Short Text Topic Recognition and Optimization Method for University Online Community

Xu Wu[1,2,3](✉), Haitao Wu[1,2,3], Xiaqing Xie[1,2,3], Jin Xu[1,2,3],
and Tianle Zhang[4]

[1] Key Laboratory of Trustworthy Distributed Computing and Service,
Ministry of Education, Beijing, China
[2] School of Cyberspace Security, BUPT, Beijing, China
{wux, wht, xiexiaqing}@bupt.edu.cn
[3] Beijing University of Posts and Telecommunications Library,
Beijing 100876, China
[4] Cyberspace Institute of Advanced Technology, Guangzhou University,
Guangzhou 510006, China
tlezhang@sohu.com

Abstract. The university online community mainly records what happens in target areas and groups of people. It has the characteristics of timeliness, regional strong and clear target groups. Compared with Weibo and post-bar, university community's text topic recognition needs to solve the problems of large text noise, fast text update and short single text content. To this end, this paper proposes a method of building university topic model based on LDA topic model. Through the steps of original text's noise reduction, LDA (Latent Dirichlet Allocation (LDA), is a topic model commonly used in the field of machine learning and is often used for text categorization.) model recognition and weighted calculation of recognition results, etc., the event themes that characterize the common characteristics for university online community are obtained. Experiments based on real university online community's data show that the topic model of university popular events established by the topic recognition model of this paper can reflect some popular events in colleges and universities, so as to provide reasonable support for university management.

Keywords: University online community · University topic model · Popular events in colleges and universities

1 Introduction

The vigorous development of the Internet has made more and more netizens gather in the online community such as Weibo and post-bar and so on. In particular, college students express their views, discussed hot topics about their school in the university online community. As the active network community in Colleges and universities, the amount of information contained in the network community has increased geometrically. Correspondingly, the information related to all aspects of College Students' life

X. Sun et al. (Eds.): ICAIS 2019, LNCS 11635, pp. 79–90, 2019.
https://doi.org/10.1007/978-3-030-24268-8_8

has become a hot research contents. This paper focuses on the contents of university online community, proposes an improved LDA method for text topic recognition, and gives weight to the results. By calculating the weight, we can judge whether the output results can represent the corresponding topic model. The text of university network community has the problem of large text noise, fast text updating and short single text. To solve these problems, firstly, this paper extends the LDA model's stop word library and filters a large number of general useless replies to decrease the text noise. Then, uses the LDA topic model to generate a word vector group which can represent this topic, and each vector is given different weights. By calculating the weights, we can judge whether the topic model can truly represent the typical topic. Lastly, this paper establishes some topic models which can represent the typical events in universities. The experimental results show that the topic recognition model based on this paper can better reflect some of the hot events in universities than the traditional LDA topic model, so as to provide reasonable support for university management.

2 Related Research

2.1 Related Work

At present, some researches have used LDA topic model or improved LDA topic model to mine topic models in different fields. Zhu Maoran et al. use LDA model to identify the topic and its key words in the whole text set, and calculated the document-topic probability distribution in each time windows. Then the topic-lexical probability distribution is calculated by LDA model for each text set under each time window, and the similarity of different topics under different time windows is calculated, thus the evolution trend of topic strength is obtained. Finally, we get the change of thematic content through the probability distribution of words under similar themes. [1] Peng et al. embed a priori knowledge of word association, global feature words and subject emotion subordinate semantics in LDA model to improve LDA's ability to recognize feature words, affective words and their relationships. And it uses LDA model to extract the feature words and emotional words of product reviews, and take advantage of affective analysis technology [2]. To classify emotional polarity of product reviews Yi et al. introduce the strategic coordinates into the subject analysis of LDA patents, and fuses the time factors, using the internal and external correlation index to measure the relationship between the various parts of the subject. [3] Tan et al. apply the LDA model to the mining of hot topics, and constructs the identification process of Weibo hot topics. [4] Zheng et al. propose an improved LDA model ST-LDA to analyze Weibo topics. This model assigns a topic to each Weibo. In particular, when assigning a topic to each Weibo, not only the influence of posting time on the topic is considered, but also the influence of posting location on the topic is added. Thus, the model can assign a topic to each Weibo. The semantic similarity of Weibo published in adjacent time and space is classified into the same topic, which improves the comprehensibility of the topic. [5] Xu et al. use LDA topic model to analyze the topic relevance of blog content, redefine the link relationship between bloggers, and then classify the topics. [6] Shi et al. mainly explore the evolution of BBS as a topic in the network media in time, so as to find hot topics and no-hot topics, and better guide the netizens to understand what is happening. [7] He et al. proposes a simple and easy symmetric learning data augmentation model (SLDAM) for underwater target radiate-noise data expansion and generation. The SLDAM, taking the optimal classifier of an initial dataset as the discriminator, makes use of the structure of the classifier to construct a symmetric generator based on antagonistic generation. It generates data similar to the initial dataset that can be used to supplement training data sets. [8] Wang et al. proposes a natural language semantic construction method based on cloud database to improve natural language comprehension ability and analytical skills of the machine [9].

