University of New Haven

Tagliatela college of Engineering

Masters in Data Science

GEO-LOCATION CLUSTERING USING THE k-MEANS ALGORITHM

Distributed and Scalable Data Engineering

Date: 05/03/2022

1. Arpitha Busireddy
2. Meghana Reddy Beemireddy
3. Sai Priya Kanuganti

Under the guidance of,

Ardiana Sula, Ph.D.

ASula@newhaven.edu

Contents

List of Figures

1. Motivation

Geolocation clustering mainly refers to the method of grouping the similar objects in order to form a cluster. Geolocation based clustering is mainly implemented in the servers in a distributed manner in order to minimize the downtime and maximize the redundancy of the data. The clustering can also be used in Marketing, Identification of fake news, Document classification etc. In this application we are using K-means clustering algorithm in order to cluster the geolocation coordinates for large dataset. Traditional approaches are unable to handle the large volume of data and can take enormous amount of time in order to perform prediction or generate the output. Therefore, in this work we are utilizing the Apache spark framework in order to perform the clustering analysis over the large dataset. The spark can be setup in two ways either in standalone mode or in cluster mode. Spark MLlib library allows us to utilize the machine learning capabilities in distributed environment over the large dataset.

GitHub url Link : https://github.com/Arpitha26/Distributed_Final_Project_6007

2. Documentation of Approach

In the problem, first we have performed the visualization of different dataset provided to solve this problem using pandas, geopandas and matplotlib library. In the provided problem, we have utilised the 3 different datasets and uploaded to the S3 bucket. The first dataset is device status data which contains the various information of mobile devices such as device id, current status, location etc. The data has been pre-processed and third-party libraries has been used for visualization. After performing visualization of device status data, we got the following map graph as shown in Figure 1.

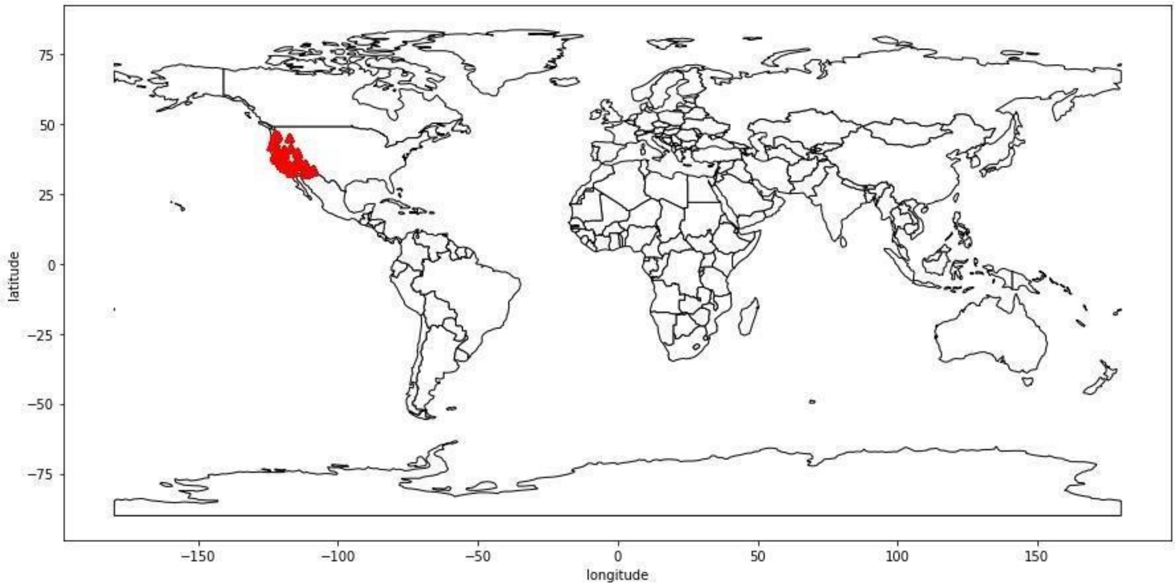


Figure 1 Device Status data Visualization

The second dataset has been used as the synthetic location dataset, which has been pre-processed and visuals has been with the help of map in Figure 2.

