GEO-LOCATION CLUSTERING USING THE K-MEANS ALGORITHM

Prof: Mr.Vahid Behzadan

Presented by Vivek Garike

Introduction and motivation:

Clustering is a process of grouping a set of data points into a set of k clustering of same objects. Those objects are similar should be in the same cluster and dissimilar objects should be in the different cluster.

Clustering has many useful applications such as group of consumers with similar prefernces,grouping of documents based on the similarity of their contents.

- Marketing: given a large data set of customers transactions and also finding the customers with similar purchasing behaviors.

- Document Classification: cluster web log data is to deter groups of consumer with similar acess patterns.

- Logistics: find the best locations for warehouses or shipinghouses in order to minimize the shipping times.

In order to slove the clustering problem by implementing the k-means algorithm. So, here k-means is a distance-based method that iteratively update the location of k-cluster. The user-defined ingredients of k-means algorithm are distance function and number of clusters k and it's need to be set according to the application domain. In a nutshell, k-means groups the data by reducing the sum of squared distance between the two data points and to their closest centroid.

Data Preparation:

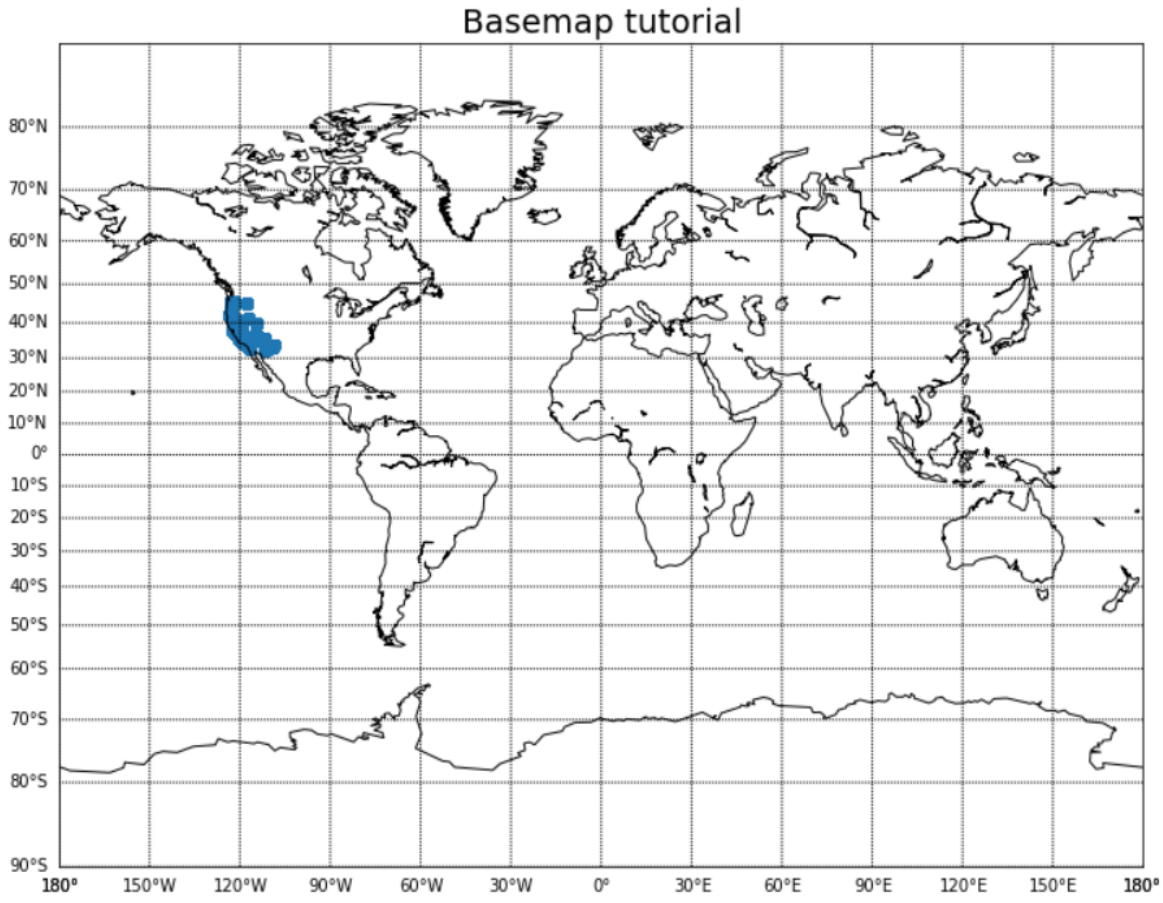
Earlier implementing the actual algorithm we need to start the pre-processing steps to convert data into standardized format for later processing. The following steps will describes the pre-processing steps:

