Boston
Area
Research
Initiative

RADCLIFFE INSTITUTE
FOR ADVANCED STUDY
HARVARD UNIVERSITY

Summer 2015, TRiCAM REU - BARI project
Hayden Fuss, Elizabeth Brooks, Jeremy McKenzie
email: whfuss@ncsu.edu

Boston Marathon Twitter Dataset

This document provides a brief overview of the files included within this folder. Given geocoded tweets from April 2^{nd} through 9^{th}, and April 12^{th} through 22^{nd}, the week of the Boston Marathon bombing, for the year 2013, we extracted tweets from the greater Boston area which we defined with the following lower left and upper right bounds respectively:
$(42.1575°N, 71.5688°W), (42.6296°N, 70.6616°W)$.

Table of Contents

1. Twitter datasets

The top directory of this dataset includes several CSV files containing Twitter data from the weeks and region described above.

- *cleaned_geo_tweets_4_02_09.csv*, *cleaned_geo_tweets_4_12_22.csv*: geocoded Twitter data from the greater Boston area during the week of the Boston Marathon bombing and the previous week.
- *training_tweets_4_12_22.csv*: $3,600$ random tweets which were extracted from *cleaned_geo_tweets_4_12_22.csv* in order to train the text classifiers we developed.
- *test_tweets_4_12_22.csv*: the remaining tweets from *cleaned_geo_tweets_4_12_22.csv* which were not extracted for training and what we used for analysis of the week of the marathon bombing.
- *twitter_criteria.yml*: a YAML file containing the list of keywords we used for simply classifying a tweet as relevant or irrelevant to the marathon bombing and manhunt. Additionally, it has the twenty-five most uniquely "tweeted-at" Twitter usernames we found within the keyword-relevant tweets, which were used to determine if a tweet may have been informative or not. Lastly, it also contains a string describe the time format associated with the tweets as well as a list useful regular expressions that be can used to clean tweets of markup. This file can be easily read into

a Python program using PyYAML.

2. Hand-classified tweets: *relevance* and *sentiment*

We developed two text classifiers which used support vector machines made with Scikit-learn. The first classifier determined if a tweet was relevant or irrelevant to the marathon bombing and manhunt, thought to be more sophisticated than our keyword dictionary. The second classifier labeled a tweet as either positive, negative, or neutral. In order to develop these classifiers they first had to be trained with hand-labeled, tokenized tweets pulled from *training_tweets_4_12_22.csv* which are in the following folders:

- *relevance*: the text of ~$1,100$ tweets were sorted into two files based on their relevance, *relevant.txt* and *irrelevant.txt*.
- *sentiment*: the text of ~$3,300$ tweets sorted into three files based on their overall sentiment, *positive.txt*, *negative.txt*, and *neutral.txt*.

3. *plots*

Using the keyword dictionary and text classifiers we determined the number of tweets per day which were keyword-relevant, relevant-classified, positive, negative, and neutral for the week of the marathon bombing and the previous week. Within this folder are CSV files containing the number of tweets for each day for each category as well as images of bar graphs of the data.

Files starting '*prev_*' were from the previous week, while files starting with '*tpd_*' were from the week of the bombing. Note, plots which contain '*kw_rel*' within their name are comparing the performance of the keyword dictionary versus the relevance text classifier.

4. *maps*

For the day of the marathon bombing, manhunt, and April 12^{th} we plotted the locations of tweets by hour over a map of the greater Boston region, with the different counties color-coded, the Boston Marathon route in white, and Suffolk county (Boston) with thicker borders.

The location of keyword-relevant, relevant-classified, and sentiment tweets were plotted for the day of the marathon bombing and manhunt; however, only the sentiment tweets were plotted for the 12^{th}. For the day of the bombing, the tweets are shown from 3:00 PM (shortly after the bombing) and for the remaining hours of the day, while the entire 24-hours were plotted for the day of the manhunt and the 12^{th}.

For the keyword-relevant and relevant-classified maps, blue dots represent relevant tweets while orange dots showed the locations of "informative" tweets, which were relevant tweets containing one of the twenty-five most "tweeted-at" usernames. For the sentiment maps, orange was used for positive tweets, while blue and green were used negative and neutral tweets respectively.

The plots were auto-cropped and combined into animated GIF's using the commands found in *auto_crop_gif_commands.txt*, which can be used from the command line on any Unix-based operating system.

Programming Resources

You can find our Python code for this project at the following GitHub repository:
https://github.com/brix4dayz/bari_reu_2015