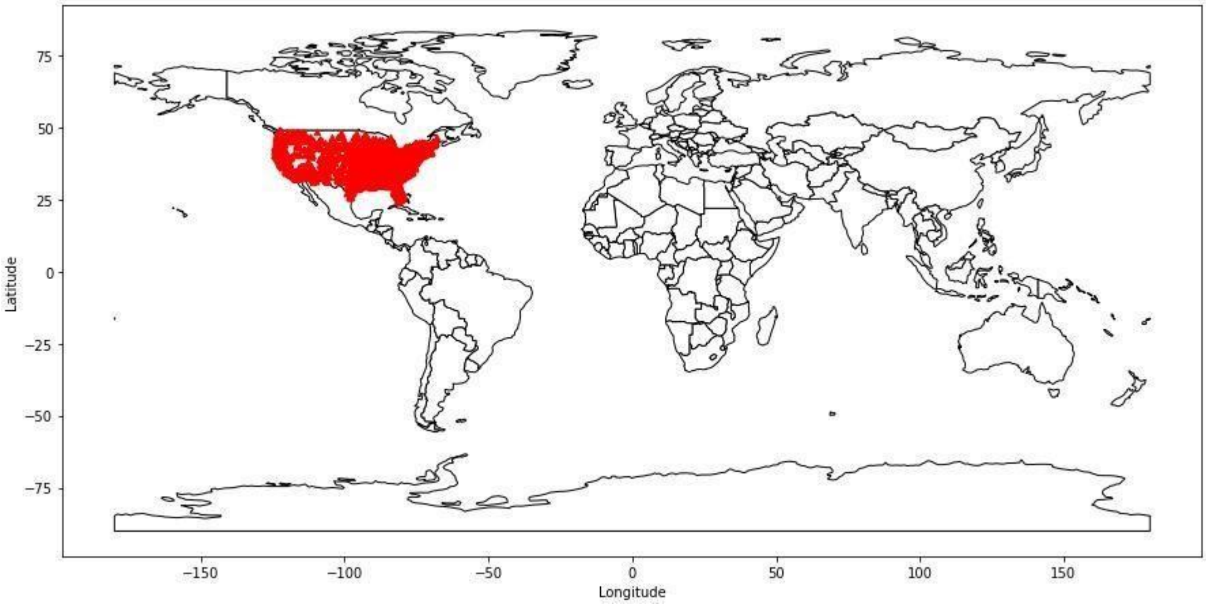


Figure 2 Visualization of synthetic location dataset

Third dataset is DBpedia location data, which has been pre-processed and the generated visuals are shown in Figure 3.

