Twitter Hashtag Predictor

Armando Banuelos, Victor Cheruiyot, Nick Tantivasadakarn

Abstract - Tweets from social networks such as Twitter and Instagram carry meaningful text that can be preprocessed using traditional NLP techniques. Users seeking to tag their tweets with the correct set of twitter hashtags can find it difficult to match their tweets with the corresponding tags. The goal of this project is to provide the user with a set of recommended twitter hashtags given some user tweet. Our model[1] uses Latent Dirichlet Allocation to identify topics of new hashtags and suggests the most relevant words in the topic were chosen as hashtag suggestions.

Keywords - Hashtag, Latent Dirichlet Allocation, Topic modelling, Twitter

1. Introduction

Given Twitter's 41.7 million user profiles, 1.47 billion social relations, 4,262 trending topics, and 106 million tweets (Kwak et al. 2010), their platform (and many others) is a playground for data scientists and artificial intelligence enthusiasts to train and test different machine learning methods. The driving motivation behind this project is to find ways to classify tweets into hierarchical social networks to enable tagging and labeling of tweets. Hashtags have become a very common method of linking ideas and thought around the world. With a structured system of hashtag selection, people will find it easy to connect their content, with content already existing in the media. This will enable Twitter better personalize their platforms to their users. In light of the Stoneman Douglas High School shooting, organizing twitter data will also help notify authorities of online threats so they can premeditatedly combat these issues.

According to Kwak et al.(2010), the Twitter platform allows users to view tweets posted by the people they "follow". To "follow" simply means to subscribe to see content posted by some other Twitter user. A user can either post their own tweets or refer to other users' tweets, a process called retweeting. This platform has a well defined markup. For example, retweeting is denoted by RT followed by @ and the address of the source, end of message is denoted by hash tags, while keywords are denoted by \# at the end of the message. Topological studies on the graph of has been done by Kwak et al.(2010) and Myers et al.(2014), which has shown to be

[1]Repository: https://github.com/vcheruiyot/cs224u-project

similar to other social media networks except that "following" in Twitter is not always reciprocal.

2. Data

The data we are using is a Twitter snapshot from 2010-2011, which was captured by Prof. Chris Potts using the now defunct TwapperKeeper service. The data has 44 million tweets grouped into 36 main topics. Despite not using current Twitter data, we believe the same methods would still apply when analyzing current Twitter data.

3. Related work

Previous work such as Goldin et al. (2013) has used LDA to first predict the topic of a specific tweet, and sample words from the dominant topic as the hashtag recommendation. This method is the most straightforward and is most similar to our method. Other approaches makes use of deep learning. Weston et al. (2014) uses a lookup table to convert tweets into a form readable by a convolutional neural networks. The networks is then fed into another look up table that create the final hashtag recommendation. Liang and Shu (2017) is similar, but used uses LSTM to process the individual characters instead of whole words. There are also methods that were not created specifically for hashtag recommendation, but may be a potential algorithm such as using VSM models to perform entity extraction (Dani et al. 2015), and semi-supervised graph methods such as Bhagat et al. (2011) and Zhou et al. (2004) which can infer labels on a graph based on seed labels.

4. Method

4.1 Preprocessing

Twitter conversations are rather colloquial, and is written such that it can pass a lot of information with in the limited space. Thus, it often comes with a lot of slang, markup short hands such as ‘RT’ for re-tweets, website urls, which will hinder the performance of our classifier. To solve this, we employ the following preprocessing scheme.

1. Setting all text to lowercase.
2. Remove hashtags.
3. Remove “RT” (standing for re-tweets).

4. Remove all user names (@ followed by a name).
5. Remove all urls.
6. Formatting basic slangs such as “coz” to “because.”
7. Text segmentation for phrases written without spaces.
8. Combining frequent co-occurring such as “New York” using a bigram classifier

4.2 Topic modeling -Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative statistical model that aims to find general topics within a large collection of documents. For our task, a document is a tweet. Each document is represented as a mixture of several topics and each word is assigned probabilities of being in a particular topic. The weight of each word will be used to classify the topic of a new document or tweet.

Since LDA is an unsupervised method, it does not know how many topics are there in the documents, and different numbers must be tested.