The research above has done a lot of work on the application level of LDA, and has also been extended based on this theory to optimize the topic recognition in specific areas. However, we can also see the shortcomings: on the one hand, LDA is an unsupervised learning model, the input content will greatly affect the accuracy of the output results, and these methods do not effectively de-noise the input data, which still has a lot of room for optimization; on the other hand, the output of the LDA model is considered the final result of the corresponding topic without considering the intrinsic correlation between the output of the LDA model. Although LDA model treats the segmentation results equally, the results of LDA model can be re-weighted to determine whether the results can represent the corresponding topic model.

2.2 LDA Model and Application

Latent Dirichlet Allocation (LDA), is a topic model commonly used in the field of machine learning and is often used for text categorization.

LDA was put forward in 2003 by Blei, David M., Ng, Andrew Y. and Jordan to predict the topic distribution of documents. It can give the topic of each document in the document set in the form of probability distribution. After analyzing some documents and extracting their topic distribution, it can cluster or classify the text according to the topic distribution.

Suppose there are M documents, corresponding to the d document, there are N_d words, that is, the input is as follows (Fig. 1):

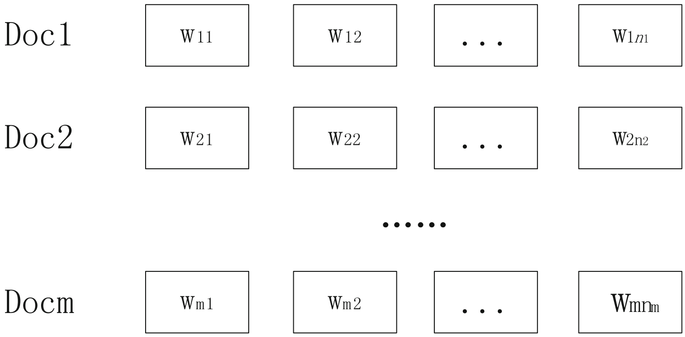


Fig. 1. The LDA Input schematic diagram.

In order to find the topic distribution of each document and the word distribution in each topic, this paper first assumes a topic number K, so that all distributions are based on K topics. Therefore, a specific LDA model can be obtained, as shown in the following Fig. 2:

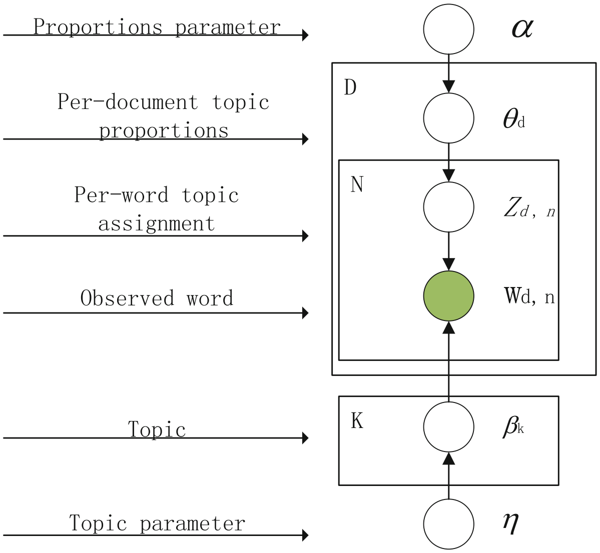


Fig. 2. The LDA model schematic diagram.

