New Event Detection Using Public Tweets

Final Presentation

F. Tuğba Doğan - M. Ali Yeşilyaprak - Simge Yücel

Outline

- Problem Definition
- Importance / Motivation
- Input Data
- Methodology
- Results

Problem Definition

- Twitter is a microblogging social media which users can share their daily life struggles, local and global events etc.
- We tried to derive the events from the tweets using these tweets' word distribution, spatial and temporal features.

Motivation/Importance

- Everyday nearly 170 million tweets are created.
- It becomes important to identify the events and bring event related tweets to forward to inform other users about the emerging events.

Input Data

- The data set contains
  - 115,886 Twitter users
  - 3,844,612 updates
- Collected from September 2009 to January 2010
- Link: https://archive.org/details/twitter_cikm_2010

Methodology

Preprocessing

- We used only 15 days' of the data between October 10, 2009 and October 25, 2009.
  - We divided this set of tweets from 15 days into smaller frames which contain 5 days.
  - First and last frames will be used for identifying event and non-event tweets.
- Tweets that contains less than 5 words were eliminated.
- The links has been removed.
- The mentions (which starts with @ sign) has been removed.
- Whole punctuations (except from #) has been cleaned.

Methodology

Stemming & Dictionary

- Stemming algorithms are computational procedures which reduce all words with the same root to a common form
- In our stemming algorithm we took the word and return the words first *n* letters. Each time we have decided the *n* value. Two special cases has been considered.
  - A word starts with ‘@’ sign were ignored and didn’t exist in our dictionary.
  - We took the whole word starts with ‘#’ sign and didn’t apply stemming to those words.
    - The letters were converted to lowercase, then this word was added to the dictionary.
- After applying stemming for each word of the tweets, a dictionary has been created from the unique words. Our dictionary contains approximately 8000 words.

Methodology

Clustering

- Cover Coefficient based Clustering Algorithm (C3M)
  - For each set the number of clusters has been calculated with using C3M Algorithm.
  - For each set binary D matrices has been created, because it is probabilistically very low to get a word two times in the tweet.
- K means clustering
  - tf-idf vectors have been calculated for each tweet
  - For each set, tweets have been clustered

Methodology

Event and Non-event Detection

- The sets has been combined, then tf-idf matrix has been calculated through all tweets in the set.
- Cluster mean vector has been calculated for each clusters in the sets.
- For each cluster center from the last frame, closest cluster center from the first frame has been selected
  - The similarity between two clusters has been calculated by Jaccard Index:
- The most closest clusters have been eliminated.
  - Tweets about the specific event cannot be remain on the agenda during 15 days.

Results

- C3M Results:
  - First set has 354 clusters.
  - Last set has 368 clusters.
- Top closest clusters in first and last frame contains spam tweets
- Mid closest clusters in first and last frame contains daily life tweets and event tweets
  - Daily life clusters are more similar to each other than event clusters
- Last closest clusters in first and last frame contains mostly noises

References

[1] Becker, Hila, Mor Naaman, and Luis Gravano. "Beyond trending topics: Real-world event identification on twitter." (2011).
[2] Abdelhaq, Hamed, Christian Sengstock, and Michael Gertz. "Eventweet: Online localized event detection from twitter." *Proceedings of the VLDB Endowment* 6.12 (2013): 1326-1329.
[3] Ozdikis, Ozer, Pinar Senkul, and Halit Oguztuzun. "Semantic expansion of hashtags for enhanced event detection in Twitter." *Proceedings of the 1st International Workshop on Online Social Systems*. 2012.
[4] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 759-768. ACM, 2010.
[5] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang. Tedas: A twitter-based event detection and analysis system. In Data engineering (icde), 2012 ieee 28th international conference on, pages 1273-1276. IEEE, 2012.
[6] J. B. Lovins. Development of a stemming algorithm. MIT Information Processing Group, Electronic Systems Laboratory Cambridge, 1968.

Thank You For Listening

Any Questions?