4.3 Choosing a suitable number of topics (*K*)

In order to find the optimal number of topics (*K*), we used a simplified grid search that iterated over a range of 2 to 70 topics. We used Python’s gensim Mallet implementation to run the LDA algorithm over this range, and selected a model based on the topic coherence score. This score measures how coherent the chosen topics are for a model. The higher the coherence score the better the model.

Although the original file has 36 different topics, each topic might be subdivided into smaller topics or are part of a larger group. After testing out different models with different number of topics (Fig 1). The best model is at 22 topics (Fig 2)

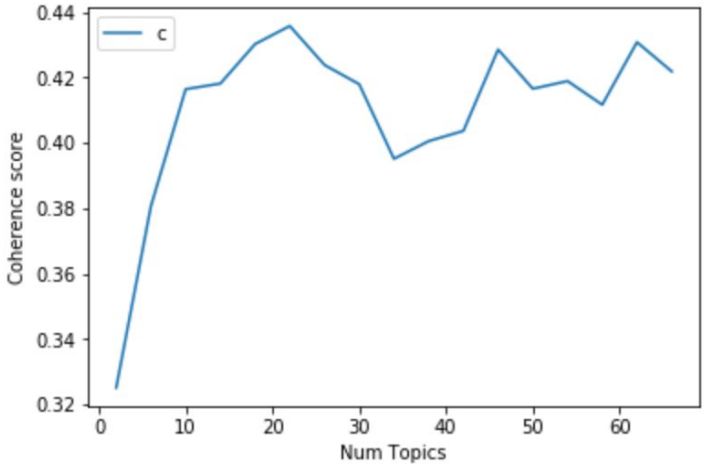

Fig 1: Coreherece scores of models with different number of topics

4.4 LDA Algorithm

LDA assumes that there exists a topic model with K topics. For each document m, containing N_m words, it has a multinomial topic distribution ϑ_m over these K topics. By using Gibbs sampling, a topic z_i of word w_i, conditioned on the used words $\sim w$ of the model and the topic-word distribution $\sim z_{\neg i}$, can be predicted as:

$$p(z_i = k \mid \sim z_{\neg i}, \sim w) \propto (n^{(t)}_{k,\neg i} + \beta t) / (\sum_{t=1}^{V} [n^{(t)}_{k,\neg i} + \beta t]) * (n^{(k)}_{m,\neg i} + \alpha_k) / (\sum_{k=1}^{K} [n^{(k)}_{m,\neg i} + \alpha_k])$$

where $n^{(t)}_k$ denotes the topic-word count of topic k and word t, and $n^{(k)}_m$ the tweet-topic distribution (Godin et al. 2013).

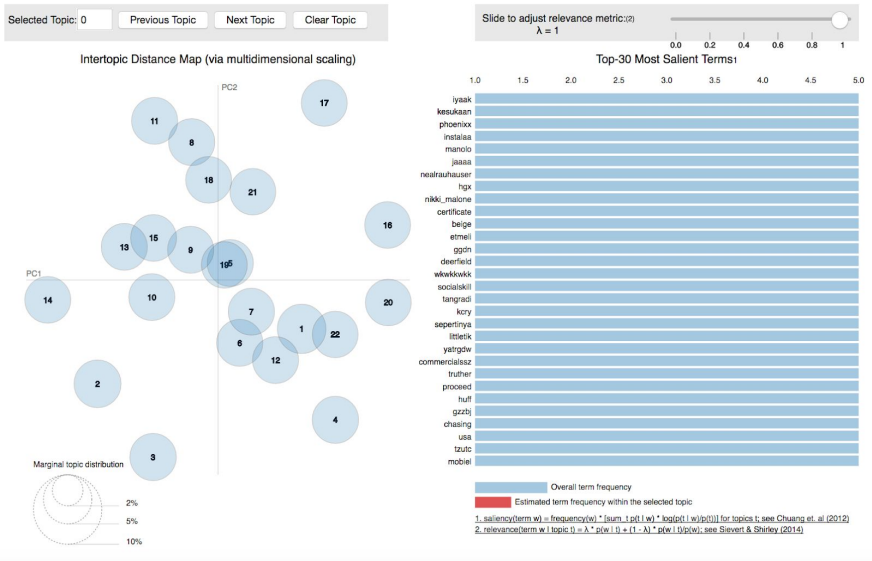

Fig 2: Visualization of the keywords clustered over the K chosen topics using pyLDAvis package

4.5 Extracting Hashtags

Using the best model from the previous step, we determine the topic of a new tweet. Then, we draw out the top keywords from the topic with the highest weight as shown in the example below.