LDA assumes that the prior distribution of a document topic is a Dirichlet distribution, that is, for any document d, the topic distribution θ_d is:

$$\theta_d = Diriclet(\vec{\alpha}) \tag{1}$$

Among them, α is a distributed super parameter, which is a K dimensional vector.

LDA assumes that the prior distribution of words in a topic is Dirichlet distribution, that is, for any topic k, the word distribution β_k is:

$$\beta_k = Dirichlet(\vec{\eta}) \tag{2}$$

Among them, the parameter η is a V dimensional vector. V stands for the number of words in the word list.

For the nth word in any document d in the data, the distribution of the subject number z_{dn} is obtained from the topic distribution θ_d as follows:

$$z_{dn} = multi(\theta_d) \tag{3}$$

For this topic number, the probability distribution of word w_{dn} can be obtained as follows:

$$w_{dn} = multi\left(\beta_{z_{dn}}\right) \tag{4}$$

Thus, for M document topics, the corresponding data has multiple distributions of M topic numbers, so that $\alpha \rightarrow \theta_d \rightarrow \vec{z}_d$ constitutes a Dirichlet-multi conjugate. A posteriori distribution of document topics based on Dirichlet distribution can be obtained by Bayesian inference.

If the number of words for topic K in the dth document is $n_d^{(k)}$, the count of polynomial distributions can be expressed as:

$$\vec{n}_d = \left(n_d^{(1)}, n_d^{(2)}, \ldots n_d^{(K)}\right) \tag{5}$$

The posterior distribution of θ_d is obtained by using Dirichlet-multi conjugate is:

$$Dirichlet(\theta_d|\vec{\alpha} + \vec{n}_d) \tag{6}$$

For the distribution of subjects and words, if there is a Dirichlet distribution of K subjects and words, and the corresponding data has a polynomial distribution of K subject numbers, then $\eta \rightarrow \beta_k \rightarrow \vec{w}_{(k)}$ constitutes a Dirichlet-multiconjugate. Bayesian inference method can be used to obtain a posterior distribution of subject words based on the Dirichlet distribution.

If the number of words V in topic K is $n_k^{(v)}$, the number of polynomial distributions can be expressed as:

$$\vec{n}_k = \left(n_k^{(1)}, n_k^{(2)}, \ldots n_k^{(V)}\right) \tag{7}$$

By using Dirichlet-multi conjugation, the posterior distribution of β_k is obtained as:

$$Dirichlet(\beta_k|\vec{\eta} + \vec{n}_k) \tag{8}$$

Because the subject-generated words do not depend on a specific document, the topic distribution and the subject-word distribution of the document are independent.

3 Establishing the Theme Model of University Online Community

3.1 Text Topic Recognition Method in University Online Community

The concept of topic area is mostly adopted in the network community of colleges and universities. The topics discussed in the same topic area can be summed up as a kind of campus-related topics. Every post in the topic area revolves around a particular topic. If you ignore the useless responses in a single post, you can assume that all the responses are thematic responses. Considering these characteristics of college network community, this paper preprocesses the input text before generating topic distribution using LDA model, as shown in the following Fig. 3:

