The Twitter Connection [Working Title]

Gabe Benjamin, Cole Harbeck, and Adrian Wan
Swarthmore College
{gbenjam1, charbec1, awan1}@cs.swarthmore.edu

1. PROJECT PROPOSAL

1.1 Problem

A description of the problem you plan to solve and the motivations for doing so (i.e., why this problem is interesting/important).

Twitter is a social media platform with 284 million users active per month, tweeting approximately 500 million tweets per day. [6] At a maximum of 140 characters per tweet and about 1 byte per character[1] this represents an informational flow of over 70GB of text per day. An average book is about 2MB in size [2], so Twitter users are collectively writing about 35,000 books a day.

With this wealth of textual information that is often supplemented with geo-location data and content-connecting hashtags, it is no surprise that a multitude of tools have arisen to harvest the information encoded in tweets. Twitter's API [7] has been used to build visualizations of tweet locations and several third party apps have been created: Tweet Ping is a website that tracks live tweets and displays them like lights on a global map [5] [4]; A World of Tweets is a similar website that builds a heatmap of tweets [3]; Tweet Beam creates a wall of tweets based on a given hashtag or search query [1]. These are just a few of the many apps that utilize the live stream of tweets from Twitter.

Existing products that leverage the informational wealth of Twitter are mostly interested in geolocation, and are either purely historical or purely live. Few, if any, current products focus on the connection between tweets or go in anyway go beyond simply using geolocation or maybe a keyword. We seek to move away from this problem space by creating a product that:

- Uses both historical and live-stream data
- Focuses on relationships between tweets, users, and content, rather than on just their location
- Displays these relationships in a visually pleasing way

[1]Issues arise with different encodings, but we ignore these for the sake of our back-of-the-envelope calculation

Course paper for Swarthmore College course CS 91: Cloud Systems and Datacenter Networks

1.2 Goals

The goals you have for the project. What constitutes success and how will you evaluate it?

Our goal for this project is to have a fully functional visualization depicting the popularity of tweets (based on re-tweets) and the connectivity of tweets being currently tweeted (based on @replies and hashtags). We will also keep a small (several week long) history of tweets, so that users have the option of viewing a graph of historical tweets as well as a new graph of tweets starting from when the program is initialized.

1.3 Hypothesis

Your hypothesis: given the work you intend to do, what are the results you expect to see? How does this work help to solve the problem?

Through this work, we expect to be able to provide a different type of Twitter visualization than what is currently available. We expect to be able to generate graphs that allow us to "watch" different Twitter conversations happen in real time. While we won't be able to see specific tweets, we will be able to see the topics of conversations (based on hashtags) as well as what high-profile accounts are taking part in the conversation in real-time.

The main difference between our visualization of Twitter data and currently existing visualization tools is that we choose to focus on the connectivity and topics of tweets rather than on location information. Current Twitter visualization tools allow users to see where and when people are tweeting. This take on visualization, while cool, may not actually provide any new information about the Twittersphere. Location data for tweets correlates closely with population distribution data (i.e. major cities have a lot of people tweeting in them). Our work will allow us to draw different conclusions from Twitter data. By visualizing popularity and connectivity of tweets, we expect to be able to see popular users (most likely celebrities), what they are talking about, what types of tweets are spawning large conversations, what hashtags are popular, and where they originate. These are only a few examples of what is possible. Although we cannot predict exactly what kind of insights can be provided by this new way of visualizing Twitter data, the above list provides a small example of the kinds of things we expect our program to be able to provide to users.

1.4 Environment

Characterize environment you intend to operate in. Does your project operate on Amazon Web Ser-

vices? on the general Internet? in a data center?

We intend to operate using Amazon Web Services. TwitCon(WT) consists of an always-on EC2 instance

2. REFERENCES

[1] T. Beam. Tweet beam, 2011.

[2] L. Daly. How big is the average epub book?

[3] F. Design. A world of tweets, 2010.

[4] D. D'Orazio. Map tweets in real time across the world with tweet ping.

[5] F. Ernewein. Tweet ping, 2013.

[6] Twitter Inc. About Twitter, Inc., 2014.

[7] Twitter, Inc. Twitter developers, 2014.