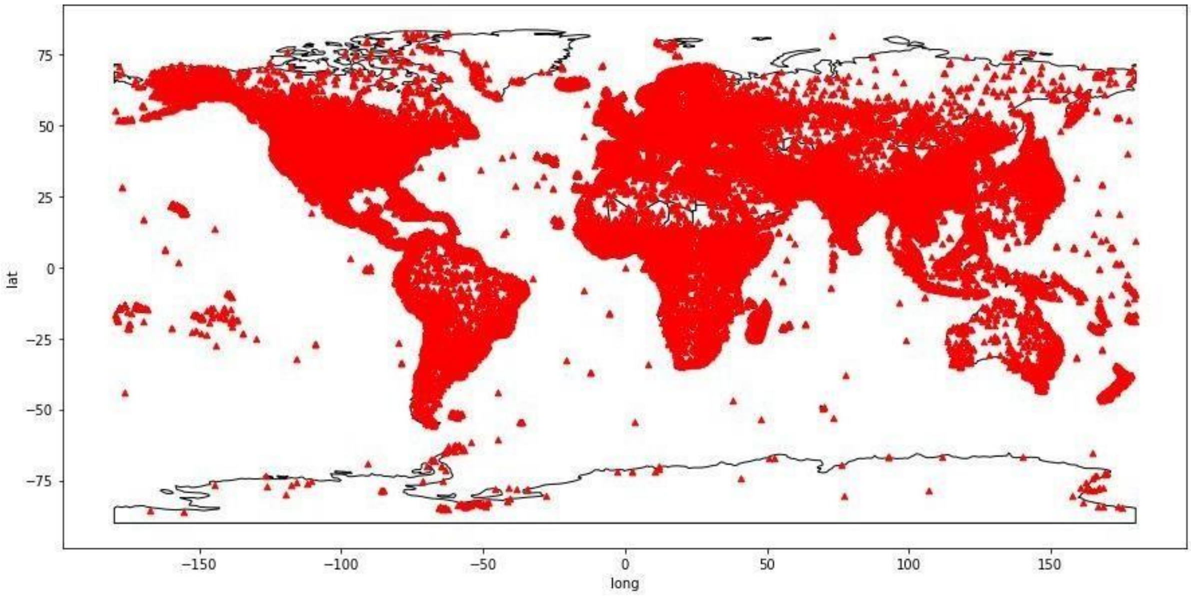


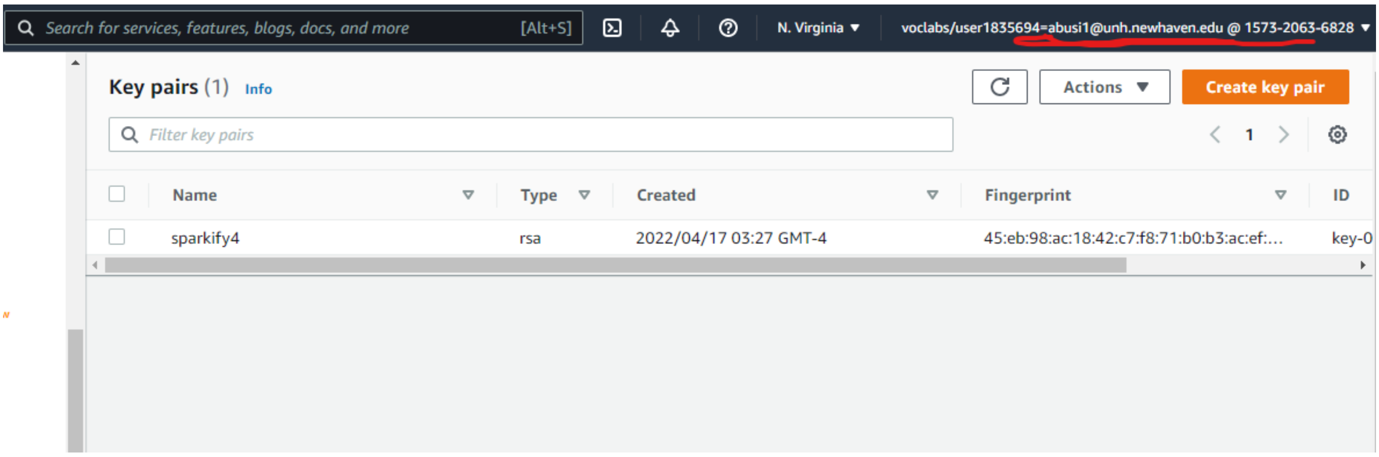
Figure 3 Visualization of DBpedia Location dataset

After performing the pre-processing and visualization, we have applied clustering mechanism using K-means clustering algorithm in Spark using RDD. After implementing the K-means clustering algorithm computation and visualization of cluster is performed and runtime analysis is done using Euclidean distance and great circle distance method.

