Topical Clustering of Tweets

Hakan Sözer, Muhammed Yağmur Şahin, Yağız Salor

Problem Description

Twitter

- Users post millions of messages everyday

on Twitter

- With different topics
- Hard to find topics that fits your interest
- Twitter has mechanisms built in like Trending Topics(TT)

Trending Topics(TT)

- Works by using frequency of
  - occurence of recent hastags
  - occurence of recent words

  in a specific period[1]

- Not uses the topic of the entire sentence(meaning)

  Examples: "Election"

  Obama wins the election.

  Obama rocks, 4 more years.

Topical Clustering of Tweets

- Automatically clustering and classifying tweets into different categories
- inspired by the approaches taken by services like Google News[2]
- will provide coherent tweet sets that has classifiers as their topics
- which will help browsing of tweets

Motivation

Social Media and Twitter

- Now, Twitter is an important information source worldwide
- Companies, politicians, mayors, celebrities, TV programs…
- Consider Melih Gökçek
- Lots of mentions about him
- By using our clustering approach he can easily see what they are about

Methodology

Approaches

- Unsupervised Clustering
- Supervised Clustering

Unsupervised Clustering

- Unsupervised Clustering approaches are used in the program TweetMotif[3]
  - Automatic extraction of topics
  - Grouping tweets in the relevant topic
  - Summarization of topics

- K means clustering algorithm with TF-IDF weighting

TweetMotif

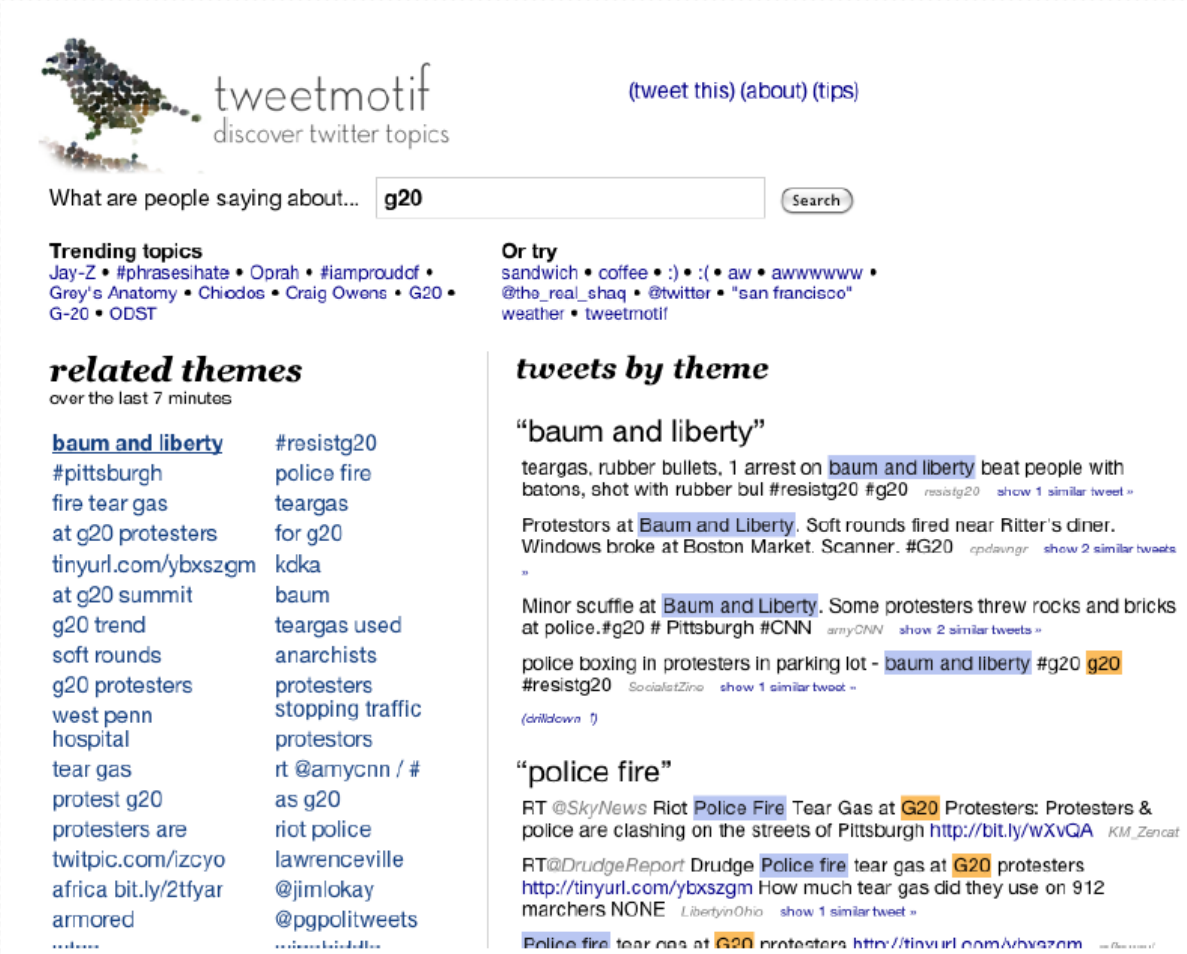

Figure 1: Screenshot of TweetMotif.

