Research on Evaluation Method of LDA Topic Model in Mail Classification

Hechen Gong*, Fucheng You and Shuren Lai

Beijing Institute of Graphic Communication. No. 1, Xinghua Street (two section), Daxing District, Beijing
Email: gonghechen123@163.com

Abstract. The LDA topic model is a document generation model, which is an unsupervised machine learning technique that can be used in keyword extraction, topic classification, and so on. The main purpose of this paper is to effectively evaluate the number of optimal topics of the LDA topic model in the message subject classification and extract the importance of the topic, so that the LDA topic model can be highly readable after subject classification. Therefore, this paper proposes a keyword matching and subjective statistical value word comparison (KM-SSVW) for the subject classification of emails, the keyword matching technique in this method uses the TF-IDF technique. This method can accurately evaluate the extracted keywords and optimize the number of topics. The data in this article is mainly from the mail gate, a total of 7,000 messages that Hillary communicates with others. The empirical results show that the proposed method of keyword matching and subjective statistical value word comparison has a good effect on subject quantity optimization and subject word readable evaluation. However, there are still some limitations in this paper. In the experiment, new methods are not validated for other types of data sets, such as microblog short text, XML documents, and WeChat public platform articles.

1. Introduction

The topic model has become a hot topic in the field of machine learning. LDA (Latent Dirichlet Allocation) [1] is a document theme generation model, also known as a three-layer Bayesian probability model, which contains three-layer structure of words, topics and documents. The so-called generation model, that is, we believe that each word in an article is obtained through a process of "choosing a topic with a certain probability and selecting a certain word from the topic with a certain probability".

In recent years, LDA topic models have been widely used in the field of natural language processing, such as text categorization, topic extraction, sentiment analysis and public opinion analysis, and the application in the field of data mining is also very hot. Mainly focused on text classification [2-5], text topic extraction [6-7], scientific literature knowledge mining [8-10], academic evaluation [11] and other research directions. Blei [12] and others proposed the CTM model, mainly to better express the potential topic information of the text collection. Li Wenbo [13] proposed the Labeled-LDA model. The innovation of this model is mainly to introduce category identification information in LDA, which reduces the training number of LDA model. Yuan Boqiu [14] applied the LDA model to feature selection technology, mainly for spam processing. A large number of empirical studies have proved the reliability and effectiveness of LDA, but there are still problems. Although LDA has a wide range of applications in topic classification, determining the optimal number of topics and the readability of extracted keywords is a key issue.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
2. LDA topic model

2.1. LDA Core Functions
The theme is the middle layer, the probability of occurrence of the word \( w \) in document \( d \) can be given by the current \( \theta_d \) and \( \phi_t \). Among them, \( p(t|d) \) is calculated by \( \theta_d \), and \( p(w|t) \) is calculated by \( \phi_t \).

\[
P_j(w_j|d) = P(w_j|t_j)*P(t_j|d)
\]  

Among them: \( w_j \) stands for words; \( d_j \) for documents; \( t_j \) for topics; \( \theta_d \) for subjects in different documents; and \( \phi_t \) for words in different topics.

2.2. Text Generation Topic Model
The model is shown in Figure 1, and the meaning of each symbol is shown in Table 1.

![Figure 1. LDA Model diagram.](image)

### Table 1. The meaning of each symbol in the LDA model

| Symbol | Meaning                      | Symbol | Meaning                      |
|--------|------------------------------|--------|------------------------------|
| \( \alpha \) | Hyperparameter of \( \theta \) | \( \omega \) | Term                         |
| \( \beta \) | Hyperparameter of \( \phi \) | \( z \)  | Subject distribution of words |
| \( \theta \) | Text-topic probability distribution | \( M \)  | Number of texts              |
| \( \phi \) | Subject-word probability distribution | \( k \)  | Number of topics              |
| \( L \) | Text length                  | \( \beta ij \) | The \( j \)-th word probability under the \( i \)-th topic |