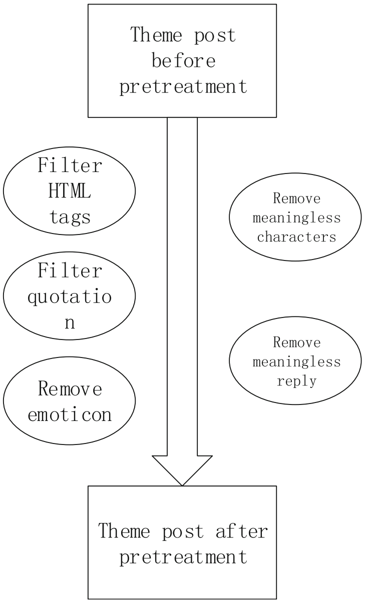

Theme post before pretreatment
Filter HTML tags
Filter quotatio n
Remove emoticon
Remove meaningless characters
Remove meaningless reply
Theme post after pretreatment

Fig. 3. Schematic diagram of text preprocessing

After pretreatment, the topic posts can be used as the input of LDA model to get the distribution of the topic words, because all the noises that affect the results are removed. And because each topic Posts contain only one topic, we can easily get the distribution of the topic words of a hot topic. As the LDA model is a bag of words model, the result can only show the frequency distribution of the subject words, and can not be divided in detail according to the weight of different words. According to the characteristics of the subject words in the topic model, this paper proposes a method of weight assignment according to the category of the subject words. By calculating the final weight results, we can get the result of weight assignment. Determine whether these topic phrases can correctly represent the corresponding topic models. Therefore, the whole process of text topic recognition in this paper is shown in the following Fig. 4:

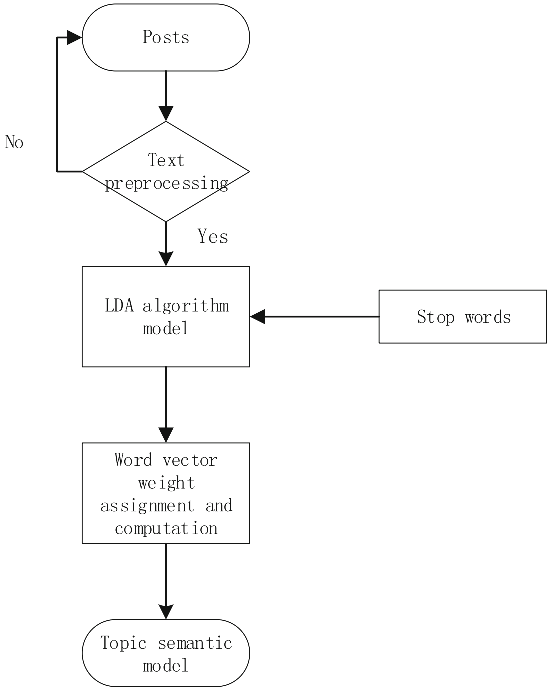


Fig. 4. Flow chart of text topic recognition

3.2 The Establishment of Popular Topic Model in University Online Community

The heat of the topic area in the network community reflects the students' attention to the corresponding events in universities. This paper investigates a large number of hot topics in the network community of colleges and universities, and obtains the topics closely related to colleges and universities shown in the following Fig. 5:

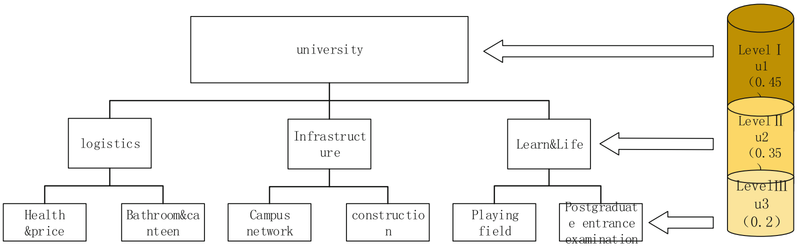


Fig. 5. Flow chart of text topic recognition

From the above chart shows, the hot topics related to the campus is divided into three categories, six small classes, which basically covers all aspects of campus life. This paper establishes the corresponding topic model bases on these six categories. The right-most side of the graph shows the weight of each level of the hot topic. For the model level, the letter u is used to represent the level, and $u_1u_2u_3\ldots u_n$ is used to represent the corresponding level. For each level, this article uses the letter w to represent the corresponding weight, $w_1w_2w_3\ldots w_n$ denotes the weight of the corresponding hierarchy, the letter $\lambda = \delta / \sum_0^n u$ denotes the compensation value, and the value of delta is related to the word frequency input by the LDA model. If the maximum word frequency is less than 0.1, then δ = 0.1, and so on. The formula for calculating weight can be obtained as follows:

$$f(u, w) = \frac{u_1 * w_1 + u_2 * w_2 + u_3 * w_3 + \ldots + u_n * w_n}{u_1 + u_2 + u_3 + \ldots u_n} * \lambda \tag{9}$$

