*Sentiment Analysis & Scenarios Analysis*

Sentiment and scenarios analysis can be combined to infer people’s attitude towards a specific topic or event. In this project, these analyses are conducted through three processes. To be brief, they include MapReducing to retrieve topic relevant tweets, pre-processing on these tweets and then analysing their sentiment attitudes. More technical details are presented in the followed part.

The first step, using MapReduce capabilities provided by CouchDB to collect all tweets that are relevant with a specific topic or event. These MapReduce functions could simply aim at searching whether a given term or string is appeared in the tweets. Or they could also be very complex, such as aiming to find out the most hot topics users discussing about.

The second step is pro-processing on tweets; and particularly, parsing ‘text’ in the tweet. Since natural language is significant complicated for a machine to understand and process, certain necessary pre-processing on ‘text’ are indispensable. In this project, this function is implemented in ‘TextParser.py’. The ‘TextParser.py’ is responsible for the following tasks.

1. Delete meaningless words appearing in ‘text’. The meaningless word means that this word is not useful and helpful to determine the sentiment of a ‘text’. In this project, these words are named as ‘stop words’. Stop words are various. For instance, the majority of them could be a pronoun (e.g. ‘she’, ‘they’…), a preposition (e.g. ‘about’, ‘to‘…). Some of them could be an adverb, such as ‘how’, ‘ever’ and so on. In this project, all the stop words are collected manually and stored in a data file named ‘stop\_words.txt’.
2. A special scenario is also been considered that some people would like to repeat certain characters in a string to express their emphasizing. (e.g. instead of using ‘bad’, people might use ‘baddddddddd’ to address their feelings.) In this case, deleting those repeated characters is necessary.
3. Deleting mentioned tweeters(@), hashtags(#) and URLs in ‘text’. Since the mentioned tweeters and hashtags show only very limited help in sentiment analysis, they are deleted in the implementation. Cause basically, they are just nouns of topics or names which could used to detect what this tweet is talking about. However, the most general way to obtain topic relevant tweets is by searching whether a certain term or string of this topic appears in the ‘text’ using MapReduce. Thus, these tweets are already associated with the topic, which make mentioned tweeters and hashtags is no longer that useful.
4. After the processes above, a new ‘text’ is generated and ready to be tagged its sentiment.

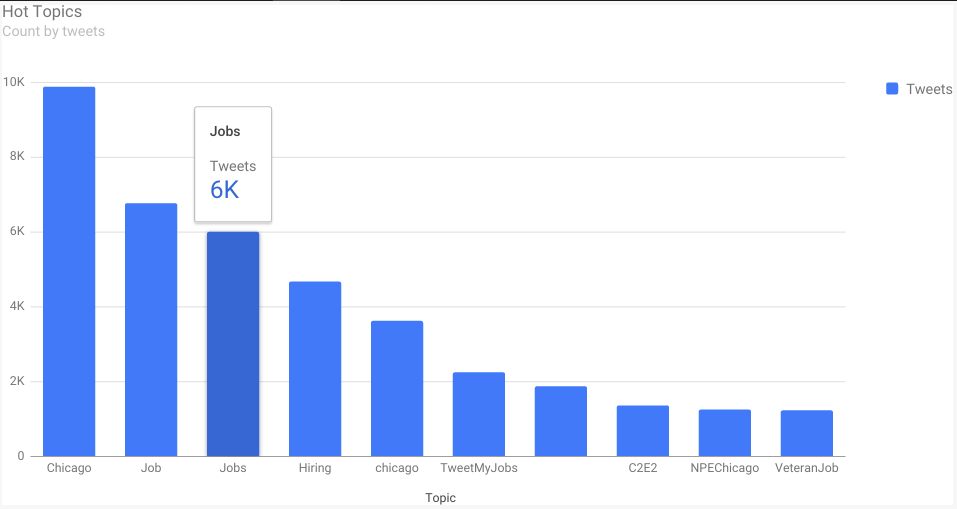
Therefore, the last step is tagging the sentiment field based on analysing ‘text’ of each tweet. In this project, TextBlob is used as the tool to tag each tweet’s sentiment. TextBlob is a Python library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as noun phrase extraction, sentiment analysis etc. [1] Both the build-in classifiers ‘PatternAnalyzer’ and ‘NaiveBayesAnalyzer’ are experimented in this project. Because of ‘PatternAnalyzer’ showing more accuracy, it is chosen as the final classifier to do sentiment analysis.

Using this analysis method, tweets retrieved from the database will be add another filed named ‘sentiment’, with the values of its sentiment classification (positive/negative/neutral) and sentiment score (where [-1,0) represents negative, 0 represents neutral, and (0, 1] represents positive).

The source code of all the processes above is under the directory of ‘sentiment’.