The basic idea of the LDA theme model is to randomly generate a document consisting of \( N \) words. Each term selects a topic with a certain probability and is selected from the topic with a certain probability. Given \( \alpha \) and \( \beta \), the LDA model is represented by a probabilistic model as shown in equation (2): 

\[
P(\theta, z, w|\alpha, \beta) = P(\theta, \alpha)\prod_{n=1}^{N} P(z_n, \beta)P(w_n|z_n, \beta)
\]  

The probability of the entire corpus is shown in equation (3):

\[
P(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha)(\prod_{n=1}^{N_d} \sum_{z_{nd}} p(z_{nd}|\theta_d)p(w_{dn}|z_{nd}, \beta))d\theta_n
\]
3. Perplexity

Through the research of scholars at home and abroad, the determination of the number of topics \( z \) is one of the most important factors affecting the topic classification effect of the LDA theme model. At the same time, it can be inferred that the readability of the subject terms in the subject is also directly related to the LDA topic model. Therefore, the results of the topic extraction are very sensitive to the \( N \) value. Through research, the most common assessment technique for topic classification is Perplexity.

Blei et al. used Perplexity as a criterion for evaluating the quality of the model. In information theory, perplexity is used to judge whether a probability distribution or probability model predicts the quality of a sample. The lower the perplexity, the better the model can predict the sample [1]. In natural language processing, confusion is a method used to measure the pros and cons of a language probabilistic model. A linguistic probability model can be thought of as a probability distribution over a sentence or segment [15]. In the LDA topic model, the confusion calculation formula (probability distribution perplexity) [1] is as follows:

\[
\text{perplexity}(D_{\text{test}}) = \exp \left( \frac{\sum_{d=1}^{M} \log p(w_{d})}{\sum_{d=1}^{M} N_{d}} \right)
\]

(4)

Among them: \( M \) is the size of the test corpus, and \( N_{d} \) is the size of the \( d \)-th text (the number of words).

\[
p(w_{d}) = \sum_{z} p(z)p(w|z, \text{gamma})
\]

(5)

Among them: \( z \) is the subject, \( w \) is the document, \( \text{gamma} \) is the text-topic distribution learned from the training set, and \( p(w_{d}) \) is the probability of the word \( w_{d} \) in the document.

Among them: Gamma function

Integer case:

\[
\Gamma(n) = (n - 1)!
\]

(6)

Real situation:

\[
\Gamma(x) = \int_{0}^{\infty} t^{x-1} e^{-t} dt
\]

(7)

4. Construction of KM_SSVW Evaluation Model

4.1. TF-IDF

TF-IDF (term frequency–inverse document frequency), It is a commonly used weighting technique used in information retrieval and data mining [16]. The importance of a word increases proportionally with the number of times it appears in the file, but it also decreases inversely with the frequency it appears in the corpus.

4.1.1. TF (term frequency). Word frequency: the frequency of a specified word or phrase in a given text.

\[
\text{TF}_{i,j} = \frac{n_{i,j}}{\sum_{k} n_{k,j}}
\]

(8)
Among them: \( n_{i,j} \) denotes the number of occurrences of word \( w_i \) in text \( d_j \), and the alphabet denotes the sum of occurrences of all characters in text \( d_j \).

### 4.1.2. IDF (Inverse document frequency)

Inverse document frequency: measure the general importance of words or phrases in text.

\[
IDF_i = \log \frac{m_i}{n_i}
\]  

(9)

Among them: \( m_i \) represents the total number of documents in the corpus, and the denominator represents the number of documents containing \( n_i \).

That is to say, combine formula (8) and formula (9).

\[
TF - IDF_{i,j} = TF_{i,j} \times IDF_i
\]  

(10)

#### 4.1.3. Subject matching rate

The high frequency words in a particular document \( d \), and at the same time the low frequency words in the entire file set \( D \), can produce a high TF-IDF. Therefore, in the \( A \) topics classified by the LDA topic model, the number of keywords (\( t_l \)) with the highest weight in each topic (\( a_i \)) is calculated using TF-IDF. The \( T \) is matched with the \( w_i \) subject word in \( a_i \) classified by the LDA topic model, and the matching word quantity \( y_{i,j} \) is obtained, and the matching accuracy rate \( S_i \) is obtained. Get the maximum keyword match rate by constantly modifying the number of topics \( A \).

\[
y_i = |t_i - w_i|
\]

(11)

\[
S_i = \left| \frac{y_i}{w_i} \right|
\]

(12)

Among them: \( A \) indicates the number of topics; \( a_i \) indicates the \( i \)-th topic (\( i=1, 2, 3... \)); \( t_l \) indicates the number of keywords in the file set of the \( i \)-th topic; \( w_i \) indicates the number of subject words extracted by the \( i \)-th topic; \( y_l \) indicates the number of subject words; \( S_i \) indicates the keyword matching accuracy.