3. Setup and Configuration of EMR Cluster

a. Creating the Keypair and s3 Bucket screenshots

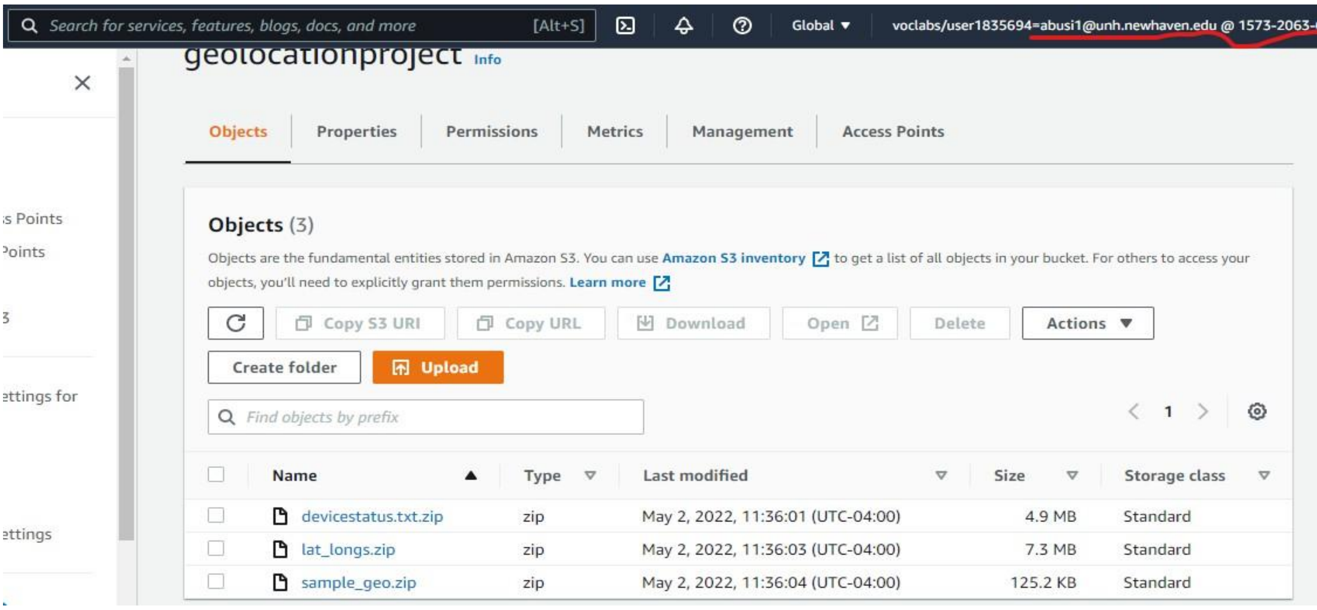
In order to access the cluster instances without any password, the keypair is generated in the form of .pem file. The .pem file is basically a private key required to access the cluster or instance without any password authentication.


Search for services, features, blogs, docs, and more
[Alt+S]
N. Virginia ▼
voclabs/user1835694=abusi1@unh.newhaven.edu @ 1573-2063-6828 ▼
Key pairs (1)
Info
Actions ▼
Create key pair
Filter key pairs
1

□	Name	Type	Created	Fingerprint	ID
□	sparkify4	rsa	2022/04/17 03:27 GMT-4	45:eb:98:ac:18:42:c7:f8:71:b0:b3:ac:ef:...	key-0

Figure 4 Creating key pair for EMR cluster

S3 Bucket Screenshot:



Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

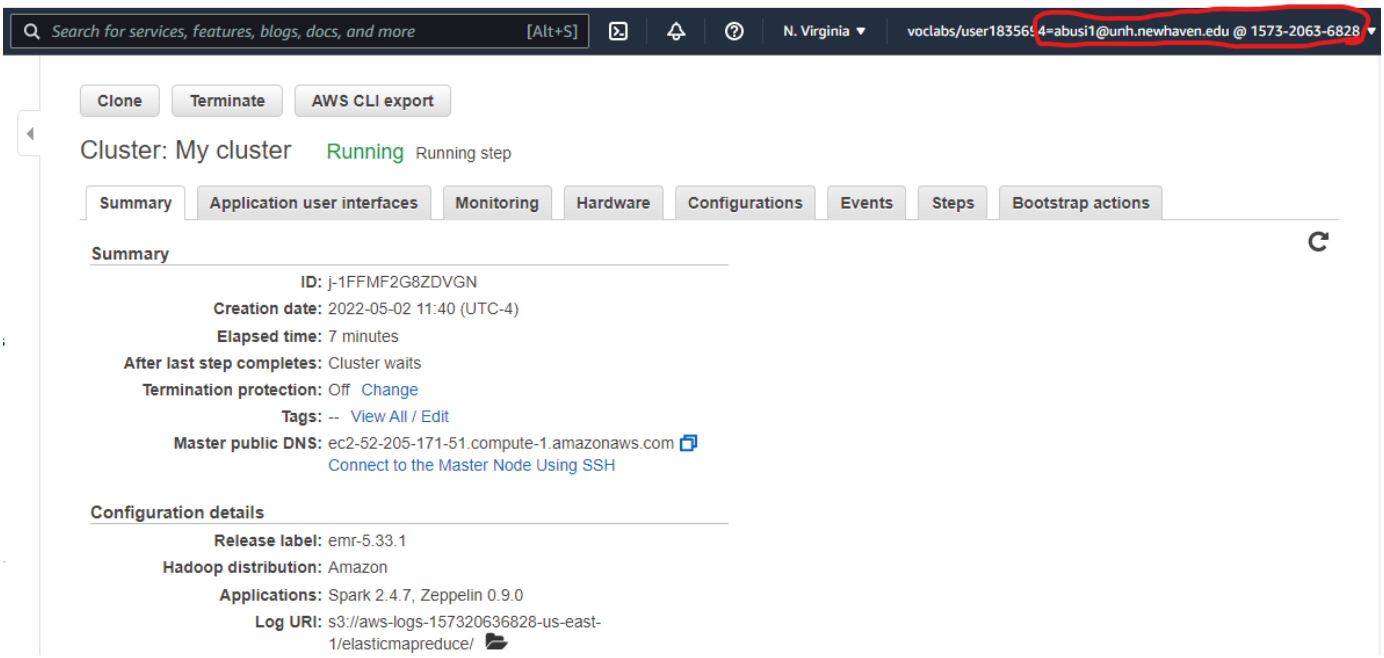
□	Name	Type	Last modified	Size	Storage class
□	devicestatus.txt.zip	zip	May 2, 2022, 11:36:01 (UTC-04:00)	4.9 MB	Standard
□	lat_longs.zip	zip	May 2, 2022, 11:36:03 (UTC-04:00)	7.3 MB	Standard
□	sample_geo.zip	zip	May 2, 2022, 11:36:04 (UTC-04:00)	125.2 KB	Standard

b. Install awscli package:

In order to enter into the AWS CLI environment, we have used this package. By providing the correct credentials and configuration files, we can easily access the aws cli environment.

c. Creating an EMR cluster

Using AWS EMR create-cluster command we have created an EMR cluster. Where the instance count is taken as 3 and instance-type is used as m5.large. Remaining all the other information and configuration setting has been provided in order to create an EMR Cluster. Once the EMR cluster is created and running successfully we have utilized it for running the pyspark code in jupyter notebook.


Search for services, features, blogs, docs, and more
[Alt+S]
N. Virginia ▾
voclabs/user1835694=abusi1@unh.newhaven.edu @ 1573-2063-6828 ▾
Clone
Terminate
AWS CLI export
Summary
Application user interfaces
Monitoring
Hardware
Configurations
Events
Steps
Bootstrap actions

Cluster: My cluster Running Running step

Summary

ID: j-1FFMF2G8ZDVGN
Creation date: 2022-05-02 11:40 (UTC-4)
Elapsed time: 7 minutes
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: ec2-52-205-171-51.compute-1.amazonaws.com
Connect to the Master Node Using SSH

Configuration details

Release label: emr-5.33.1
Hadoop distribution: Amazon
Applications: Spark 2.4.7, Zeppelin 0.9.0
Log URI: s3://aws-logs-157320636828-us-east-1/elasticmapreduce/

Figure 5 Creation of EMR cluster

4. Big Data application/dataset

a. Description

In this application/task the analysis has been performed over three different datasets. The device status dataset contains the information such as date, model, device ID, latitude and longitude. Whereas the synthetic location data contains the attributes such as latitude, longitude and location id. The third dataset DBpedia contains the latitude, longitude, and name of the page.

b. Implementation

In this task K-means clustering algorithm has been implemented over Euclidean distance and great circle distance algorithm over all the 3 datasets. Initially 5 random point has been chosen as the centroid and the distance of each point has been calculated with respect to all the centroids and data points has been assigned to the closest cluster, this process is iterated multiple times in order to get the better cluster analysis.

c. Results & Discussion

As already described in the previous section, we have performed the pre-processing and visualization of all the 3 datasets. After performing these steps, we have applied the K-means clustering algorithm with two different methods on all the 3 datasets. The first method uses the Euclidean distance whereas, the second method is GreatCircle Distance method. For each dataset we will analyse the results in detail.