The weighted result of the topic model can be obtained by using the above formula. If the result is greater than 0.5, the input text can be used to represent the corresponding topic model.

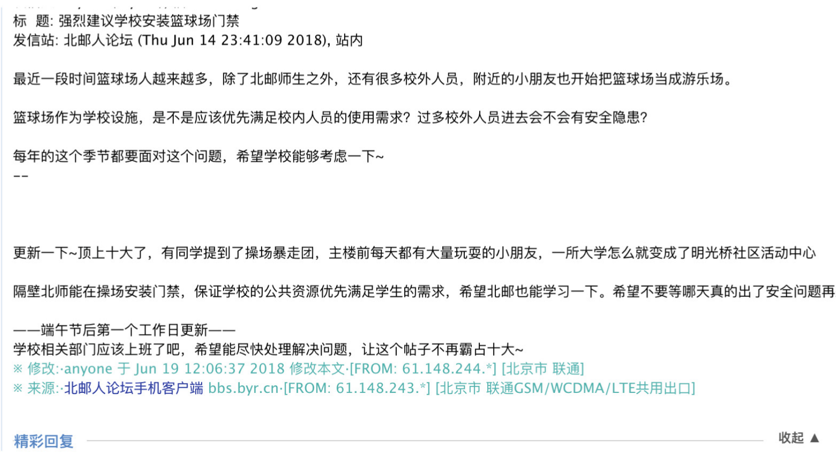
4 Experimental Results and Analysis

4.1 Text Topic Recognition Analysis of University Network Community

The University Forum of Beijing University of Posts and Telecommunications is one of the typical network communities in Colleges and universities. According to the different topics, the forum is divided into eight discussion areas, each of which includes the website, the campus of Beijing University of Posts and Telecommunications, academic science and technology, information society, humanities and arts, life style, leisure and entertainment, physical fitness, games and love of hometown. There are other discussion areas behind those areas. By analyzing the content of the discussion area, this paper draws a conclusion that posts posted on specific plates have obvious characteristics. For example, in the life fashion section, the discussion is mostly about life-related topics, such as canteen, bathroom, supermarket, and so on, while in the health care section, the discussion is mostly about hospitals, nursing and other topics. These topic posts also have a strong attribute: that is, the response below a topic post, except for some irrelevant replies, the rest of the replies are related to the topic of the topic post discussion.

This paper selects the posts related to campus life in the life style area of the forum of Beijing University of Posts and Telecommunications, and obtains the contents of each posts in the zone by means of web crawler.

Web crawler is also called web spider, or web robot. It is a kind of programs or scripts according to certain Rules that automatically grab information from the World Wide Web. In other words, it can automatically get the content of the web page according to the link address of the web page. If we compare the Internet to a big spider web, it has a lot of web pages, web spiders can get all of webpage's content. [10] In this paper, we can find the origin contents on the website like this (Fig. 6):



