Social Media Community Detection

MAIS 202 – Final Project

Dataset: https://www.kaggle.com/kazanova/sentiment140 [1]

Please note that the dataset was changed. Mainly because the previous dataset had major flaws that would have made it very difficult to clean.

1. **Problem statement:**
   The new updates in the project model (mentioned below), require slight adjustments to the definition of a community on social media. Throughout this project, I will assume that a community is defined/characterized by the similarity of tweets between members.

2. **Data Preprocessing:**
   The new dataset contains 1.6 million tweets and around 659 thousand users. After considering the new machine learning model, I have decided that I would only need the username and tweet columns. Since the users are not unique, and having separate tweets does not make sense, I decided to concatenate all the tweets of each user. This way, I will not be losing any valuable information. Additionally, the tweets were lightly cleaned, removing most non-alphanumeric characters. It is good to point out that hashtags (#) were left in the dataset. Hashtags are probably the strongest indication of communities. Please have a look at cleaning.py for more information.

3. **Machine learning model:**
   Previously, I was planning to have a basic KNN model. There were two main issues with that idea. Firstly, datasets with exact geolocations do not exist (publicly). Secondly, if we use geolocation, then we are not really using any of the “social media” aspect of our data. Therefore, the new model had to be some sort of NLP model.
   New Model: tf-idf will be used to vectorize the tweets, and then K-means will be implemented to cluster the tweets. The hyperparameter, K, will be finalised after the model is fully implemented. The initial value of K is preferred to be around 4500, making the size of each community ~150 users.
   Note: Having an unsupervised model means the validation process is far from trivial.

[1] Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford, 1(2009), p.12*.