Bairi Sandhya Rani
14CO205
(+91) 9035769725
sandhya.bairi154@gmail.com

Interpreting the Public Sentiment Variations on Twitter

Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, Chun Chen, and Xiaofei He

WORK FOR IMPLEMENTATION

- In this paper, twitter sentiment variation is tracked and analyzed to determine the possible genuine reason behind particular sentiment with the help of emerging topics.
- I would like to perform the sentiment analysis for a chosen target and then implement the two Latent Dirichlet Allocation models described in the paper namely Foreground and Background LDA, Reason Candidate and Background LDA.
- Python language is used for implementation.

PLAN OF IMPLEMENTATION

- First standard LDA need to be implemented. For this purpose, initial corpus i.e., set of documents can be obtained.
- With regard to the paper, these documents are the tweets extracted from the twitter and then preprocessed. Then sentiment analysis is done to assign labels to each individual tweet.
- These are then used as corpus for analyzing sentiment variation. For the first evaluation, an existing corpus is used.
- Corpus used
  - A text corpus is a large, structured collection of texts. NLTK library comes with many corpora.
  - Reuters Corpus, nltk.corpus.reuters or Brown Corpus, nltk.corpus.brown are used.
  - Reuters Corpus contains 10,788 news documents totalling 1.3 million words. The documents are classified into 90 topics and grouped into two sets, called “training” and “test”.
  - Categories in Reuters corpus overlap with each other because a news story often covers multiple topics.
  - Brown Corpus contains text from 500 sources and the sources have been categorized by genre, such as news, editorial, and so on.
- Sentiment Polarity Datasets for training and testing
  - includes 1000 positive and 1000 negative processed reviews. Coutesy of Pang & Lee.
- Python libraries needed :
  - numpy
    - ✔ consists of multidimensional array objects and a collection of routines for processing those arrays.
  - nltk
    - ✔ Natural Language Toolkit is a leading platform for building python programs to work with human language data with easy-to-use interfaces and over 50 corpora, used for classification, tokenization, stemming, tagging etc.
    - ✔ corpus used – stopwords, wordnet, Reuters, Brown
  - tweepy
    - ✔ easy-to-use Python library for accessing the Twitter API.
  - textblob
    - ✔ library for processing textual data
    - ✔ provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation etc.
  - Scikit-learn, pandas
    - ✔ efficient tools for data mining and data analysis
  - json

- ✔ parses JSON from strings or files, JSON into a Python dictionary or list, converts Python dictionaries or lists into JSON strings etc.
- Others like sys, re, random, OptionParser

- Parameters of LDA:
  - num_topics – specify many topics you would like to extract from the documents
  - corpus file name
  - iteration count
  - alpha – document-topic density; greater this value, the document will be assigned to more topics and viceversa
  - beta – topic-word density; greater this value, each topic will contain more words and viceversa
  - threshold of document frequency to cut words
- Data preprocessing
  - stopwords : remove general words
  - punctuation : remove punctuations
  - lemmatize : reduce related forms of a word to a common base
- Data preparation is not simple and only a list of documents is needed. The shorter each document, lesser the time to complete a topic model.
- The standard LDA is then modified to form the FB-LDA since for the background tweets set, FB-LDA follows a similar generative process with the standard LDA.
- Additional functionality is required for foreground tweets set since each tweet has two topic distributions namely foreground topic distribution and background topic distribution.

WORK FOR FIRST EVALUATION

- Implementation of standard LDA on Reuters corpus or Brown corpus.
- Extracting tweets from twitter, preprocessing and assigning sentiment label to each individual tweet based on classifiers like Support Vector Machine, Naive Bayes theorem or Maximum Entropy.

WORK FOR FINAL EVALUATION

- Implementation of FB-LDA and RCB-LDA from the standard LDA
- Corpus used will be the actual tweets extracted and preprocessed with assigned labels (positive, negative, neutral etc.)

REFERENCES

- Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.
- http://www.cs.cornell.edu/people/pabo/movie-review-data/