标 题: 强烈建议学校安装篮球场门禁
发信站: 北邮人论坛 (Thu Jun 14 23:41:09 2018), 站内

最近一段时间篮球场人越来越多，除了北邮师生之外，还有很多校外人员，附近的小朋友也开始把篮球场当成游乐场。

篮球场作为学校设施，是不是应该优先满足校内人员的使用需求？过多校外人员进去会不会有安全隐患？

每年的这个季节都要面对这个问题，希望学校能够考虑一下~
--

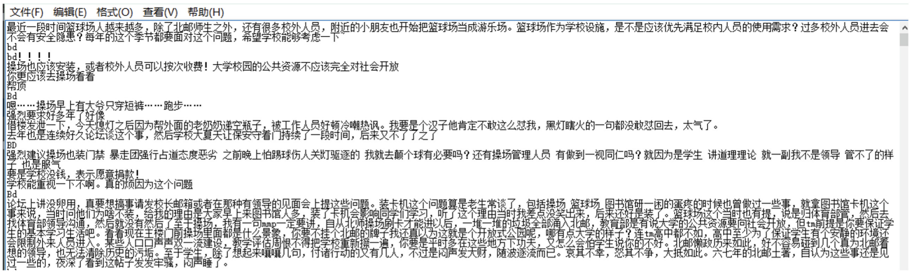
更新一下~顶上十大了，有同学提到了操场暴走团，主楼前每天都有大量玩耍的小朋友，一所大学怎么就变成了明光桥社区活动中心

隔壁北师能在操场安装门禁，保证学校的公共资源优先满足学生的需求，希望北邮也能学习一下。希望不要等哪天真的出了安全问题再来补救。

——端午节后第一个工作日更新——
学校相关部门应该上班了吧，希望能尽快处理解决问题，让这个帖子不再霸占十大~
※ 修改:·anyone 于 Jun 19 12:06:37 2018 修改本文·[FROM: 61.148.244.*] [北京市 联通]
※ 来源:·北邮人论坛手机客户端 bbs.byr.cn·[FROM: 61.148.243.*] [北京市 联通GSM/WCDMA/LTE共用出口]

Fig. 6. The origin text on the website

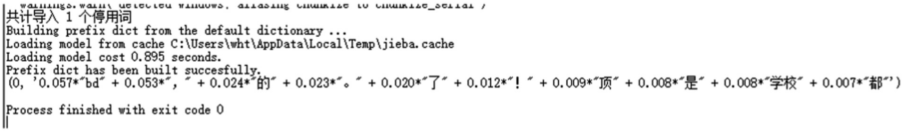
The origin contents above can’t use by LDA model before text preprocessing completed. We can see the whole text below (Fig. 7):



最近一段时间篮球场人越来越多，除了北邮师生之外，还有很多校外人员，附近的小朋友也开始把篮球场当成游乐场。篮球场作为学校设施，是不是应该优先满足校内人员的使用需求？过多校外人员进去会
不会有安全隐患？每年的这个季节都要面对这个问题，希望学校能够考虑一下~
bd
bd！！！！
操场也应该安装，或者校外人员可以按次收费！大学校园的公共资源不应该完全对社会开放
你更应该去操场看看
帮顶
Bd
嗯……操场早上有大爷只穿短裤……跑步……
强烈要求好多年了好像
借楼发泄一下，今天熄灯之后因为帮外面的老奶奶递空瓶子，被工作人员好顿冷嘲热讽。我要是个汉子他肯定不敢这么怼我，黑灯瞎火的一句都没敢怼回去，太气了。
去年也是连续好久论坛谈这个事，然后学校大夏天让保安守着门持续了一段时间，后来又不了了之了
BD
强烈建议操场也装门禁 暴走团强行占道态度恶劣 之前晚上怕踢球伤人关灯驱逐的 我就去颠个球有必要吗？还有操场管理人员 有做到一视同仁吗？就因为是学生 讲道理理论 就一副我不是领导 管不了的样
子 也是服气
要是学校没钱，表示愿意捐款！
学校能重视一下不啊。真的烦因为这个问题
Bd
论坛上讲没卵用，真要想搞事请发校长邮箱或者在那种有领导的见面会上提这些问题。装卡机这个问题算是老生常谈了，包括操场 篮球场 图书馆研一闲的蛋疼的时候也曾做过一些事，就拿图书馆卡机这个
事来说，当时问他们为啥不装，给我的理由是大家早上来图书馆人多，装了卡机会影响同学们学习，听了这个理由当时我差点没笑出来，后来还好是装了。篮球场这个当时也有提，说是归体育部管，然后去
找体育部领导沟通，然后就没有然后了至于操场，我有一句mmp一定要讲，自从北师操场刷卡才能进以后，一堆一堆的垃圾全部涌入北邮，教育部是有说大学的公共资源要向社会开放，但tm前提是你要保证学
生的基本学习生活吧。看看现在主楼门前操场里面都是什么景象，你要不挂个北邮的牌子我还真以为这就是个开放式公园呢，哪有点大学的样子？连tm高中都不如，高中至少为了保证学生有个安静的环境还
会限制外来人员进入。某些人口口声声双一流建设，教学评估周恨不得把学校重新擦一遍，你要是平时多在这些地方下功夫，又怎么会怕学生说你的不好。北邮懒政历来如此，好不容易碰到几个真为北邮着
想的领导，也无法清除历史的污垢。至于学生，除了想起来嚷嚷几句，付诸行动的又有几人，不过是闷声发大财，随波逐流而已。哀其不幸，怒其不争，大抵如此。六七年的北邮土著，自认为这些事还是见
过一些的，夜深了看到这帖子发发牢骚，闷声睡了。

