LDA-based Topic Mining of Microblog Comments

Xu Liu, Ying Gao, Ziyi Cao, Guang Sun*
Hunan University of Finance and Economics, Changsha 410205, China

*aurorawhale@vip.qq.com

Abstract. Microblog comments express the public's views and attitudes on Microblog hot topics and hot events. Therefore, by excavating the theme of Microblog comments, we can help people understand the trend of public opinion and the attitude of the public towards the event. In view of Microblog comment topic mining, the problem we need to solve is how to use simple methods to quickly and efficiently obtain Microblog comment content, and how to use what methods to mine the theme of Microblog hot events to a great extent. First of all, we construct dynamic links through static web pages, write Microblog comment web crawler program in Python language, and solve the problem of crawling Microblog dynamically loading web page comments. Secondly, we use the Latent Dirichlet Allocation model, which is the implicit Dirichlet distribution, to mine the text.

Keywords: Latent Dirichlet Allocation model, text processing, topic mining

1. Introduction
Microblog is a platform for information release, sharing, comment, and dissemination based on social network relationships between users. Compared with the traditional news media, the information provider of microblog is no longer a single news media individual and organization, and the new hot events and news can be sent, shared, and commented on microblog by the general public for the first time. And because of the amplification affect the speed of microblog faster and wider than the traditional news media, the influence is more profound. How to quickly and accurately extract the topic and public opinion in Microblog comments to track the change of people's attitude towards a particular event, individual, company, or government behavior is the focus of our research[1].

2. Related works
Pal proposed an algorithm to find key people in Twitter topic[2]Sankaranarayanan used Twitter data to create a news system Twitter Stand [3]that can capture current hot news topics.
Weng et al[4] The Tweets sent by the same user are aggregated into a long document, which is then modeled to train the Latent Dirichlet Allocation model. Wang et al[5] [6] Using keywords (hashtag) in Twitter, tweets with the same or collinear keywords (hashtag) are aggregated to train Latent Dirichlet Allocation model.

But the research literature on Chinese Microblog is not many, because the difference between Chinese and English languages leads to the great difference between Chinese Microblog and English Microblog, and the text processing in Chinese and English is very different. So that the Twitter research can not be directly applied to Chinese Microblog.

Zheng Feiran uses the vector space model to detect the keywords in Microblog, and on this basis, by calculating the growth rate of word frequency and word frequency, to judge whether a word is a news word or not. Finally, the clustering method is used to find the news topic[7].

Zhang Chenyi of Zhejiang University pointed out that although extensive research has been carried out on the theme mining of traditional texts, for special short texts such as Microblog reviews, traditional text mining algorithms have inherent shortcomings in modeling them. Therefore, MB-LDA model is put forward to mine Microblog theme [8]. By analyzing the short text of Microblog, Peng Zeying thinks that the short text of Microblog has the characteristics of "long tail distribution", and puts forward the cluster analysis of Microblog short text using incomplete clustering. To some extent, it is difficult to deal with the problem of Microblog short text by propagating the clustering analysis algorithm[9].

Sun Shengping combined vector space model and SinglePass incremental clustering algorithm to propose SP&HA clustering algorithm [10]. By analyzing the characteristics of Microblog information, such as sparsity, multi-dimension, and massiness, Tang Xiaobo put forward the text preprocessing technology according to the characteristics of Microblog information, and then used the Latent Dirichlet Allocation model to mine the topic[11].

LDA theme model is selected in our experimental model, which is used to complete a series of text analysis and topic mining for Microblog comments.

3. Lda model

3.1. Overview

In natural language processing, the latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. LDA is an example of a topic model and belongs to the machine learning toolbox and in a wider sense to the artificial intelligence toolbox.

3.2. Model Definitions

LDA is a typical word bag model, that is, it thinks that a document is a set of words, there is no order and sequence between words. A document can contain multiple topics, and each word in the document is generated by one of the topics.

Additionally, just as the Beta distribution is the conjugate prior probability distribution of the binomial distribution, the dilikre distribution is the conjugate prior probability distribution of the
polynomial distribution. Therefore, as described LDA the Bayesian network structure, a document is generated in an LDA model as follows:

- The $\theta_i$ of topic distribution is the document is generated by sampling from the Topic distribution $\alpha$.
- Sampling from the topic polynomial distribution $\theta_i$ to generate document $i$ topic $Z_{i,j}$ of the $j$ word.
- The word distribution of the subject $Z_{i,j}$ from the sample of the Topic distribution $\phi_{z_{i,j}}$.
- Sampling the final word $w_{i,j}$ from the polynomial distribution of words $\phi_{z_{i,j}}$.

So the joint distribution of all visible and hidden variables in the whole model is

$$p(w_i, z_i, \theta, \Phi | \alpha, \beta) = \prod_{j=1}^{N} p(\theta_i | \alpha) p(z_{i,j} | \theta_j) p(\Phi | \beta) p(w_{i,j} | \phi_{z_{i,j}})$$

(1)

The maximum likelihood estimate of the word distribution in the final document can be summed

$$p(w_i | \alpha, \beta) = \int_{\theta} \int_{\Phi} \sum_{z_i} p(w_i, z_i, \theta, \Phi | \alpha, \beta)$$

(2)

by integrating the $\theta_i$ and $\Phi$ of the upper form.