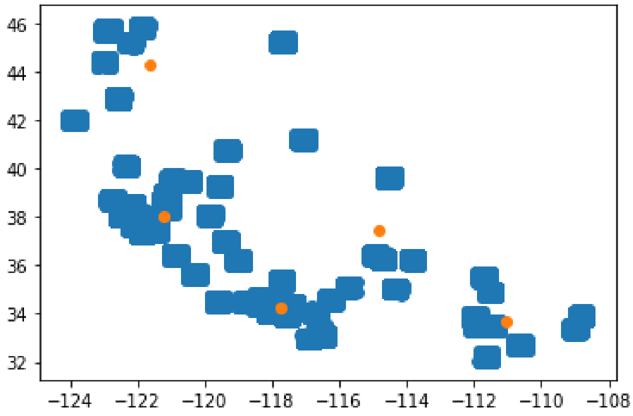


Figure 6 Cluster analysis of Device data using Euclidean Distance method

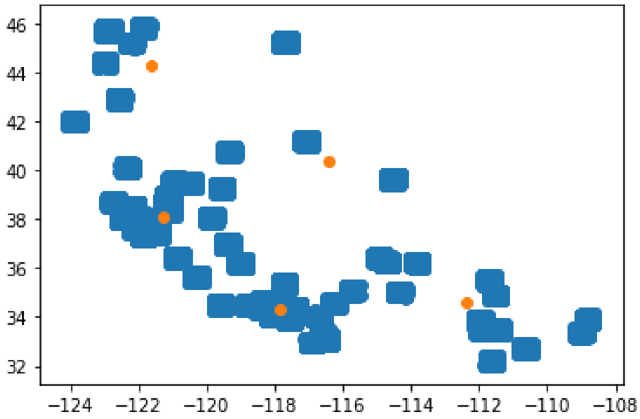


Figure 7 Cluster analysis of Device data using Great circle distance method

After analysing both the clusters, it can be said that Euclidean distance method was able to efficiently form the clusters. On performing a comparative analysis between these two methods with respect to time. Time taken by Euclidean distance method is 38.05 seconds. Whereas, Time taken by Greatcircle distance method is 20.87 seconds.

Now we will checkout the cluster analysis for synthetic location data for K=2 and K=4.