Supervised Clustering

- According to article "Twitter Content Classification" hastags (#election) are approximate indicators of topics[4]
- But not all of the tweets have hastags
- Use these hastags as classifiers
- Rocchio Classifier[2]
  - broadly used in document classification
  - quick to train
  - handle feature sparsity more robustly than other models(as the words in tweets are very sparse)

Supervised vs Unsupervised

- These methods will be tested and investigated deeply during our process
- The best one in terms of
  - Performance
  - Effectiveness
  - Cost of implementation

  will be used

- Maybe hybrid method?

Summarization of Tweets

- Find the most representative tweet in cluster
- Use TF-IDF similarity for this selection
- That will be approximate summary of the cluster
- Google News uses this approach. It displays most relevant Story in a cluster of online news articles.

Summarization of Tweets

The algorithm we used to select the top N most representative tweets from a given cluster is as follows[2]:

-Construct a centroid, V, for the cluster of tweets, C

-Initialize an empty list for the selected tweets, T

-Sort all tweets in C according to their TF-IDF similarity with V, where the highest ranked ones are the ones that are most similar

-Loop N times

    -Pick the highest ranked tweet, t, from C, whose TF-IDF similarity against all tweets in T is below some threshold k

        -Add t to T, and remove t from C

Google News


Google
Haberler
Türkiye sürümü
Modern
En Çok Okunan Haberler
S&P'den açılıma tam not!
Vatan - 1 Saat önce
Standard and Poor's da Türkiye'nin notunu 'BB+'ya yükseltti. Not artırımında Kürt açılımının sağlayacağı ekonomik faydaya dikkat çekildi. Uluslararası kredi derecelendirme kuruluşu Standard&Poor's (S&P), Türkiye'nin uzun vadeli kredi notunu BB'den, BB ...
Vatan
Gerçek zamanlı
Star Gazete
Milliyet
Internet H...
Internet H...
Haber Aktuel
Sabah
Star Gazetesi yazarları Elif Çakır ve Miroğlu kaza geçirdi
Akşam - 1 Saat önce
Close. Star Gazetesi yazarları Elif Çakır ve Miroğlu kaza geçirdi. 27 Mart 2013 Çarşamba - 22:18 - Aksam.com.tr. Star Gazetesi yazarları Orhan Miroğlu ile Elif Çakır, Mardin'de geçirdikleri trafik kazasında yaralandı. Hastaneye kaldırılan yazarların durumlarının ...
medya73....
BDP'li Sebahat Tuncel'e polise tokat faturası
Akşam - 2 saat önce
Silopi'de 2011 yılında Nevruz kutlamaları sırasında çıkan olaylarda Emniyet Müdürü Murat Çetiner'e tokat atan BDP'li vekil Sebahat Tuncel 25 bin lira manevi tazminat ödemeye mahkum edildi.(Aksam.com.tr) ...
Adana Ha...
Erteleme maçında kazanan Karşıyaka!
Fotomaç - 45 dakika önce
Maç, iki takımın karşılıklı basketleriyle başladı. Pınar Karşıyaka'dan Thomas, Caner ve Aminu, Banvit'ten Erkan, pota altında başarılı isimler oldu. İlk 3 dakika konuk ekibin 6-2 üstünlüğüyle geçildi. Pınar Karşıyaka, Dixon ve Can'ın skora katkı yapmasıyla 5.
Haber FX
Uzman çavuş eşini öldürdü
ntvmsnbc - 18 dakika önce
Şanlıurfa'da, görev yaptığı Şırnak'taki birliğinden izinsiz olarak ayrılan uzman çavuş, tartıştığı eşini sokak ortasında kurşun yağmuruna tuttu. Talihsiz kadın olay yerinde hayatını kaybederken, kaçan zanlı polis tarafından yakalandı.
Urfa Haber
Tarihi siber saldırı
Vatan - 2 saat önce
Bir grup, istenmeyen e-postalarla mücadele için çalışıyordu ve sunucular arasındaki çekişme sonucunda merkezi alt yapının zarar görmesine sebep oldu. Bu durum Netflix gibi sıklıkla kullanılan hizmetler üzerinde ciddi etkisi olunca uzmanlar bankacılık ve ...
Haberciniz
Dünya »
Kuzey Kore'den Savaş Çağrısı
medya73.com - 1 Saat önce
Kore Yarımadası Güney ve Kuzey Kore arasında yaşanan sıcak saatlere ve

Expected Results

Expected Results

- Results can be rated by comparing for each cluster
  - for each tweets in that cluster
  - the summary statements

- giving a rating in terms of relativity for each comparison

References

1. Algorithms Behind Trending Topics http://www.ignitesocialmedia.com/twitter-marketing/trending-on-twitter-a-look-at-algorithms-behind-trending-topics/
2. Topical Clustering of Tweets,K.D. Rosa, R Shah, B. Lin, A. Gershman, R. Frederking, Language Technologies Institute, Carnegie Mellon University
3. TweetMotif: Exploratory Search and Topic Summarization for Twitter , B.O. Connor, M Krieger, D. Ahn
4. Twitter Content Classification http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2745/2681