
Team Dahlia Capstone Project Proposal

Instant visualization of Twitter data using an online dashboard

Meihao Chen
Center for Data Science
New York University
New York, NY 10003
mc5283@nyu.edu

Yitong Wang
Center for Data Science
New York University
New York, NY 10003
yw652@nyu.edu

As Twitter has become a global and real-time mirror of what is happening in the world at any given moment, being able to access tweet information opens up a world of possibilities: from monitoring mentions of your brand or event (think of it as real-time market research) to performing demographic analysis to using Twitter activity as research material, uses of this feature are endless. With an online dashboard that combines all this with the analysis and visualization capabilities, you can execute super advanced analyses of what is happening around you.

1 Related Works

There exists many sophisticated works that can bring us insights. CartoDB, the commonly adopted geographical visualization tool has recently launched CartoDB Twitter Maps to perform geospatial analysis and visualization of Twitter activity. With this new feature users can search for the Twitter activity they'd like to visualize and analyze, select the period of tweets they want, and start mapping all in a matter of minutes. Another successful toolkit would be the Digital Methods Initiative Twitter Capture and Analysis Toolset (DMI-TCAT), which is a set of tools to retrieve and collect tweets from Twitter and to analyze them in various ways. It is written mostly in PHP and runs in a webserver (LAMP) environment.

2 What to do

We plan to create an online dashboard for visualizing tweets with both basic functions such as file uploading and downloading, field selection, query parsing, descriptive analysis and visualization, as well as more advanced natural language understanding.

2.1 Basic Dashboard Functions

The dashboard will contain a UI to conduct file read-in and write-out. After the users upload data in JSON format, they will be able to select field and instances that they would like to analyze and visualize.

Ideally, the dashboard will have a query system for users to query on interesting terms with respects to the data they upload. Apache Solr is a decent candidate for query processing.

One interesting thing about tweet is that it reflects hotspot. Therefore the dashboard should analyze information infusion networks, or the most retweeted tweets and most popular accounts. This online dashboard will also have the ability to conduct most of the common descriptive visualization options, including pie chart, histogram, time series, bubble chart, etc.

2.2 Natural Language Understanding

We would also like to implement natural language processing on tweets. Topic modeling and visualization would be a very powerful and one of the most important feature of this dashboard as it will bring great insights about trending topics to users.