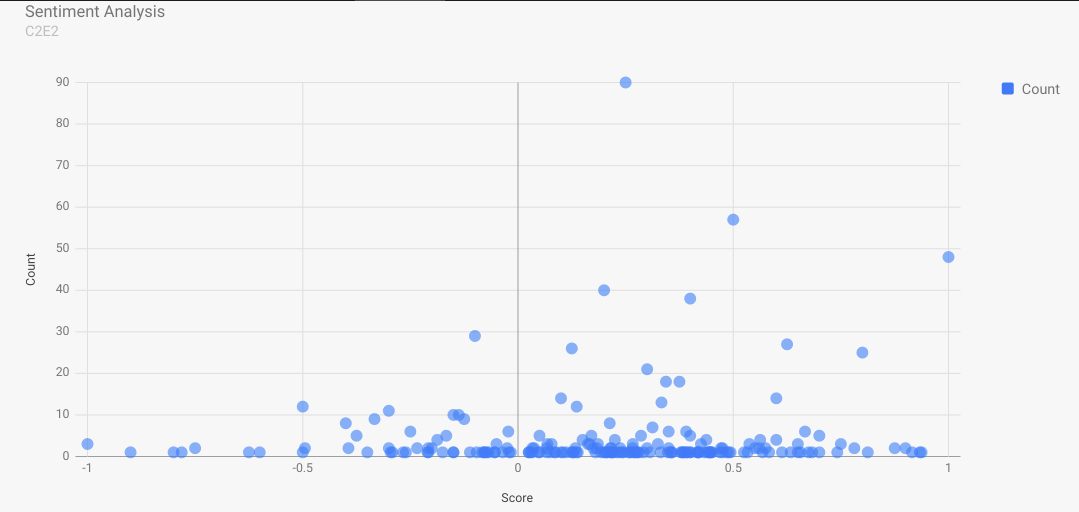
*Scenarios Analysis 1: Hot Topic*

The hot topics are statistical hashtags in tweets that users mentioned most frequently. Figure 1 illustrates the most popular 10 topics in Chicago. It shows that three topics (They are Job, Jobs, and Hiring respectively) are highly associated with finding jobs among the top five ones. This phenomenon could probably reflect a fact that the employment market in Chicago is still weak and certain amount of people are experiencing the difficulties to find a job.



*Scenarios Analysis 2: Topic of C2E2*

C2E2 is an example of a specific topic chosen from the hot topics above. C2E2 is the abbreviation of Chicago Comic & Entertainment Expo. Figure 2 shows the sentiment distribution of tweeters regarding C2E2. Overall, it can be seen that people in Chicago hold more positive attitudes towards C2E2 than negative one, which expresses that they love Comic and the Comic-relevant things.

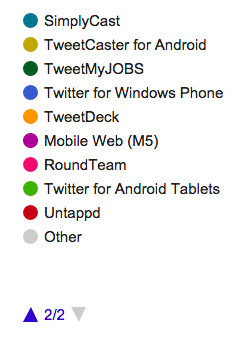
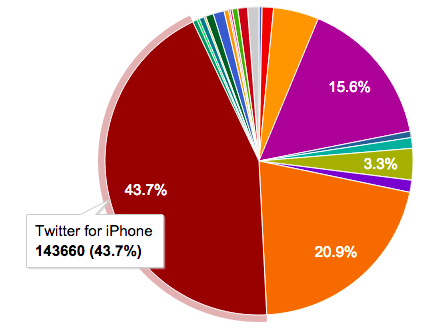


Based on the overall sentiment results, a further analysis is conducted to see what specific opinions hold by users to C2E2. This function is implemented in ‘text\_analysis.py’ and the result is shown below. The result presents that lots of users have feelings such as ‘thanks’, ‘love’, ‘great’, ‘best’ to C2E2.

[(u'c2e2', 694), (u'chicago', 219), (u'mccormick', 189), (u'comic', 114), (u'table', 90), (u'entertainment', 67), (u'expo', 63), (u'panel', 60), (u'thanks', 59), (u'time', 54), (u'booth', 51), (u'love', 44), (u'move', 44), (u'great', 43), (u'best', 41), (u'convention', 38), (u'trooper', 35), (u'storm', 35), (u'clone', 34), (u'chicks', 34)]

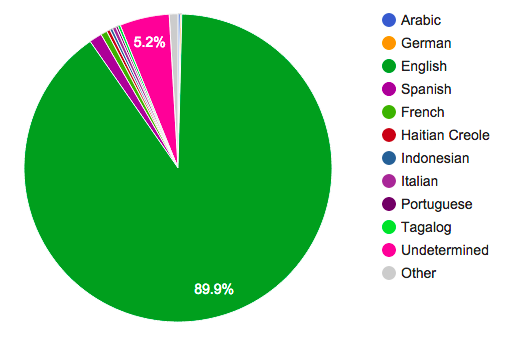
*Scenarios Analysis 3: Device Distribution*

As the development of industry technology, devices used to access Internet are more and more various. In aspect of product popularity or other analysis, it is significant to detect what devices are used most by people. Figure 2 demonstrates the details of the distribution of devices people used to post tweets. As shown in the figure, iPhone is the most popular device with 43.7% users using it, following by android device that owns 20.9% users. An interesting phenomenon that only 3.3% Twitter users post tweets from the Twitter Web Client can be found from this figure. Compared with the percentages of either Twitter for iPhone or Twitter for Android, this number could be an evidence that mobile device is the overwhelming tendency in the future’s IT industry development.



*Scenarios Analysis 4: Language Distribution*

Since Chicago is a famous metropolis, people from around the globe bring their cultures here and throw them into this big pot. Thus, an analysis based on the language distribution is also conducted to see what kinds of language are daily used by people in Chicago. Figure 2 shows the language distribution of those tweets. Instead of English, the most widely used language is Italian (1.3%), then followed by French (0.7%), which means lots of Italian and French are living in Chicago. However, there are 5.2% users cannot be determined using which language. The reason could lies in that these users probably are Chinese, Japanese, Korean or other foreigners that their languages are dramatically different with English and therefore are difficult to be detected and recognized.



[1] <http://textblob.readthedocs.org/en/dev/>