Bursty Topic Detection with application to Twitter

Rizwan Ali and Ali A. Ghorbani

Intelligent and Adaptive Systems

UNB

Introduction

- Twitter is a social network where users post microblogs which can be related to a diverse field of topics.
- With Twitter's informal fashion of text, traditional topic detection methods which are more focused on news articles and blogs cannot be applied.
- We propose a solution to detect hot topics by analyzing sudden burst in keywords used in the tweets.
- We also keep track of the performance of topics, using a scoring system, for a longer period of time and give topics based upon it.

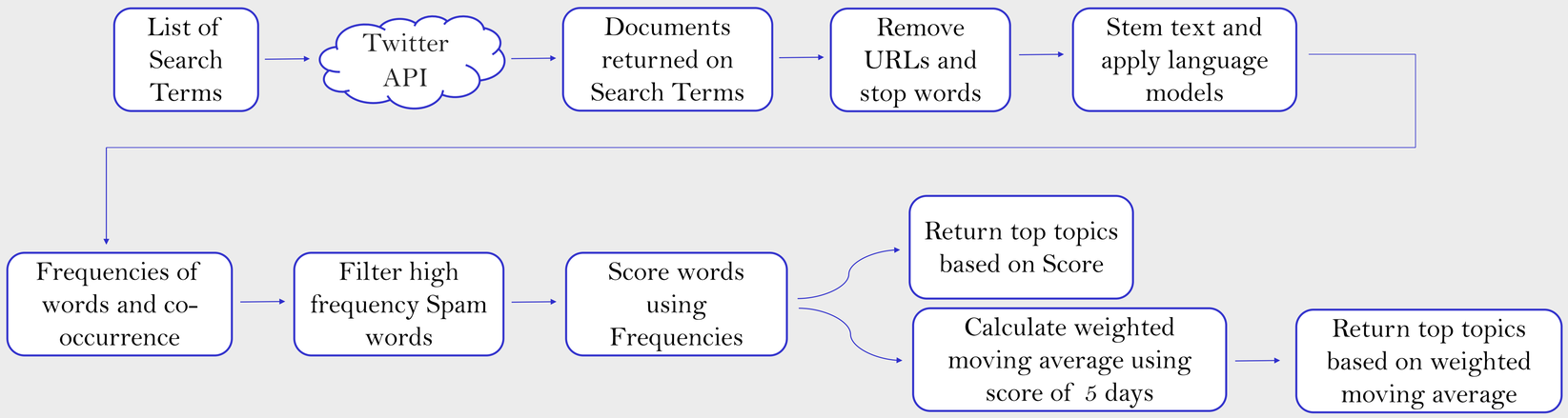
Problem formulation

- Given a set of documents posted on Twitter related to a particular search term, we determine the hot topics being talked about in them.
- Topics reported should be of high quality and easily understandable for the user and reported with their search terms. Multiple search term can have same topics.
- The topics should be tracked over a period of time. Related documents (tweets) to a keyword can be retrieved as well.
- Topic should be detected when there is a sudden surge in the frequency of its popularity, or it is reasonably popular over a long period of time which means it is a regularly talked about topic.

Data Set

- Our data set was gathered through Twitter Search API. We passed pre decided search terms to it. The search terms were focused on republican presidential nomination of US presidential race of 2016. Candidate names and popular hashtags associated with them were used as search terms.

Our Approach



Scoring

- Top topics are scored using bigram and unigram frequency score. The score is formulated using bigram frequency which is adjusted using the below equation. The adjusted frequency F_{ij} is decreased if the individual word frequency is much higher than the bigram frequency.

$$F_{ij} = f_{ij} - w\left[\frac{f_i - f_{ij}}{f_{ij}}\right] - w\left[\frac{f_j - f_{ij}}{f_{ij}}\right]$$

- Where f_i and f_j are individual word frequency and are always equal to or larger than the bigram frequency f_{ij}. A weight w is used to increase or decrease the effect of the equation and can adjust depending on the dataset.

Tracking

- The topics are tracked using weighted moving averages. Highest weight is assigned to the most recent frequency. We track the frequency of five days. The table below gives the weight of each day.

Day	Weight
1st recent	0.5
2nd	0.2
3rd, 4th, 5th	0.1

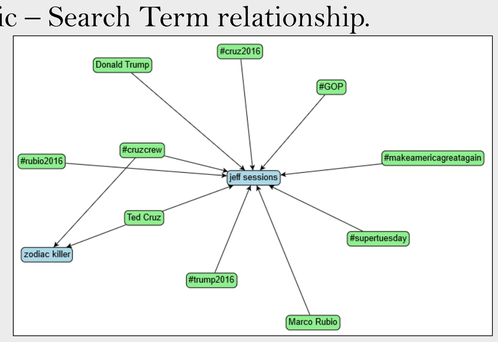
Evaluation

- Experimenting with the system for data set of US republican presidential nomination acquired from Twitter Search API, we saw positive results. The implemented framework was consistently able to identify the top topics of the day, which were verified using news media and twitter trends only related to our data set.
- A single topic was detected against multiple search terms and a combined score was assigned to it.
- Our framework gathered data using Twitter Search API and only specific content related to the search terms returned from the API were used. Other topic detecting frameworks give generic trending topics, but in our framework we can focus on search terms.

Day	Topic	Search Terms
1st March	david duke	Ted Cruz, Donald Trump, Marko Rubio, #MakeAmericaGreatAgain, #SuperTuesday, #GOP
2nd March	chris christie	Donald Trump, #MakeAmericaGreatAgain, Marko Rubio, #SuperTuesday
3rd March	zodiac killer	Ted Cruz, #CruzCrew
4th March	super tuesday	Donald Trump, Ted Cruz, Marko Rubio, Ben Carson, #Trump2016, #MakeAmericaGreatAgain, #SuperTuesday, #CruzCrew, John Kasich
5th March	hillary clinton	Donald Trump, Ted Cruz, #Trump2016, #MakeAmericaGreatAgain, Marko Rubio
6th March	ben carson	Donald Trump, #Trump2016, Marko Rubio, Ted Cruz, #MakeAmericaGreatAgain
7th March	mitt romney	Donald Trump, #Rubio2016, Marko Rubio, #MakeAmericaGreatAgain
8th March	puerto rico	Marko Rubio, #Rubio2016

Results

- Topic – Search Term relationship.



- Document frequencies of top topics of 1st March