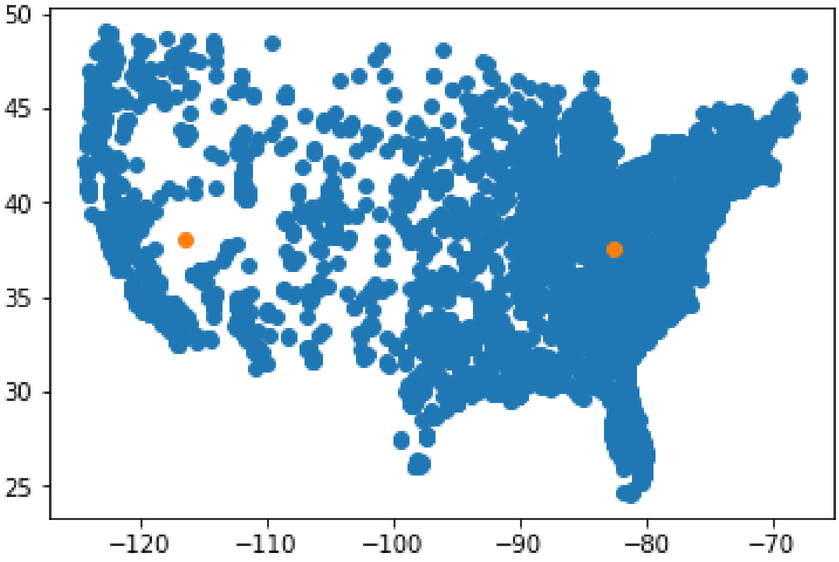


Figure 8 Synthetic data analysis using Euclidean distance with K=2

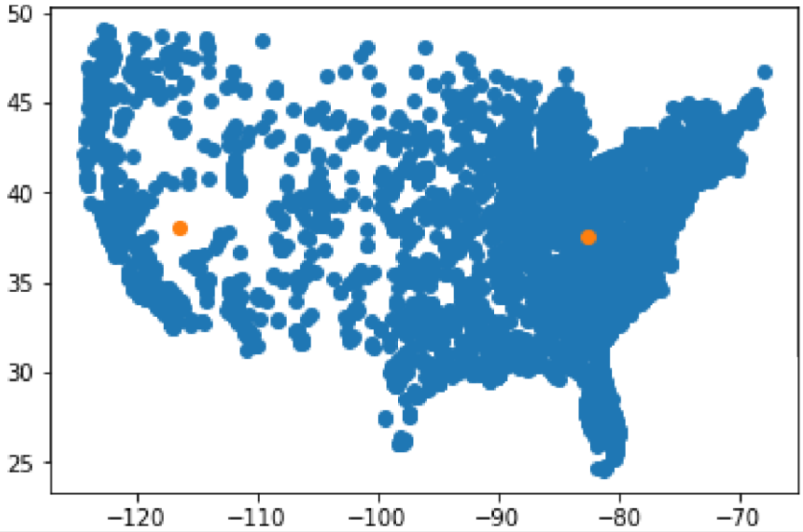


Figure 9 Synthetic data analysis using GreatCircle distance with K=2

Time taken by Euclidean distance for analysis is 1.31 seconds. Whereas, the greatcirlce distance method took 0.649 seconds which is almost half of the time taken by Euclidean distance method for K=2.

Now we will analyse the results for K=4 with both the method over synthetic data.

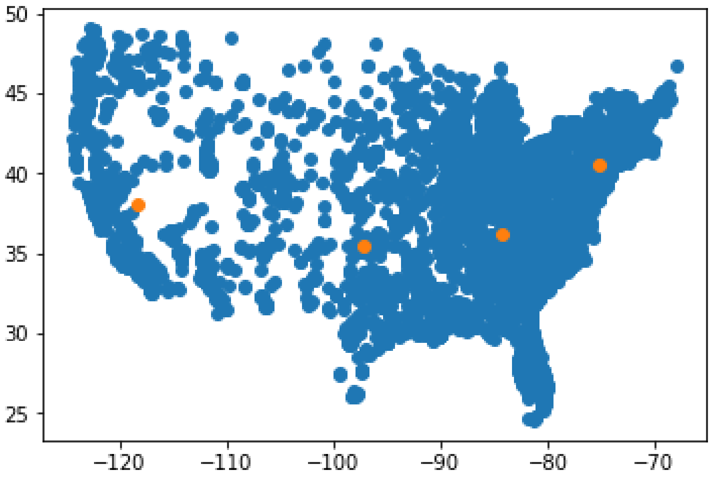


Figure 10 Synthetic data analysis using Euclidean distance with K=4

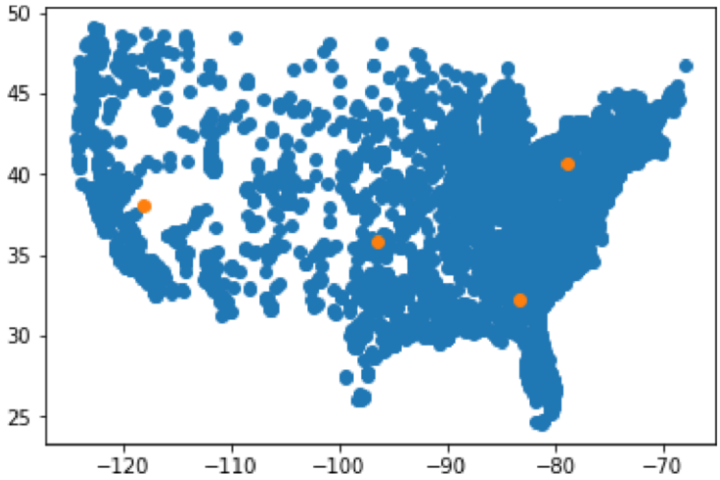


Figure 11 Synthetic data analysis using GreatCircle distance with K=4

Time taken by Euclidean distance for analysis is 5.51 seconds. Whereas, the greatcircle distance method took 1.55 seconds which is very less as compared to time taken by Euclidean distance method for K=4.