Fig. 7. The noise text without text preprocessing

So, if we use LDA directly, the results obtained by directly using LDA model are as follows (Fig. 8):



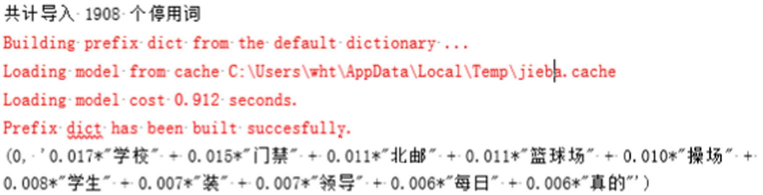
```
  warnings.warn("detected Windows; aliasing chunkize to chunkize_serial")
共计导入 1 个停用词
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\wht\AppData\Local\Temp\jieba.cache
Loading model cost 0.895 seconds.
Prefix dict has been built succesfully.
(0, '0.057*"bd" + 0.053*"，" + 0.024*"的" + 0.023*"。" + 0.020*"了" + 0.012*"！" + 0.009*"顶" + 0.008*"是" + 0.008*"学校" + 0.007*"都"')

Process finished with exit code 0
```

Fig. 8. LDA result without text preprocessing

As you can see, the phrases with the highest frequency are in turn “bd”, “的”, “了” and other characters that have nothing to do with the subject.

After the text preprocessing, the output of the LDA thematic model is shown in the following Fig. 9:



```
共计导入 1908 个停用词
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\wht\AppData\Local\Temp\jieba.cache
Loading model cost 0.912 seconds.
Prefix dict has been built succesfully.
(0, '0.017*"学校" + 0.015*"门禁" + 0.011*"北邮" + 0.011*"篮球场" + 0.010*"操场" +
0.008*"学生" + 0.007*"装" + 0.007*"领导" + 0.006*"每日" + 0.006*"真的"')
```

Fig. 9. LDA result using text preprocessing

Furthermore, a detailed comparison of the above results is shown in the following Table 1:

Table 1. A detailed comparison of the two results

Comparison item	Origin LDA	Improved LDA
Topic relevance	Weak	Strong
Accuracy rate	Low	High
Able to represent events	No	Yes

We can see that after the text preprocessing process proposed in this paper, the correlation between the topic distribution obtained by LDA model and the actual topic distribution involved has been effectively improved.

