Final Report

Title: Data Mining on Twitter and Visualization

Name(s) Wenjun Zhu, Ke Zhao

Abstract

1. Introduction and motivation

Twitter is a popular social network where users can share short SMS-like messages called tweets. Users share thoughts, links and pictures on Twitter, journalists comment on live events, companies promote products and engage with customers. The list of different ways to use Twitter could be really long, and with 500 millions of tweets per day, there’s a lot of data to analyze and to play with.

In order to have access to Twitter data programmatically, we need to createan app that interacts with the Twitter API. Tweepy provides the convenient Cursor interface to iterate through different types of objects

By using Twitter API, we will collect data with foucus on Syria Striking topic. We used MySQL database process the data, then transferring data into Excel files. The next thing we did is analyzing the dataset, extracting the frequency of each word, then we will use NLP to categorize words according to sentimental analysis. We used TextBlob to do the sentimental analysis, TextBlob is a python library and offers a simple API to access its methods and perform basic NLP tasks.  Finally we visualized the data according to the result from sentimental analysis.

1. Problem definition (what you try to do in the project)

Social media app is a very popular platform for people to share their thoughts and ideas. In Twitter, there are often many breaking news such as emergency accidents, political news and etc. We are choosing the most controversy topic about US strike of Syria. We collected data from Twitter with focusing on Syria air strike. With these data, we can analyze the data of some valuable users, find out what opinions they hold towards the air strike. According to the geographic location of the specific tweets to classify the altitude of the whole state.

1. Your approach and method description

**Accessing the Data**

In order to authorise our app to access Twitter on our behalf, we need to use the OAuth interface:

import tweepy

from tweepy import OAuthHandler

consumer\_key = 'YOUR-CONSUMER-KEY'

consumer\_secret = 'YOUR-CONSUMER-SECRET'

access\_token = 'YOUR-ACCESS-TOKEN'

access\_secret = 'YOUR-ACCESS-SECRET'

auth = OAuthHandler(consumer\_key, consumer\_secret)

auth.set\_access\_token(access\_token, access\_secret)

api = tweepy.API(auth)

**Removing stop-words**

Given the nature of our data and our tokenisation, we should also be careful with all the punctuation marks and with terms like RT (used for re-tweets) and via(used to mention the original author of an article or a re-tweet), which are not in the default stop-word list.

**NLP tasks using TextBlob Sentiment Analysis**

The sentiment property returns a named tuple of the form Sentiment(polarity,subjectivity). The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective.

**WordCloud and BaseMap for Viusalization**

Matplotlib‘s main tool for this type of visualization is the Basemap toolkit, which is one of several Matplotlib toolkits which lives under the mpl\_toolkitsnamespace. Basemap is a useful tool for Python users to have in their virtual toolbelts.

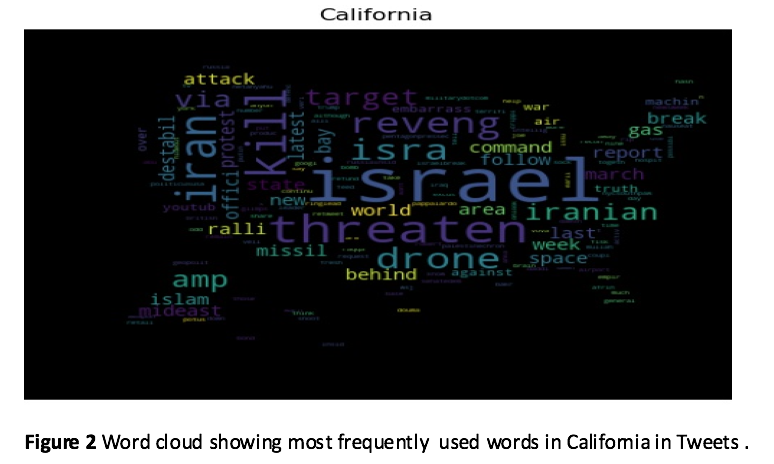
Word clouds are a very information-dense representation of the frequency of all words in a given text. Word clouds are more effective than just using bar charts displaying the counts of words for large amounts of text, as the chart would be difficult to parse if there are too many bars.

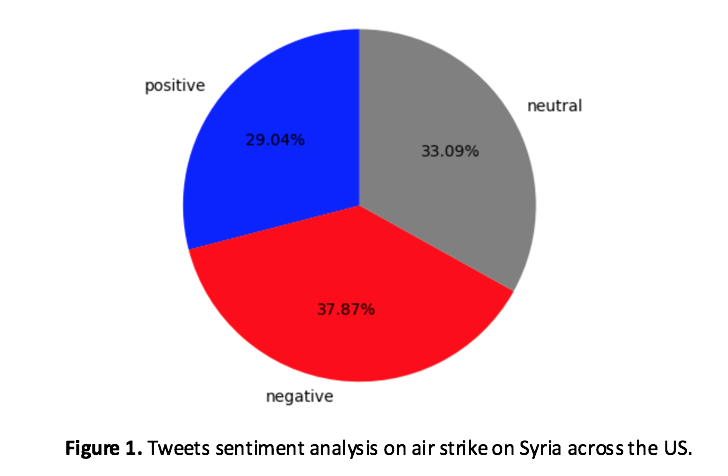
4. Results



**Chart 1.** States Opinion towards Syria air strike.

Red: oppose, Blue: support, Gray: neutral, White: no data





We successfully obtained users attitude of air strike on Syria. In the map above, the red states represent opposing opinion toward this air strike and blue states represent consensus opinion toward air strike. Besides, we conduct specific analysis of each state most frequently used words in the context of air strike on Syria, the above word cloud picture showed most frequently used words in California.

5. Conclusion

Through this project, we successfully conducted a basic data analysis using some basic sentiment analysis techniques. The NLP textblob can only return back a polarity value which is not accurate when deciding the attitude of users’ tweets. So in the future experiment, we are going to use deep network to analyze user data, this will result in a more accurate attitude of each user toward a specific event.

6. References

1. *Mining Twitter Data with Python (Part 1: Collecting data)*
2. The wordcloud library is MIT licenced, but contains DroidSansMono.ttf, a true type font by Google, that is apache licensed. The font is by no means integral, and any other font can be used by setting the font\_path variable when creating a WordCloud object.
3. Geographic Data with Basemap
4. TextBlob is a Python (2 and 3) library for processing textual data. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more. http://textblob.readthedocs.io/en/dev/