Same type of analysis has been performed for DBpedia dataset with the value of K=6. After analysing the K-means clustering, for both the methods we got the following output.

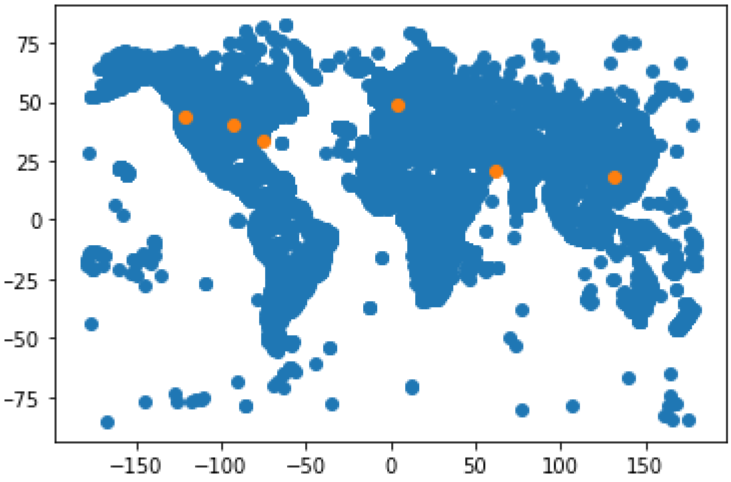


Figure 12 DBpedia data analysis using Euclidean distance with K=6

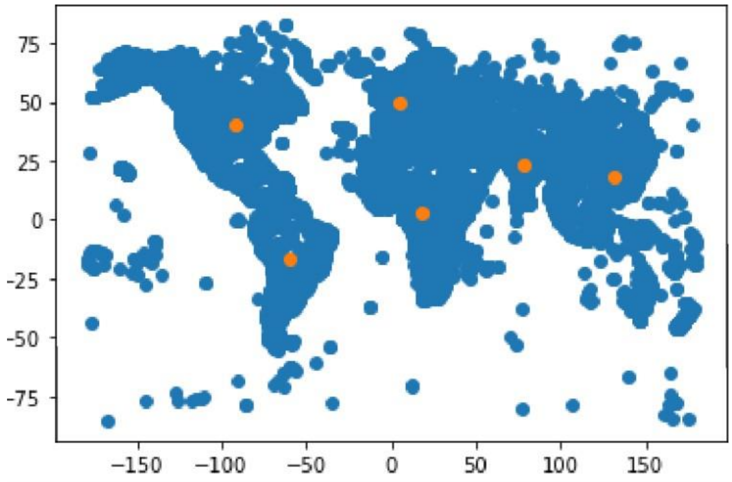


Figure 13 DBpedia data analysis using GreatCircle distance with K=6

Time taken by Euclidean distance for k=6 is 98.71 seconds. Whereas, GreatCircle distance method took 36.54 seconds for the same.

5. Conclusion and Future Work

Overall, we can conclude that time taken by GreatCircle distance method is very less as compared to Euclidean distance method and results are also very close to the Euclidean distance method. K-mean clustering is very powerful algorithm for geolocation clustering, it can be used for various applications such as Identification of fake news, Document classification etc. Spark is very powerful framework in order to perform the analysis over the large dataset, the scalable behaviour of the spark allows it to handle the large volume of data very efficiently. The in-memory computation capability allows to execute the process faster as compared to the traditional approaches. Also, HDFS (Hadoop distributed file system) is very efficient object-based storage method for data storage and analysis. It’s scalable and fault-tolerant nature makes it a better choice for storing the data over distributed cluster.