4.2 Verification of Popular Topic Model in University Online Community

After obtaining the frequency of the above-mentioned topic distribution words, different weights are assigned to them, and a topic model similar to the one shown in the Fig. 10 is obtained:

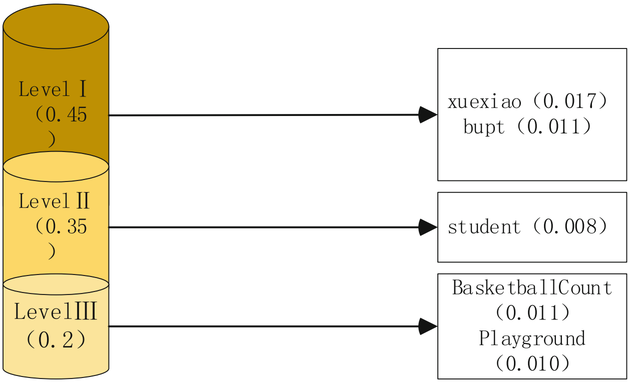


Fig. 10. An example model using the theory of this paper

So the weight of the vector frequency of the above words can be calculated like this:

$$\begin{aligned} \mathrm{f}(\mathrm{u},\mathrm{w}) &= \frac{((0.017+0.011)*0.45+0.008*0.35+(0.011+0.010)*0.2)*0.1}{(0.017+0.011+0.08+0.011+0.010)^2} \\ &= 0.6033 \end{aligned} \tag{10}$$

Because of 0.6033 > 0.5, this paper can draw the conclusion that the above topic models can well represent the corresponding themes. That is: the content of the theme is related to the basketball court and playground closely related to the life of college students.

This conclusion also applies to other similar campus themes if we strictly fellow the steps above.

5 Summary and Prospect

The university online community is closely related to schools and students. How to dig up the information from it and serve the schools and students is the current concern of all colleges and universities. This paper focuses on the characteristics of university online community, proposes an improved LDA method for text topic recognition. Firstly, this paper extends the LDA model's stop word library and filters a large number of general useless replies to decrease the text noise. Then, uses the LDA topic model to generate a word vector group which can represent this topic, and each vector is given different weights. By calculating the weights, we can judge whether the topic model can truly represent the typical topic. Lastly, this paper establishes some topic models which can represent the typical events in universities. In the future, on the basis of this study, we can also establish other new thematic models according to the characteristics of universities, improve the calculation method of thematic model weight, and explore the development trend of thematic models of typical events in universities.

Acknowledgements. This work is supported by the National Key Research and Development Plan(Grant No. 2017YFC0820603), BUPT's Informatization Innovative Application Project, " privacy protection and data release on Campus big data" and the Open Project Fund of Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education.

References

1. Zhu, M., Wang, Y., Gao, S., Wang, H., Zhang, X.: Evolution of topic using LDA model: Evidence From Information Science Journals. J. Beijing Univ. Technol. **07**, 1047–1053 (2018)
2. Peng, Y., Wan, H., Zhong, L.: Fine-grained sentiment analysis algorithm of product reviews based on semantic weakly-supervised LDA. J. Mini-micro Syst. **05**, 978–985 (2018)

3. Yi, H., Wu, H., Ma, Y., Ji, F.: Technical topic analysis in patents based on LDA and strategic diagram by taking graphene technology as an example. J. Intell. **05**, 97–102 (2018)
4. Tan, C., Wang, C., Zhang, Y.: A hot topic identification based on LDA for Chinese microblog. J. Suzhou Univ. **04**, 71–77 (2014)
5. Zheng, Z., Jin, B., Cui, Y.: Study on recognition of spatial-temporal events based on microblogs. J. Comput. Sci. **10**, 214–219 (2016)
6. Xu, B., Zhao, C., Zhang, Y.: Topic community mining in blogosphere based on LDA. J. Comput. Digit. Eng. **11**, 40–43 (2012)
7. Shi, D., Zhang, H.: LDA model-based BBS topic evolution. J. Industr. Control Comput. **05**, 82–84 (2012)
8. He, M., Wang, H., Zhou, L., Wang, P., Ju, A.: Symmetric learning data augmentation model for underwater target noise data expansion. CMC: Comput. Mater. Continua **57**(3), 521–532 (2018)
9. Wang, S., et al.: Natural language semantic construction based on cloud database. CMC: Comput. Mater. Continua **57**(3), 603–619 (2018)
10. A simple web crawler for crawler learning. https://www.cnblogs.com/chenkun/p/5653459.html