- Firstly, load the dataset
- Determine which delimiter to use.
- Filter out the records which do not have parse correctly; each record should contain the 14 values.
- Extract the date as first field, model as 2nd field, device id as 3rd field, latitude as 13th field and finally longitude as 14th field.
- Store the latitude and longitude as first two fields.
- Filter out the locations of latitude and longitude whose values are zero.

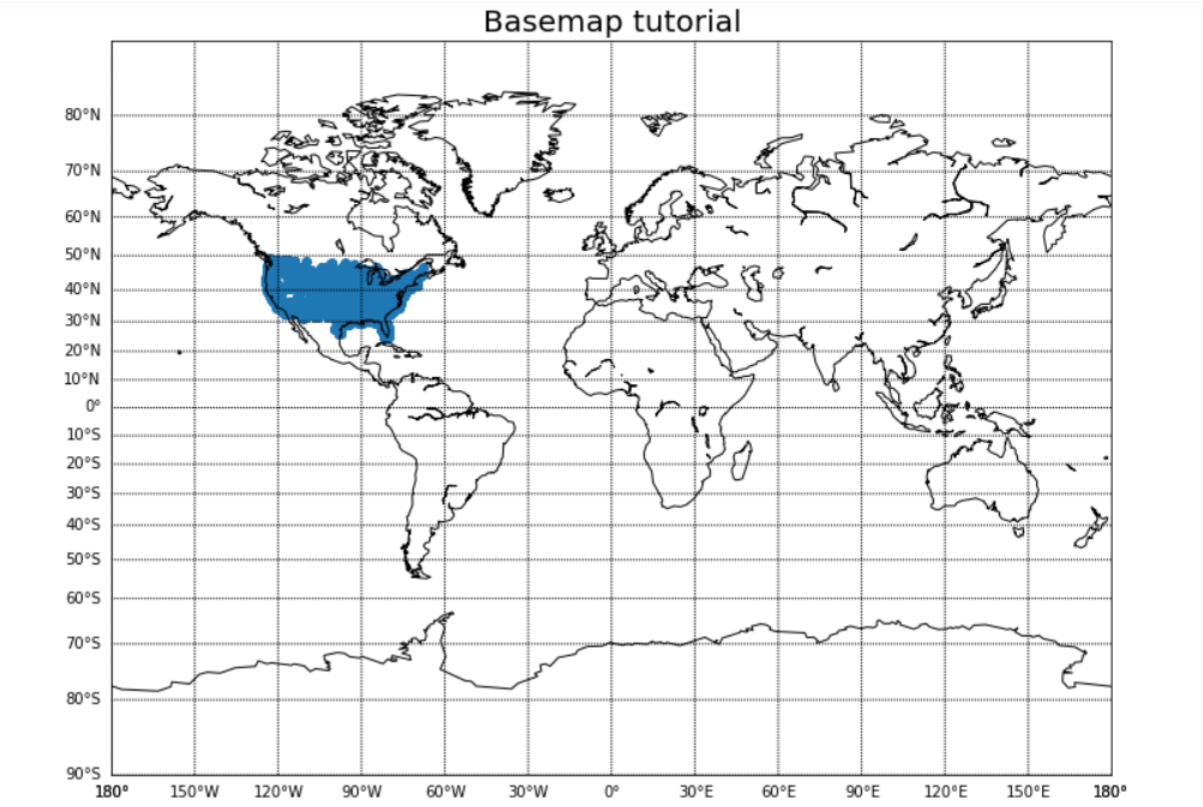
- Split the model field that contains the device manufacturer and model name by spaces.
- Save the extracted data as comma separated values file.
- Just make sure whether the files was saved correctly.