"Thinking about making a serious stand for your future and your income for 2010."

'0.083*"**future**" + 0.042*"lose" + 0.031*"du" + 0.031*"fun" + 0.031*"miss" + '
'0.031*"ger" + 0.031*"sumpah" + 0.031*"good" + 0.031*"lady" + 0.021*"nap"'

5. Results

Our method of evaluation is calculating the recall of the hash tags given the hashtags of a held out test set. Which reaches a score of 0.13 . This method is chosen over other metrics such as, precision and F_1 because it is impossible calculate false positives, i.e. when a recommended hashtag is irrelevant to a tweet, given the current data set. The fact that is project only suggests hashtags also make irrelevant tweets less of a problem for our purposes.

Past work such as Goldin et al. (2013) uses subjective evaluations. This method is the most ideal metric for our task, but is impractical given the time frame and available resources of our project.

6. Discussion

The recall of our algorithm are rather poor. We suspect that it is because a lot of hashtags are combinations of many words such as “#handsoff.” which our algorithm does not consider. The algorithm might also have suggested a relevant hashtag in slightly different words, and does not increase the recall.

7. Future work

In our initial plans for the project, we aimed to use convolutional neural networks to suggest new hashtags. The model would capitalize the structured nature of twitter graphs to inferrelations from previous tweets to the new one in order to help narrow down the relevant topics. The LDA model was intended only as a baseline comparison to the convolutional neural network as there are previous works that have done similar methods (Godin et al. 2013). However, we have gone through major circumstances that have hindered our ability to work on the project. Thus, the LDA baseline became our only possible implementation.

Other possible improvements to our method include making use of Sparse Graph methods proposed by (Nallapati et al. 2007). This method uses the topic assignments from LDA to help capture the relationship between words in a text. The method will help disambiguate word such as “bank” that can either mean a monetary institution or a river bank.

References

Bhagat, Smriti, Graham Cormode, and S. Muthukrishnan. "Node classification in social networks." In Social network data analytics, pp. 115-148. Springer, Boston, MA, 2011.

Godin, Fréderic, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. "Using topic models for twitter hashtag recommendation." In Proceedings of the 22nd International Conference on World Wide Web, pp. 593-596. ACM, 2013.

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. "What is Twitter, a social network or a news media?" In Proceedings of the 19th international conference on World wide web, pp. 591-600. ACM, 2010.

Li, Yang, Ting Liu, Jing Jiang, and Liang Zhang. "Hashtag recommendation with topical attention-based LSTM." Coling, 2016.

Liang, Davis, and Yan Shu. "Deep Automated Multit-task Learning." arXiv preprint arXiv:1709.05554 (2017).

Myers, Seth A., Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. "Information network or social network?: the structure of the twitter follow graph." In Proceedings of the 23rd International Conference on World Wide Web, pp. 493-498. ACM, 2014.

Nallapati, Ramesh, Amr Ahmed, William Cohen, and Eric Xing. "Sparse word graphs: A scalable algorithm for capturing word correlations in topic models." In Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on, pp. 343-348. IEEE, 2007.

Stilo, Giovanni, and Paola Velardi. "Hashtag sense clustering based on temporal similarity." Computational Linguistics 43, no. 1 (2017): 181-200.

Weston, Jason, Sumit Chopra, and Keith Adams. "\# tagspace: Semantic embeddings from hashtags." In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1822-1827. 2014.

Yogatama, Dani, Daniel Gillick, and Nevena Lazic. "Embedding methods for fine grained entity type classification." In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), vol. 2, pp. 291-296. 2015.

Zhou, Denny, Olivier Bousquet, Thomas N. Lal, Jason Weston, and Bernhard Schölkopf. "Learning with local and global consistency." In Advances in neural information processing systems, pp. 321-328. 2004.