CS 527 Project Proposal

Matthew Zinke Raymond Swannack

1 Introduction

With the advent of social media, social data have become an important source of insight into the activities of consumers of these services. Websites like Facebook, Twitter, Pinterest, LinkedIn, and others hold vast quantities of information about the people using them. This data can be used to track and predict social trends, map relationships between users, and make product recommendations. For this project we will research different ways Twitter data can be stored and categorized to make useful observations and build a system that can be used to gain insight into social trends and user preferences.

2 Objectives

The objective of this project is to explore previous research in the area of reading, categorizing, and storing data from twitter and implement our own system based on this research. We hope to learn more about the possibilities and limitations of collection methods as well as potential conclusions that can be drawn from the collected dataset.

3 Previous work

Favorite, retweet, and reply data is important to organizing and ranking the popularity of tweets. One published work describes utilizing top-k algorithms to rank social media objects from real-time streams [1]. The research targets social media in general but applies to our use case.

Social graphs may contain valuable information about relationships between users. One publication proposes Twigraph, a schema for graphing relationships between users based on word use [2]. These relationships are

then used to recommend user profiles, advertisements, and articles to users based on profile activity.

Sentiment Analysis on social data is another popular topic right now. Research that has targeted twitter specifically has run into issues not found elsewhere, such as the 140 character limit on tweets and the use of texting shorthand to overcome this limitation [3]. This sentiment analysis can be used to monitor wider social trends, and has been used to attempt to predict political elections and outcomes [4].

4 Proposal

Using Twitter's Streaming and REST API, we will fetch streams of tweets based on search criteria and filters, then store this information into a MySQL database. This will allow us to test different collection and storage techniques, and then experiment with the data ourselves. The following section outlines the system we will build to do this.

4.1 Technologies and Tools

We will use some common technologies and tools to achieve our goals.

1. HTML5, Javascript (web client)
2. Javascript (Node.js v8.x), npm (server)
3. npm packages: oauth (authorization), mysql (database driver)
4. MySQL
5. git
6. Chrome
7. Twitter

4.2 Modules

The following section contains a description of the main modules we will build for data collection, storage, and presentation.

4.2.1 Twitter Stream Database

A MySQL database will be used to store tweets and metadata. Current proposed database tables follow, these are subject to change based on research.

- Tweets
- Hashtags
- Users
- Follows (showing follow relationships between users for social graph)

4.2.2 Twitter Stream Reader

A Node.js application will open connection to twitter stream, receive tweets, and store them to MySQL database. The stream reader will be able to filter out unwanted tweets so they are not recorded to the database. Example filters follow.

- Blacklist users (filter out users)
- Whitelist users (filter for certain users)
- Filter for keyword (contains/does not contain)
- Filter for hashtag (contains/does not contain)
- Location filters (within N miles of set location(s))

4.2.3 Application Interface

We will provide a simple application for querying and exploring the data. It will consist of a REST API and an HTML5 Client application.

Rest Interface We will write a RESTful application server in Node.js to make tweet database contents available to other applications, mainly an HTML5 client application that will be served via HTTP.

HTML5 Client Application A client application will be developed to display calls to the REST interface and navigate the data. Some proposed client views follow.

- Total Statistics (tweets stored, uptime, etc)
- Timeline plots showing activity over time periods
- Rankings (most active users, most liked tweets, etc)
- Server Configuration

References

[1] N. Vouzoukidou, B. Amann, and V. Christophides, "Continuous top-k queries over real-time web streams," *CoRR*, vol. abs/1610.06500, 2016.

[2] D. Sundararaman and S. Srinivasan, "Twigraph: Discovering and visualizing influential words between twitter profiles," *CoRR*, vol. abs/1706.05361, 2017.

[3] E. MARTNEZ-CMARA, M. T. MARTN-VALDIVIA, L. A. UREA-LPEZ, and A. R. MONTEJO-REZ, "Sentiment analysis in twitter," *Natural Language Engineering*, vol. 20, no. 1, p. 128, 2014.

[4] J. Kulshrestha, M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, K. P. Gummadi, and K. Karahalios, "Quantifying search bias: Investigating sources of bias for political searches in social media," *CoRR*, vol. abs/1704.01347, 2017.