Visualization:

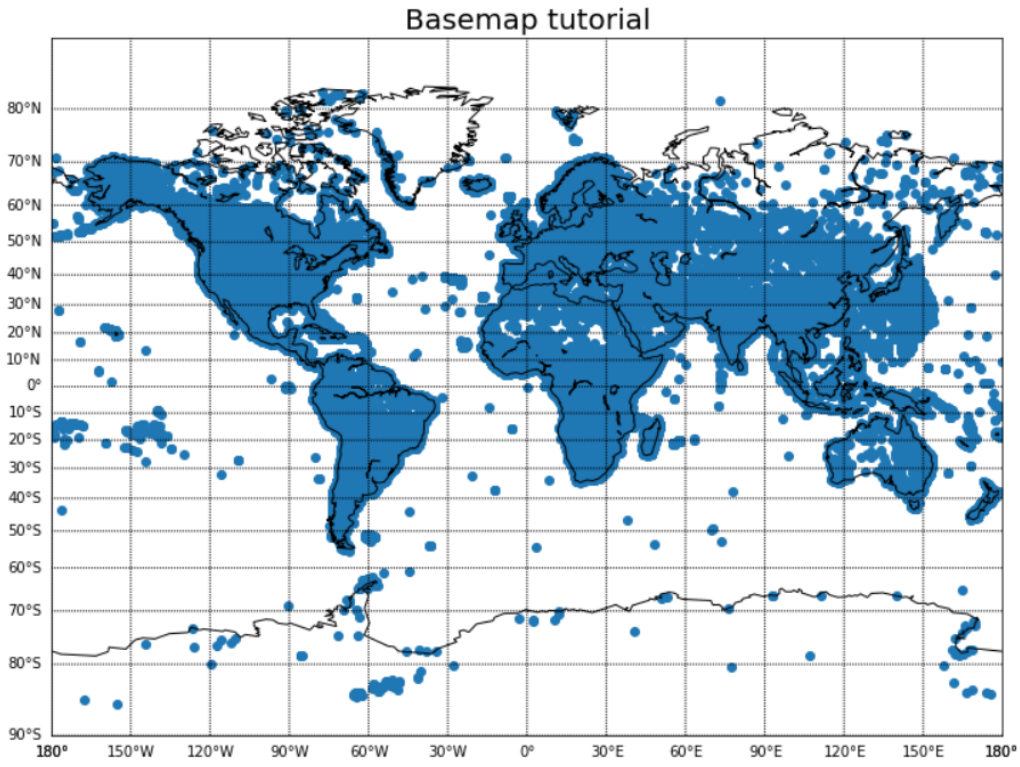
Mobilenet Location Data:



Synthetic Location Data:



DBPedia Location Data:



Approach for Clustering:

The k-sized centroids contains all random sample points in a dataset. On each and every iteration the algorithm assigns each point to its closest centroid and then calculate the new centroids by taking the mean of all points in a centroid's cluster. The difference between the points and centroid can be measured by Euclidean distance. A true algorithm would iterate untile the changes in centroid location converges to 0., this algorithm continues iterating until the sum of all change of centroid location converges to alpha=0.1km.i.e, the algorithm calculates the distance between the new centroid's location and former location on each iteration. We observed that this method gave us better results for the clusters without an extremely long runtime.

Implementing:

- The latitude or longitude point and an array of current centre points returns the index of the closest centre array of given point.
- Sum of two points which returns a point is called addingpoints is used to compute the new cluster centers.
- Euclidean distance gives the difference between the two data points or two clusters.
- GreatCircleDistance returns the great circle distance between the two data points or clusters.
- The used distance and k parameter should be read an input from the command line.

- A suitable convergence criterion and create a variable convergedist is used to decide when the k-means calculations is done.
- Plan and implement the main part of kmeans algorithm.When the iteration is complete, display and return the final k-center points and store the k-clusters.

Conclusion:

The k-means algorithm has many applications as demonstrated by testing algorithm on data sets with different characteristics. This clusters based on geo-locations and that can be applied to many different kinds of dataset with latitude and longitude as features. To determine the meaning information for each application so, it was mandatory to perform the experiment with different values for number of cluster k. After several attempts the mean difference between the cluster and data points was dropped significantly.

Visualization was promptly major part of the project and it is able to process the large number of data in the end.For humans it was mandatory to interpret the meaning of data and clusters yield by the algorithm.

Geo-Location also experiments with different method to alter the run-time i.e; RDD persistence and cloud execution. Where the RDD persistence decreased the run-time of the algorithm. It was different story for cloud significantly improved the runtime. Running algorithm on big

data with amazon elastic mapreduce services was faster by a few degreed of magnitude and it was allowed us to observe power of large-scale cluster computing.