CS 690V – VISUAL ANALYTICS
HOMEWORK 6

Submitted by:

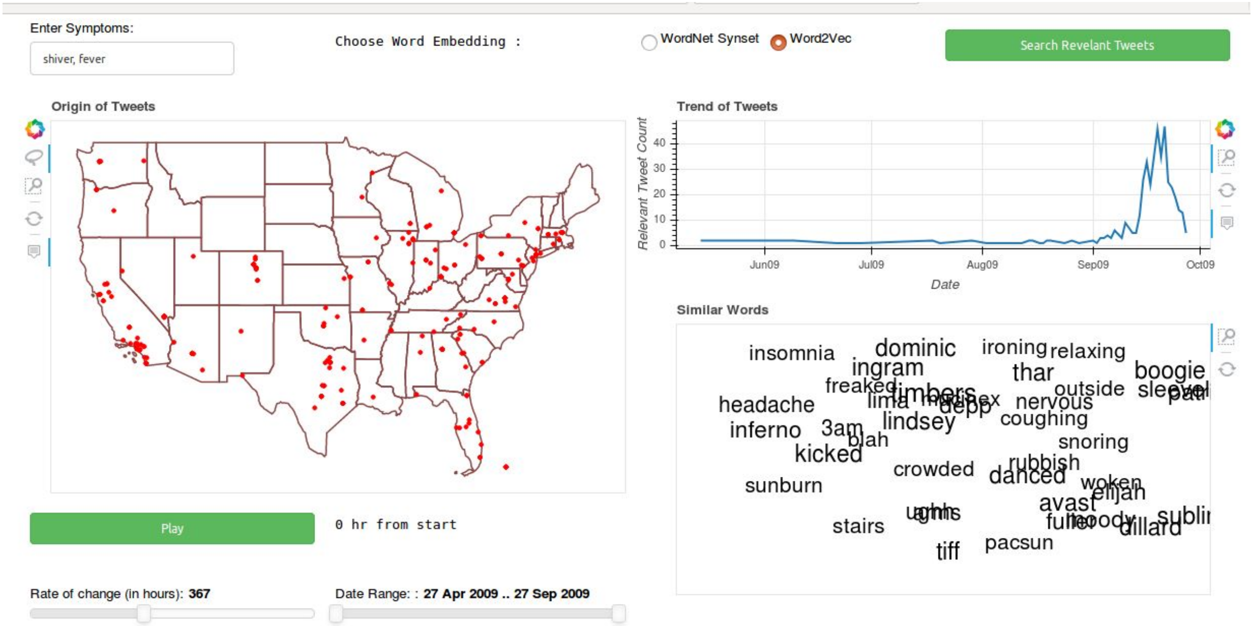
- Suhas Keshavamurthy
- Kriti Shrivastava

VAST Challenge 2011: MC 1- Characterization of an Epidemic Spread

http://hcil2.cs.umd.edu/newvarepository/VAST%20Challenge%202011/taskdescription-of-all2011challenges-printfromoriginalwebisteofchallenge.pdf

Solution Idea:

Using this dashboard, the user can analyze the spread of epidemic across the city. A text input is provided at the top where the user can enter the symptoms/keywords (*in a comma separated format*) using which the tweets will be filtered. The filtered tweets are then plotted across the map. The play button at the bottom of the map gives the flexibility to see the trend in tweets across the country over time. The video can be paused at the point of interest. Option for selecting a date range is also provided to examine the location and tweets over the selected time period more closely. Hovering on the point on map displays the tweet text. The line graph shows the trend in number of relevant tweets over time. The word cloud gives an idea of words similar or closely related to the searched symptoms/keywords. Lastly, there are two approaches that the user can select from, WordNet synset and Word2Vec.



Enter Symptoms: shiver, fever

Choose Word Embedding : ○ WordNet Synset ● Word2Vec

0 hr from start

Rate of change (in hours): **367**

Date Range: : **27 Apr 2009 .. 27 Sep 2009**

Figure 1. Overview of the system

Sample Data:

We are using Twitter data for our analysis. (You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users in CIKM 2010.) The dataset contains a collection of over 4 million tweets. We have taken a subset of this data as the dataset provided for the MC1 challenge has only around 1 million tweets. The dataset contains the UserId, tweet text, location, and the timestamp when the tweet was created.

Data Preprocessing:

The following tasks were performed to preprocess the data-

1. Get longitude and latitude from tweet location: Original data contained User Id and location (City name). We are retrieving the Latitude and Longitude data from Nominatim (tool from Open Street Map) and saving it locally which is later used.
2. Remove the URLs and emojis from the tweet text.
3. Remove stop words from the tweet text.
4. Remove punctuation from the tweet text.
5. Convert all words to lower case words.
6. Tokenize and extract words from the tweet text.

