Sentiment Analysis of the Top Trends of the week around Chicago in Twitter

Done by,

Shivakumar Vinayagam
A20341139
svinayag@hawk.iit.edu

Problem Statement

* Analyze a week of tweets around chicago from twitter and get the top 5 most trending topics based on the hash tag analysis of the tweets.
* On the Highest trending topics do a lexicon based sentiment analysis .
* Also train a classifier on the week of tweets to help predict future tweets.
* Determine the Best settings for the Classifier so that it is not too biased to the data and has some decent accuracy of prediction.

Approach

* First collect a week of twitter data using the Twitter REST API ,curate it into tokens, hash tags and other details for each tweet and store it in file format.
* Use a sample amount of tweets to find the optimal settings for the vectorizer and Logistic Regression Classifier.
* Use SentiWordNet to classify the tweet as either positive(+1), negative (-1) or neutral(0) , since it gives more details than AFINN.
* Then use the hash tags to get the top 5 most trending topics in the dataset collected.
* Vectorize the dataset and train the classifier without the tweets from the top trends and based on this predict the top trend tweets.

Data collection

* The data will be collected from twitter using Twitter Search API which is a REST API.
* Used the Location tag for chicago to search the tweet and the tweet id and time of it to collect distinct tweets of one week.
* About 22k tweets were collected from around chicago, then curated and stored
* Each record is for one tweet and contain a list of tweet(text), hash tags, mentions and urls each.
* Will tokenize each tweet and also collect the hash tags separately and database it.

Problems Faced

* The twitter search api sent only the recent tweets so used tweet id and date to get distinct tweets of one week time.
* The various setting of the vectorizer were experimented over to find the optimal values for n_fold, min_dt, max_dt, tokenize function and binary.
* Also trained and tested on entire dataset to estimate the full accuracy possible.
* Tried using tokens to find the top trends in dataset but found too much of noise to get proper results. Hence used Hashtags to get the top trends.
* Also used the top trends and exclude them from the dataset to train on remaining and predict on the excluded trends.

Results

* The optimal setting found for the vectorizer are 5 folds, min_dt=2, max_dt=0.5, tokenize func=tokenize(w/o punc) and binary as true.
* Also trained and predicted on entire dataset itself to get vectorization of 21422 tweets with 11441 terms and an accuracy of 0.959667562943.
* Result when using tokens of tweets to find top trends where
    1. The Term chicago has count 10220.0
    2. The Term il has count 5147.0
    3. The Term in has count 4990.0
    4. The Term the has count 4979.0
    5. The Term job has count 3846.0

    Hence decide to not use the tokens to get top trends.
* Thus used Hashtags to get the top trends and found

    [(u'Chicago', 4093), (u'Hiring', 2558), (u'job', 1922), (u'Job', 1900), (u'CareerArc', 1747), (u'Jobs', 1696), (u'chicago', 777), (u'hiring', 773), (u'ncaa', 531), (u'Hospitality', 512)] as the top 10 trends.

    Tried changing the value of equalize the tags “job” and “Job” using lowercase but found it confounded tags like “CareerArc”.
* Also used the top trends and excluded them from the dataset to train on remaining and predict on the excluded top trends. Got an average accuracy of 0.815919242915637 for the top 5 trends.

Conclusion

* Used REST API to get tweet around Chicago of 1 week period.
* Used the optimal setting on vectorizer and got an accuracy of 0.7404999999999994.
* Used Hashtags instead of tokens to get top trends and got “Chicago” as top. This is expected as search tags using location around a city includes geotagged tweets and the city tagged tweets mainly.
* Didn’t want to bias the classifier too much towards the dataset and hence calculated the accuracy excluding the top trends to predict the top trends and got a value of 0.815919242915637