Analysis of Current Approaches in Topic Modeling for Twitter Data

Brian Gillespie
Northeastern University
Seattle, WA
bng1290@gmail.com

Shailly Saxena
Northeastern University
Seattle, WA
saxena.sha@husky.neu.edu

ABSTRACT

Twitter is a dynamic source of text data, reflecting the changing world and how it is viewed by those who in inhabit it. Proper analysis of this data can yield very interesting insights into public opinion and the popular sentiment towards events as they unfold in real time. The ability to see which events matter to people as they emerge is highly desirable, however topic modeling of Twitter data remains somewhat elusive due to the nature of Tweets themselves. Non-standard vocabulary, limited character length, and the wide variety of tweet content often make it difficult to extract meaningful, generalized topics from the data.

In this paper, we investigate two of the more recent approaches to topic modeling of Twitter data, and compare their effectiveness against a basic TF-IDF model of raw Tweets as a base reference. In the first approach, user profiles are generated by aggregating Tweets per user. This aims to leverage the consistency of each user's Tweets, thus increasing the relevance of words within each topic. The second method attempts to modify Latent Dirichlet Allocation (LDA) such that the distribution of topics for each user is considered alongside the distribution of topics for the entire Twitter dataset.

Keywords

Twitter, LDA, topic modeling, NLP, TF-IDF, social media

1. INTRODUCTION

Twitter is an excellent medium for the development of hot topics, and these topics often gain a lot of momentum as they pan out to reach a wider variety of people. Analysis of these trending topics as they develop and impact users around the world is a popular area of research in text analysis. Many times, Twitter has been a nexus for people from all over to come together and offer sympathies during tragic events such as the terrorist attacks in Brussels. Users often rally behind simple messages, and weigh in their opinions, popular or unpopular, facilitating global discourse on a brand new scale. As a whole, the variety of information contained within Twitter is loaded with the random musings of over 305 million active users. This is a great obstacle that must be overcome in order to make an generalizations about the data. To be able to distill essential topics is very desirable as business attempt to cater to an increasingly diverse and dynamic consumer base. Such knowledge can also assist with attracting new markets, and to gain a better understanding of their own place in the minds of the general population.

There has been much research in the structure and dynamics of microblogging networks, and many insights have been gained by looking into how users interact with one another and with the world. Attempts to glean detailed information from the actual contents of Tweets however, has met with limited success. Recent developments in topic-modeling approaches to Twitter data such as Hong and Davison[1] and Zhao, Wayne Xin, et al.[3] have further refined traditional topic-modeling techniques so that they better cater to the specific structure of Twitter data. In this paper, we explore some of the more promising approaches to topic modeling in tweets, and investigate their implications to further research into developing a more successful Twitter topic model.

In [1], various preprocessing methods were employed to increase the accuracy of Latent Dirichlet Allocation(LDA) applied to a series of tweets from September 2009 to January 2010. Tweets were aggregated by the User IDs of their authors, and a new set of documents was created where each document is a *user profile* of a unique user id and his or her combined set of tweets. Hong and Davison saw a significant improvement in the <metric> of their topic models for the user profile approach, however the other alternative approaches employed did not appear to improve <metric>. Thus we will employ and evaluate the user profile aggregation in this paper, to further investigate its effectiveness.

In [3], researchers attempted to modify LDA so as to adapt the algorithm to model Twitter data more closely. By not only considering the distribution of total topics in the data, topics were also drawn for each user, based on the understanding that each user's set of interests are relevant to how their tweets are composed.

2. DATASET AND DATA PREPROCESSING

The datasets used were obtained from Cheng, Caverlee, and Lee [2]. In order to reduce noise in the text data, a set of stop words were removed from the tweet bodies, the remaining words were then stemmed, and any hyperlinks were also removed. A bag of words model was created, and TF-IDF vectors were generated. For the *user profile* LDA approach [1] and the Twitter-LDA model [3], the vectors were aggregated based on user IDs. We also processed each tweet for the documents ”as-is” for the Twitter-LDA model. The training and tests datasets were merged, and then repartitioned to prepare for 10-fold cross-validation.

2.1 The Dataset

The training dataset from Cheng, Caverlee, and Lee [2], contains 3,844,612 tweets from 115,886 Twitter users over the time period of September 2009 to January 2010. The test set, also from [2], contains 5,156,047 tweets from 5,136 Twitter users over the same time period. Each line of the dataset contains a unique user ID and tweet ID, text content, and a timestamp. The text content of each tweet is limited to 140 characters, and can contain references to other users of the form @username, as well as popular #hashtags. Many tweets also include hyperlinks which are often passed through URL shorteners (e.g. http://goo.gl/uLLAe). The implications of these more anomalous text instances are considered in the next section.

2.2 Data Preprocessing and Reduction

Due to the unconventional vocabulary in tweets and the large amount of noise inherent in such a diverse set of text, we decided to perform several data reduction steps. A regular expression was used to remove any odd characters and any URL links, to reduce noise in the data. After pruning the tweet contents, the text was tokenized, stop words were removed, and the remaining tokens were stemmed using the Porter Stemming library from the Natural Language Toolkit (NLTK). The prepared corpus of tweets was then converted to a set of Term Frequency (TF) vectors, using the corpora library from gensim. Gensim also provided a map of words to their hashed keys so that the LDA results could be interpreted.

An additional document corpus was generated from each dataset by aggregating tweets by user ID. This reduction is referred to as the *user profile* model [1]. After aggregation, the corpus contained a total of <115,886> documents. The largest *user profile* contained (after preprocessing) <NUMBER> tokens, and the smallest contained <NUMBER> tokens. To further reduce the data, any profile containing <THRESHOLD> words was removed. After Term Frequency hashing was applied, words occurring less than <NUMBER> times were removed and more than <NUMBER> were removed[2].

3. REFERENCES

[1] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88. ACM, 2010.

[2] J. C. Z. Cheng 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)*, October 2010.

[3] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011.