According to the maximum likelihood estimation of the $p(w_i | \alpha, \beta)$, the parameters in the model can be estimated by Gibbs sampling and other methods.

4. Lda microblog review mining experiment

After text processing, you can already directly mining topics. There are many models of topic mining, we choose Latent Dirichlet Allocation model. There are special related Lda packages in the Python to download directly. We use lda packages directly to mine the text.

After simple noise reduction, word segmentation, and deactivating words, the document can not be directly used for text analysis, but the document information should be vectorized first. We use the most basic Bag-of-words model to quantify the information of the document and use the Python sklearn library to generate the document entry matrix. The program first reads the processed text information, then calls the fit and transform functions to learn and process the document. The result of the program is to generate a sparse matrix. To facilitate reading, we turn it into readable results.

We are implemented in the PyCharm editing platform, its LDA implementation is implemented with the class in the use sklearn.decomposition. Latent Dirichlet Allocation. Since LDA is based on word frequency to do modeling and analysis, it is generally not suitable to use TF-IDF to do document features in theory, but it is not impossible to do such an attempt. We just use TF-IDF to do document features. After mining the LDA model, we output the top five most likely topics, each of which outputs ten keywords (the number of topics and possible keywords can be set as needed).

For the whole text analysis, the general theme of netizens' comments is the criticism of Zhai Tianlin's academic fraud and the impact of such incidents on their future papers. Another is the
recognition of Zhai Tianlin's acting skills. I hope he can learn from it. Stand up again and continue to be a good actor.

5. Conclusion
Through the related introduction and operation, we mainly solve three problems, first: on the basis of Python, we realize the simple, fast and effective crawling of Microblog comments by constructing a web crawler algorithm for dynamic web page comments; Second, using text processing technology, including noise reduction, word segmentation and deletion of stop words, to a certain extent noise pollution processing; third, using LDA model mining program, through the completion of related operations, text topics, and simple analysis. Based on the limitations of this work, we do not define the relevant variables to test the model, but take the subjective feeling questionnaire to test. By comparing the high-frequency vocabulary of the questionnaire, the subject result vocabulary obtained by the LDA model in this experiment has a coincidence degree of 76.4 percent. Combined with the subjective feelings of ten testers on the questionnaire and LDA models, the results of this experiment are also highly consistent with the actual phenomenon.

6. Future work
There are still many imperfections in our research contents and methods. The further research directions are:

(1) In Microblog comments, there are plenty of words with exactly the same meaning, and the later work can construct a Chinese word dictionary to similar the words with the same meaning.

(2) There are still some shortcomings in the theme mining effect of the LDA model when it is aimed at short text such as comments, such as not expressing the theme well, and improving the LDA model for short text in the future.

(3) LDA model is not easy to carry out back-testing, so as to verify the accuracy of the model. The next step will define a related variable, such as an emotional index is used to test the model.

Acknowledgment
This work was supported in part by the National Natural Science Foundation of China, grant number 72073041. Open Foundation for the University Innovation Platform in the Hunan Province, grant number 18K103; 2011 Collaborative Innovation Center for Development and Utilization of Finance and Economics Big Data Property, Universities of Hunan Province. Open project, grant number 20181901CRP03, 20181901CRP04, 20181901CRP05. 2020 Hunan Provincial Higher Education Teaching Reform Research Project under Grant HNJG-2020-1130, HNJG-2020-1124. 2020 General Project of Hunan Social Science Fund under Grant 20B16.

References
[1] Jinxiang W 2015 Microblog topic mining based on topic model
[2] Pal A and Counts S 2011 Identifying topical authorities in microblogs WSDM (2011 Hong Kong China)
[3] Sankaranarayanan J Samet H and Teitler B E 2009 Twitterstand: news in tweets GIS(2009 Seattle/ Washington)
[4] Weng J Lim E P Jiang J and Qi H 2010 Twitterrank: finding topic-sensitive influential twitterers WSDM(2010 New York)

[5] Yuan W Jie L Yalou H and Feng X 2016 Using Hashtag Graph-Based Topic Model to Connect Semantically-Related Words Without Co-Occurrence in Microblogs TKDE D 28 2531661

[6] Wang Y Liu J Jishi Q Yalou H Jimeng C and Xia F 2014 Hashtag graph based topic model for tweet mining ICDM (2014 Shenzhen China)

[7] Feiran Z Zhiqian M Zhifei Z Can G 2012 A method to detect news topics in Chinese microblogs Computer Science D 39 138141

[8] Chenyi Z Jianling S Yikun D 2011 MB-LDA model based microblog topic mining CRAD D 48 17951802

[9] Zeying P Xiaoming Y Hongbo X Chunyang L 2011 Incomplete clustering of large scale short texts CJOIS D 25 054059

[10] Shengping S 2011 Research on hot topic detection and tracking technology for Chinese microblogs

[11] Xiaobo T Hongyan W 2012 Research on microblog topic mining model based on latent semantic analysis LIS D 56 114119