Approach:

For the first approach, we used NLTK WordNet to get the synonyms of the symptoms/keywords entered by the user. We initially also included the hypernyms, holonyms, meronyms and hyponyms of the keywords but later found that it resulted in a lot of irrelevant words which were out of context. It was not possible to manually inspect these as the list of words changes with user input. Thus we decided to stick with only the synonyms for now. Once we had a dictionary of relevant words, we filter the tweets using this dictionary. For example, for the word "flu", synset returns the list- *influenza, grippe, flu.*

As our second approach, we used Gensim Word2Vec. We train our model on all the tweets. When the user enters the list of symptoms/keywords, we use the already trained model to find the list of top 20 most similar words in the data for each keyword. The model also returns a score of similarity between the keyword and the returned word. We use this similarity score to set the size of the words displayed in the word cloud. Lastly, we filter the tweets based on the presence or absence of the keywords and their similar words. For example, for the word "flu", Word2Vec returns the list- *swine, vaccine, seasonal, diagnosed, pandemic, h1n1, tested, influenza, etc.*

The filtered set of relevant tweets is then available for analysis. Looking at the results from our preliminary experiments for both the approaches, the Word2Vec model seems to perform better. The words returned by Word2Vec are more relevant to the context as compared to Synset which just uses the literal meanings from English dictionary to find the synonyms of the word.

Visualizations:

- **Time series:** Provides information regarding the count of tweets which contain words similar or closely related to the input search terms. The x-axis is the date of creation of the tweet and the y-axis is the count of relevant tweets for that particular day.
- **Word Cloud:** The word cloud provides a visual representation of the words most similar or closely related to the search term. The size of the word in word cloud depicts its occurrence frequency in the relevant tweets for Synset model and similarity score using the Word2Vec model.
- **Geospatial Scatter Plot:** The plot displays the origin of tweet against the background of the map of the United States of America.

  Interactions:
  - On hover over the tweet, the tweet text is displayed. *This takes time to render, please wait to see the text.*
  - The plot can be zoomed in with the use of the Box Zoom Tool to enlarge and identify the location of tweet in a particular region.
- **Widgets**:
  - Text Input : Provides space for the user to enter the list of symptoms or desired keywords.
  - Toggle Button: To choose between the 2 approaches: “Synset” and “Word2Vec”.
  - Buttons:
    1. Search Tweets button: Used to filter the tweets using the criteria selected by the user (symptoms and the approach). All the graphs are updated on click of this button. *Please wait for a while after clicking this button to see the new results.*
    2. Play/Pause buttons: to play or pause the time-series plotting in the scatter plot.
  - Slider : Used to control the rate of plotting the scatter plot (speed for play button).
  - DateTimeSlider : User can select a date range within which he/she would like to investigate the data more closely. DateTimeSlider affects the scatter plot and line plot only.

Observations:

- Preprocessing of twitter data is slightly different from normal text, as people tend to use a lot of short forms to convey words. This may be because of the limited number of characters available. Sentence structure is also different in twitter data for the same reason. The tweet may contain URLs, emojis and other special characters.
- Correlating location with user is not always accurate and might not yield relevant results. For example, in the tweet “My friend is very ill”, the speaker is talking about a third person, which may or may not be from the same location of the tweet. Hence, making an

assumption that the tweet location is the accurate location for disease spread might be erroneous. Our method is not capable of handling such scenarios. We might need to perform sophisticated analysis on the tweet to find if the speaker is talking about a third person before considering the tweet as relevant.

- User is provided with the flexibility to identify relevant tweets by searching for terms related to the topic of interest. This would be helpful in observing the trend of a particular symptom. For example, nausea may be a common symptom during the initial spread of the disease but may not be as evident later on.
- Our approach also can not correctly identify tweets where the keyword is used but in a different context. For example, in the tweet "I am sick of this job!! I need to switch!!" the word "sick" is used but the context is different. Our approach would mark this tweet as relevant because of the presence of "sick".

Bokeh Version: 0.12.10

Python Version: 3.6.0

Package Dependencies:

1. Tweet-Preprocessor: This library makes it easy to clean, parse or tokenize the tweets.
   - ○ Install using: $ pip install tweet-preprocessor
2. Geopy:
   - ○ Install using: $ pip install geopy
3. NLTK: NLP library
   - ○ Install using: $ pip install nltk
4. Gensim: To use Word2Vec
   - ○ Install using: $ pip install gensim
5. Cython:
   - ○ Install using: $ pip install cython
6. Other packages: numpy, pandas, sklearn, random

To run the code use the command: bokeh serve --show vis.py
Caution: The data takes time to upload/update. Please wait for the visualizations to render.

Google Drive Link: https://drive.google.com/open?id=0B7_M0GqgA3Puc2J6dVlNal9IeFE

References:

1. Z. Cheng, J. Caverlee, and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In Proceeding of the 19th ACM Conference on Information and Knowledge Management (CIKM), Toronto, Oct 2010