Bursty Topic Detection with application to Twitter

Rizwan Ali and Ali A. Ghorbani

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

1. List of Search Terms
2. Twitter API
3. Documents returned on Search Terms
4. Remove URLs and stop words
5. Stem text and apply language models
6. Frequencies of words and co-occurrence
7. Filter high frequency Spam words
8. Score words using Frequencies
9. Calculate weighted moving average using score of 5 days
10. Return top topics based on Score
11. Return top topics based on weighted moving average

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.

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.

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

Results

- Topic – Search Term relationship.

- Document frequencies of top topics of 1st March.