Twitter New Term Recommendation System

Project Brief

Aifeng Yun August 21, 2015

## Summary

We would like to develop a system to actively detect the most popular words and phrases which are burgeoning on the Twitter platform, and automatically recommend related contents to users based on community detection and sentiment analysis results of these terms.

## Backgrounds

Nowadays, social networks are the most efficient platform for information dissemination. New ideas, new trends and new language are transmitting on these platforms freer and faster than any traditional platforms, making larger audiences and having more interactivities with users. Twitter, given it’s light-weighted nature, is popular among individual storywriters and involving more “live” language where new terms originates. Its short paragraph nature, also makes it efficient for mining the keywords from.

For any of the natural language used by people, there’s never a static set of vocabulary. Languages are changing everyday. New words, phrases are burgeoning every minute, to descript new things, new feelings, new ideas and new trends. Studying of these new terms can give us great hints about where a language is going, and even to predict some changes in society. Recommendation of related information based on analysis of these new terms can be a more efficient method than the traditional way which based solely on relationships between people.

## System Architecture

The system is made up of the following components:

* **Crawler** – Crawling of raw data from Twitter, using both API and web spiders.
* **Term Filter** – Extracting qualified new terms from the raw data.
* **Database Maintenance** – Routines to access and maintain the database.
* **Community Detection** – Detect relationships between users who uses the new terms.
* **Sentiment Analysis** – Analysis the sentiments people towards the new terms.
* **Recommendation** – Determine what information to recommend to users based on CD and SA results.
* **Presentation** – Graphic User Interfaces for easy of access.

## Current Development Status

We have already developed a basic version which can run on a Linux server, with the following limitations:

* The system can only extract single words. Phrases detection require further implementation.
* The system cannot provide real-time results due to Twitter API rate limits.
* The database maintenance routine seems to have defects, which is not capable for large flow of data and long term running.
* The Community Detection and Sentiment Analysis components are running asynchronously, making difficulties to the recommendation to work.
* The Recommendation component is still under development. The web UI is merely displaying Community Detection and Sentiment Analysis results.

## Development Plan

* Changing the host environment from Google Cloud Platform to Amazon Web Services. This is merely an economic decision. The migration may takes some time as the platforms have differences.
* Rewrite the database routines for better performance and reliability.
* Implement the new phrases detection functionality. This may require complex logic and external vocabulary databases for filtering.
* Use web spider technologies such as Scrappy to bypass the API rate limit issue.
* Design mechanism to synchronize the running of the Community Detection and Sentiment Analysis components. This may require some kind of refactoring.
* Fully implementation of the Recommendation component.
* Redesigning of the web user interface, may add app UI for using on mobile devices.

## Task arrangement

I am in charge of the following components:

Crawler, Term Filter, Sentiment Analysis, Recommendation.

While Qian Ma is mainly for the rest of the components, listed as follows:

Database Maintenance, Community Detection, Presentation.