NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

SCE20105 – Track Trajectory of Topic on Geo-Text Data

Presented by Zeng Shijia

Supervised by Prof Cong Gao, Co-supervisors Dr Chen Zhida & Mr Liu Shang

Background

Big data which are geo-tagged and contain textual, temporal information are being generated on an unprecedented scale. The purpose of this project is to mine useful information from such data.

Problem Definition

Given a set of data which has attributes of time, location, text and query keywords we want to track. We aim to develop an algorithm to find the spreading path of this topic from large geo-textual temporal data.

Dataset Description

A sample dataset of 16,205 geo-tagged tweets with timestamp was provided. The data spans from Jan 12 2020 to Jan 30 2020.

Experiment Results

Query Topic	volcanic ash
Clustering Time (5656 data)	20 clusters 1152.51s
Retrieval Time	7 top locations 0.36s

Future Work

In our future work, we plan to try other clustering methods and modify scoring function for better accuracy

Algorithm

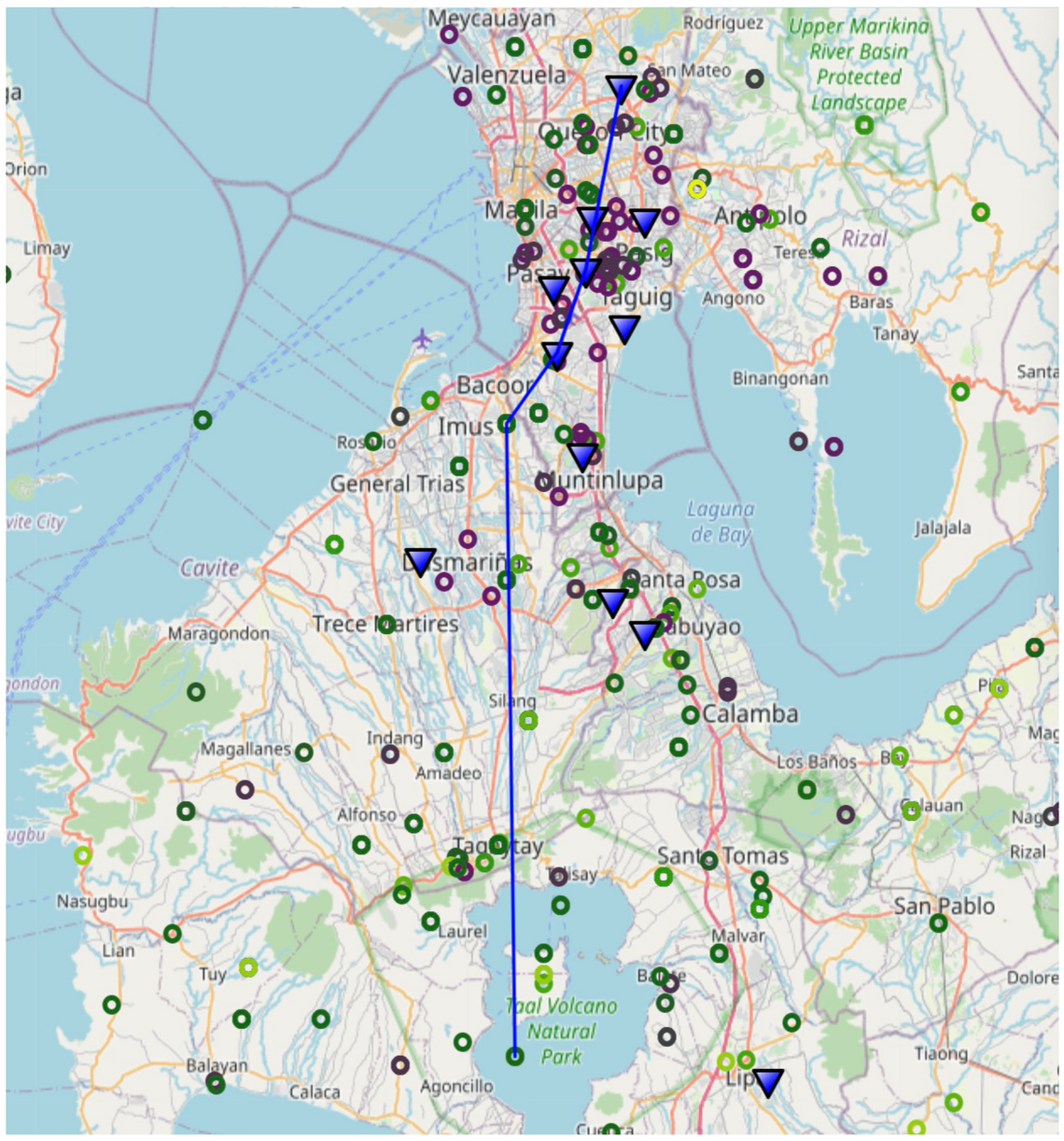
Step 1: cluster all tweets on semantic, geographic distance, and time with defined loss function using unsupervised machine learning method

$$L = \sum w_1 \times dis + \ w_2 \times time + w_3 \times text$$

Step 2: calculate average score of each cluster with query topic based on text and time

$$S = w_1 \sum textSim + w_2 |cluster| + w_3 \sum e^{\alpha |currTime - time|}$$

Step 3: retrieve top-k clusters and their centroids. Then link them on temporal order to build the trajectory of the query topic



Trajectory Graph