Latent Dirichlet Allocation as a Twitter Hashtag Recommender System *

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

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

ABSTRACT

In this paper, we investigate an unsupervised approach to hashtag recommendation, to aid in the classification of tweets. Two corpora were generated from a sample of tweets collected from September 2009 to January 2010. The first corpus was composed simply of individual tweets, and the second corpus was made by aggregating tweets by UserID into what is referred to as the USER PROFILE model. Latent Dirichlet Allocation(LDA) was used to generate topic distributions for each corpus, and then Collapsed Gibbs Sampling was used to generate topic distributions for new test tweets. By sampling a topic from each test tweet, and then sampling one of the top terms from that topic, a set of words can be selected as recommended hashtags. This recommendation process was applied to several example tweets, and the relevancy of the suggested hashtags were evaluated by human observers. The results of this would be briefly mentioned here if there were any.

Keywords

Twitter, LDA, topic modeling, NLP, social media

1. INTRODUCTION

Twitter is an amazing source of text data, both in content and quantity. With over 305 million monthly active users across the globe, there are a wide variety of subjects being discussed. Cataloging this data, and finding ways to make it more search-able is very desirable, as this data can be an effective resource for business knowledge and studying social trends. Hash-tags appears as the natural categorization index for tweets, however since only 8% of tweets contain a hashtags, they cannot be used as a direct categorization for all tweets. Compounding this issue, there are no prescribed hashtags; any sequence of characters can be a hashtag as long as it has a # in front of it. A hashtag recommendation system can be implemented to encourage users to utilize more hashtags, and to provide a more consolidated hashtag base for tweet categorization. But in order to avoid prescribing hashtags, and thus limiting the free expression of Twitter users, it is more reasonable to generate the suggested hashtags from the content of Twitter itself. This has the added gain of providing an adaptable system that can develop alongside the vocabulary and interests of the Twitter user base.

*There are no legal restrictions! Break all the rules! We don't own this, and wouldn't even dare claim it if our lives depended on it. Plagiarize to your heart's desire, but maybe hook us up with a job interview if you think we had some good ideas in here.

2. RELATED WORK

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 the networks however, has met with limited success. Recent developments in topic-modeling approaches to Twitter data such as Hong and Davison[2] and Zhao, Wayne Xin, et al.[4] 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 [2], various preprocessing methods were employed to increase the accuracy of Latent Dirichlet Allocation(LDA) applied to a series of tweets from the first and second weeks of November 2009. 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 Precision, Recall, and F1 scores of their topic models for the user profile approach, however the other alternative approaches employed did not appear to improve these same metrics as significantly. Due to the success of the *user profile* approach, we will employ and evaluate this method, to further investigate its effectiveness.

In [4], 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.

Table 1: Twitter Sample Sep 2009 to Jan 2010

UserID	Tweet
66	For all the everything you've brought to my internet life #mathowielove
53717680	What's good Tweeps this Friday
16729208	"I bet you all thought Bert & Ernie were just roommates!"
86114354	@miakhyrra take me with you! lol

3. DATASET AND PREPROCESSING

The datasets used were obtained from Cheng, Caverlee, and Lee [3]. 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 [2] and the Twitter-LDA model [4], 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 re-partitioned to prepare for 10-fold cross-validation.

3.1 The Dataset

The training dataset from Cheng, Caverlee, and Lee [3], 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 [3], contains 5,156,047 tweets from 5,136 Twitter users over the same time period. In general, tweets contain between 1 and 20 words, with an average of 7 words per tweet. A smaller proportion of the tweets contained more than 20 words as seen in Figure 1. 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.

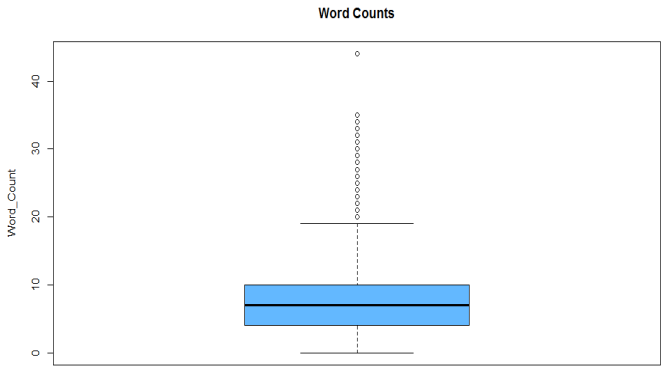

Figure 1: Boxplot for number of words in a tweet.

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 [2]. After aggregation, the corpus contained a total of 115,886 documents. Data preprocessing was performed on both of these corpora, and is described in the next section.

3.2 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 preprocessing steps. A regular expression was used to remove any non-latin characters and any URL links, as URL links are generated randomly and cannot be easily related to a topic. 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.

We noticed that many tweets contained excessive repetition of words (for instance one tweet read: "@jester eh eh eh eh eh eh eh eh eh..."), so in order to reduce any bias towards overused words in the data we removed words appearing over 5 times in a tweet, as well as using the TF-IDF model to reduce emphasis on such words. To reduce bias to common words across the corpus, we removed terms appearing in over 70% of the documents as per [4]. As seen in Figure 2, before preprocessing the data was dominated by common words that do not confer much meaning (e.g get, u, just, one). After preprocessing, however, we begin to see more meaningful words (e.g people, twitter, home, blog in Figure 3). After preprocessing, there were 26,542,006 total words and 184,002 unique words in the vocabulary.

