

K’ iterations

Stage1 Stage2 Stage3

(initialization) (training model) (visualization)

We break down our approach into three stages. We name first stage as data

initialization, second stage as training model and last stage as visualization.

Stage1: we choose proper number of k clusters as using random choice where we

restrict the range of randomized picking to achieve proper values of

centroids. A local dataset will be chosen for experimental result for

initial practice.

Stage2: k-mean will be fit by geographic data and will make k’ iterations to form

k clusters

Stage3: we will visualize different clusters with different colors by using matplot

package in python

3. Methodology

3.1 Program Design

Work Flow:

closest\_point(): given a (latitude/longitude) point and an array of current

center points returns the index in the array of the center

closest to the given point

add\_points(): given two points, return a point which is the sum of the

two points.

euclidean\_distance(): calculate the eculidean distance between two

points.

great\_circle\_distance(): the shortest distance between two points on the surface of a sphere, measured along the surface of the sphere

WCSS(): within-clusters sum-of-squares measure

kmeans\_cluster(): aggregating functions above and make k clusters

visualize(): make color for k groups of data points and make label ‘x’ on

k centroids

read\_data(): extract data from given datasets

3.2 Data Description

3.2.1: Small Dataset

// which dataset/application – creativity (5%)

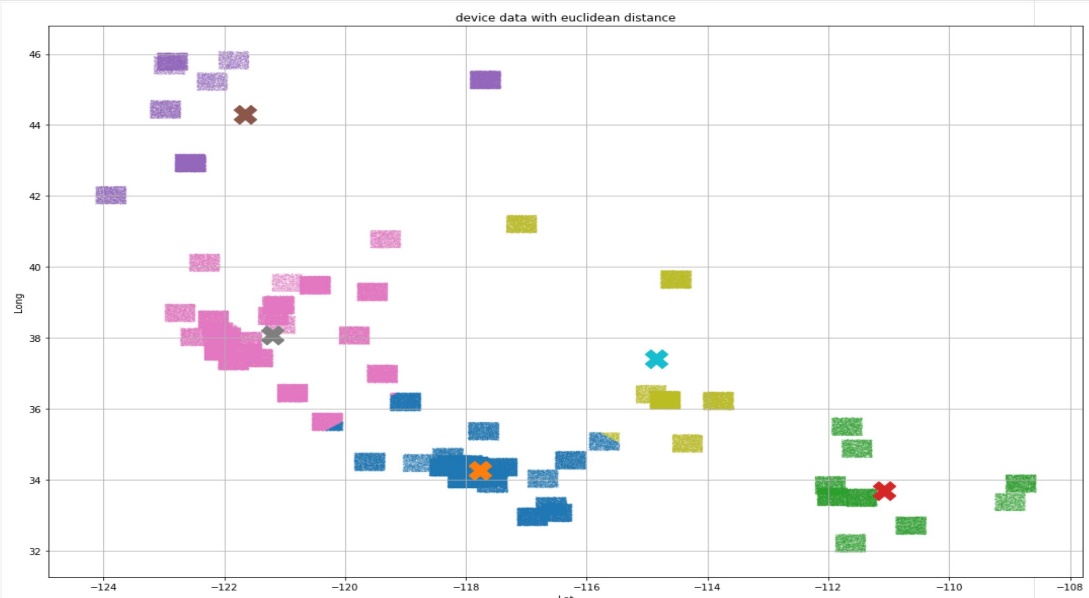
// description (5%)

