Social Media Community Detection

MAIS 202 – Final Project

Dataset:

https://archive.org/details/twitter_cikm_2010 [1]

The dataset is divided into two parts, training set and test set.

1. Training set contains 115,886 Twitter users and 3,844,612 updates.
2. Test set contains 5,136 Twitter users and 5,156,047 tweets.

The geolocation of each user is in the dataset. Given locations are limited to cities in the United States. The content of the tweet is included as well.

Methodology:

- Firstly, combine the user table with the tweets table. This will make the analysis much easier moving forward.
- K-Nearest Neighbours is the most fitting model, since we want to cluster the data based on location and similarities in tweets.
- The final result should be represented in a graph like diagram, with nodes' size and color depend on the population of each community.

[1] Z. Cheng, J. Caverlee, and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In Proceeding of the 19th ACM Conference on Information and Knowledge Management (CIKM), Toronto, Oct 2010. (Bibtex)