4. MODEL AND EVALUATION

Latent Dirichlet Allocation, first proposed by Blei, Ng, and Jordan [1] is a generative probabilistic model, that seeks to generate mixtures of topics drawn from some distribution of topic probabilities. Each word appearing in these topics is itself probabilistically drawn from some mixture of these topics. This multilevel approach to topic generation has met with significant success in the analysis of text data, and is the subject of focused study in data mining of microblogs such as Twitter and Facebook.

Talk about the why and how of LDA. Then maybe a brief intro into the math, throw in a figure from the paper maybe?

We are generating several different topic models in this work, and so we decided to see if the various approaches had any significant effect on the topic distributions we made. To do this, we calculated the Jensen-Shannon (JS) divergence for each topic distribution. TALK ABOUT JS DIV.

5. REFERENCES

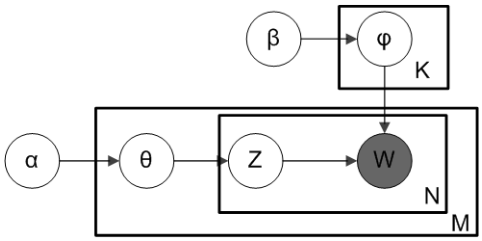

Figure 2: Plate Diagram of LDA.

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] 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.

[3] 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.

[4] 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.

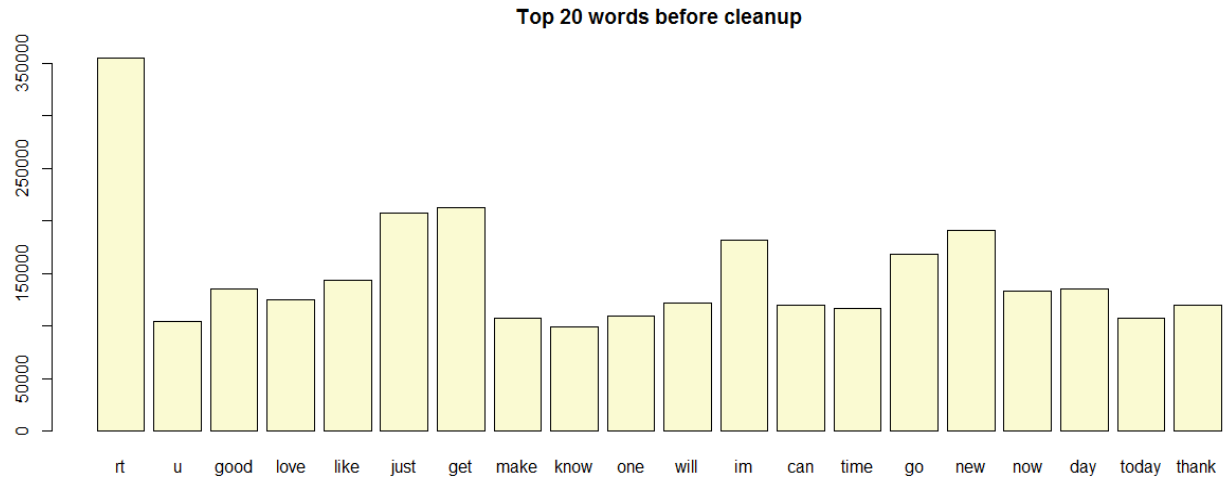

Figure 3: Original barchart of the top 20 words and their counts.

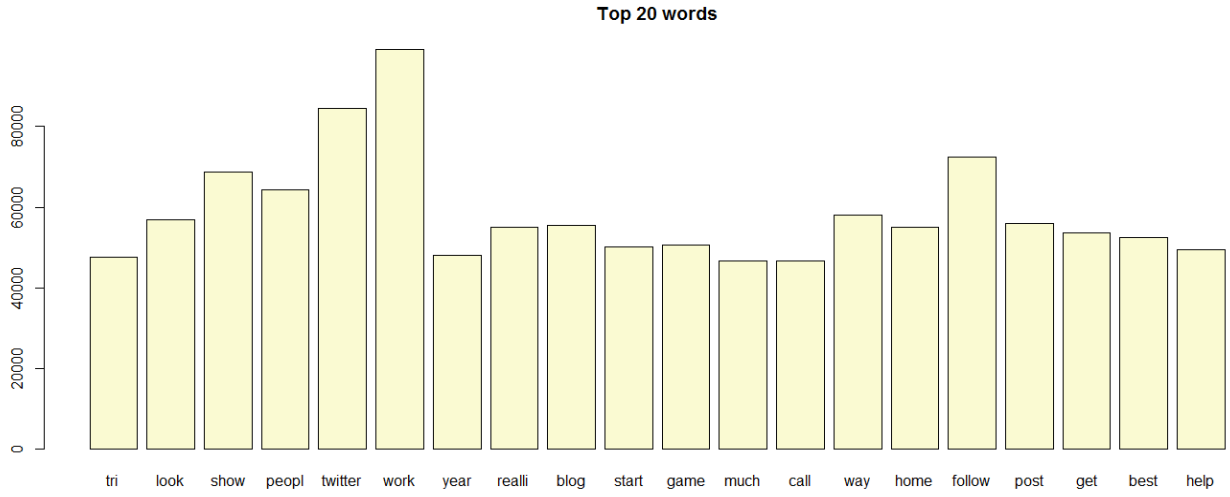

Figure 4: Barchart of the top 20 words and their counts after preprocessing.