3.2.2: Big Dataset

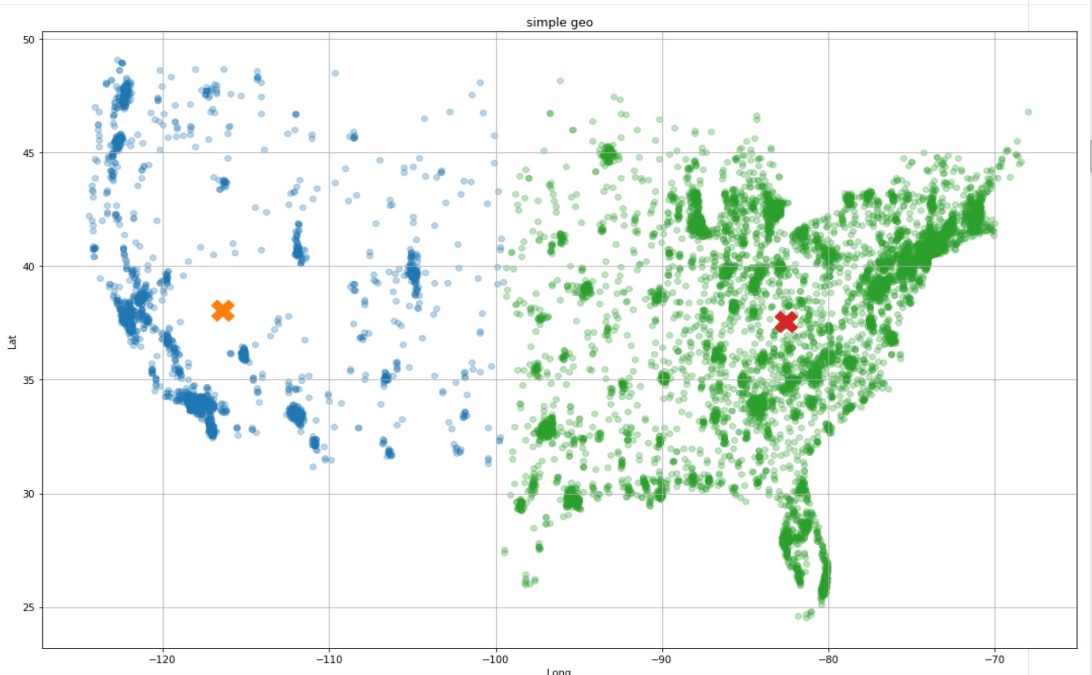
4. Results

4.1 Small Datasets

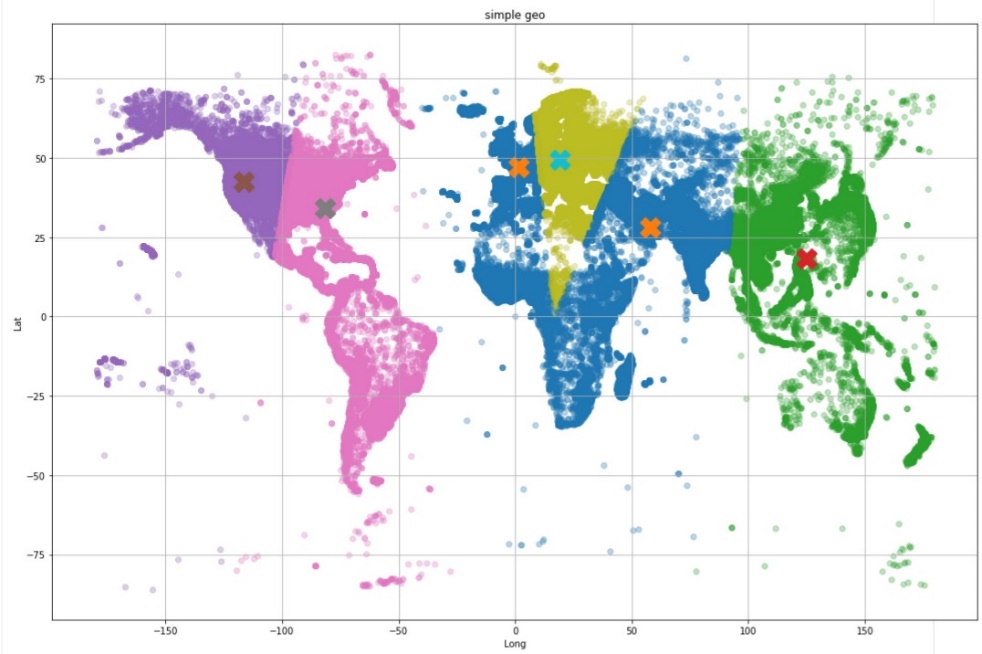
4.1.1 Device Location



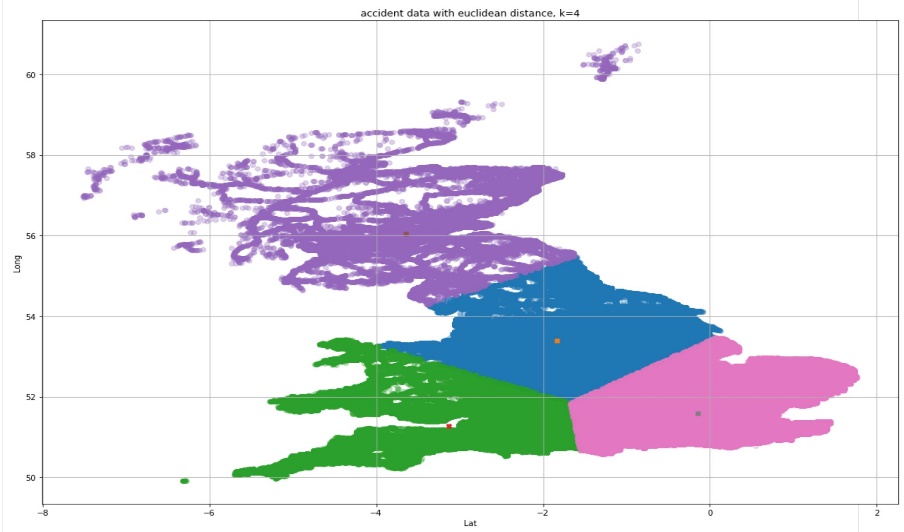
4.1.2 Synthetic Location

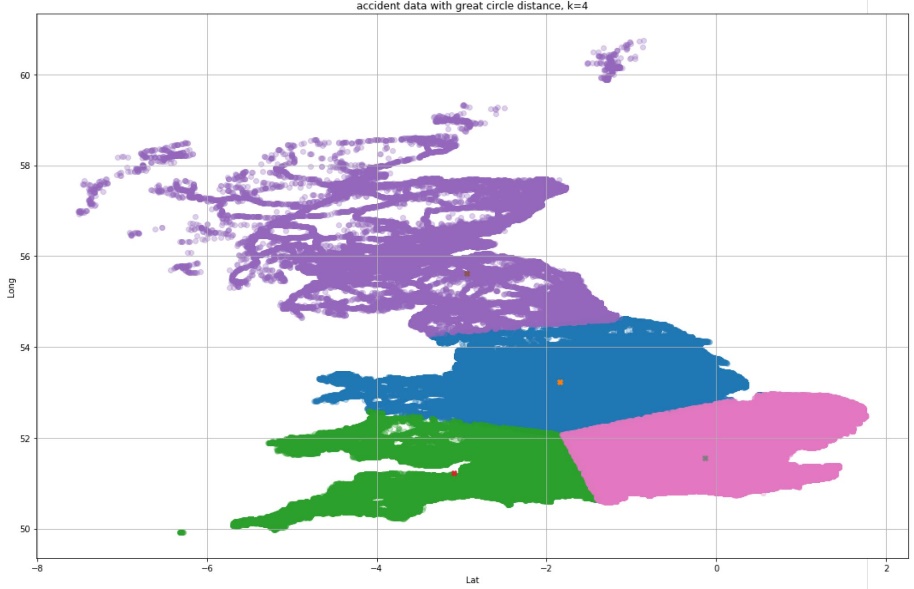


DBpedia\_location\_data



3.3.2 Big Dataset





5. Discussion

The result on 4 clusters indicates that two different distance measurement results in different distances between clusters. The clusters in great circle measurement are closer to each other than in Euclidean distance measurement, however, great circle measurement is more complex than Euclidean distance measurement from perspective of algorithm analysis. Our big dataset derives from some area of European. The computation advantages could not be evaluated from the datasets we applied.

6. Conclusion

3. small data/pseudo cluster

1) [code] implementation (20%)

2) [write-up] results/discussion (10%)

4. Big data application/dataset

1) which dataset/application – creativity (5%)

2) [write-up] description (5%)

3) [code] implementation/execution (10%)

4) [write-up] results/discussion (10%)

5. [write-up] final conclusion/lessons learned/future work (5%)

**Team Member Roles and Responsibilities**

* **project manager**
  + communication/coordinator among **all team members**
  + manages submission to SVN repository
  + set up and maintain team repo (optional)
  + [report/write-up] template
  + [report/write-up] motivation and introduction – coordinate with **key user**
  + [report/write-up] documentation of approach – coordinate with **developers**
  + [report/write-up] conclusion – coordinate with **key user**
* **developer local**
  + implementation
  + [report/write-up] documentation of implementation
  + testing locally
  + testing pseudo-cluster together with **developer cloud**
* **developer cloud**
  + assists **developer local**
  + testing pseudo-cluster
  + cloud execution
  + [report/write-up] documentation of cloud execution
  + assists **key user** with documentation of results
* **key user**
  + data preprocessing
  + find and preprocess Big data application/real-world dataset
  + [report/write-up] documentation of real world data
  + assists **developer cloud** in executing implementation on real-world data
  + [report/write-up] documentation of results