#### 4.2. The Proportion of Subjective Value Term Statistics

Before the document is categorized, the text data is preprocessed by the data, which leads to the classification of each topic in which the topic is not the meaning of each word. Therefore, through the statistical method, the valuable words (readability) of the subject words under each topic generated by the LDA topic model are obtained, thereby obtaining the proportion of the valuable words.

\[
F_i = \frac{k_i}{w_i}
\]

(13)

Among them: \( w_i \) indicates the number of subject words extracted by the \( i \)-th topic, \( k_i \) represents the number of valuable words in the \( i \)-th topic, and \( F_i \) represents the proportion of the \( i \)-th topic obtained by the LDA subject model classification.

That is, combining the formula (12) and the formula (13), the KM_SSVW evaluation model is obtained:

\[
\begin{aligned}
&F_i - S_i = \frac{k_i}{w_i} - \frac{y_i}{w_i} \rightarrow 0 \\
\text{KM_SSVW} = \begin{cases} 
F_i \rightarrow 1 \\
S_i \rightarrow 1
\end{cases}
\end{aligned}
\]

(14)
5. Experiment Procedure

5.1. Data Sources

The data in this article is mainly from the mail gate incident. There are 7,000 emails that Hillary communicates with other people. The mail gate is the hacker attack on the computer of important people in Hillary and the surrounding area. After the insider broke the news, the mail is published online through Wikileaks. Because there are many null values in the original data, the meaningless data is deleted directly, as shown in Table 2, which is the first 15 pieces of the original data of the mail.

| Number | Id  | ExtractedBodyText                                      |
|--------|-----|--------------------------------------------------------|
| 1      | 2   | B6\nThursday, March 3, 2011 9:45 PM\nH: Latest...     |
| 2      | 3   | Thx                                                    |
| 3      | 5   | H <hrd17@clintonemail.com>\nFriday, March 11,... |
| 4      | 6   | Pis print.\n- -\nH <hrd17@clintonemail... |
| 5      | 8   | H <hrd17@clintonemail,cor>\nFriday, March 11... |
| 6      | 9   | FYI                                                    |
| 7      | 10  | B6\nWednesday, September 12, 2012 6:16 PM\nFwd...     |
| 8      | 11  | FYI\nB6\n— —— |
| 9      | 12  | B6\nWednesday, September 12, 2012 6:16 PM\nFwd...     |
| 10     | 13  | FYI                                                    |

5.2. Data Preprocessing

5.2.1. Dictionary extraction, word segmentation and filtering. In natural language, text preprocessing is very important, mainly including extracting dictionaries, word segmentation and filtering. It can be understood from the original data that there are many time, date, website, email address and meaningless characters and numbers in the mail content. Get the dictionary through Python programming, use Python's split participle [18] package to segment the English raw data, and write regular expressions in Python to filter meaningless characters and numbers. Table 3 is the result of preprocessing data.

| Number | Id  | ExtractedBodyText                                      |
|--------|-----|--------------------------------------------------------|
| 1      | 2   | Thursday March PM Latest How Syria is aiding Qaddafi and more Sid hrc memo syria aiding libya docx hrc memo syria aiding libya docx March For Hillary |
| 2      | 3   | Thx                                                    |
| 3      | 5   | Friday March PM Huma Abedin Fw Latest How Syria is aiding Qaddafi and more Sid hrc memo syria aiding libya docx Pis print |
| 4      | 6   | Pis print Wednesday September PM Fw Meet The Right Wing Extremist Behind Anti f\m\s\l\t\i\r\m Film That Sparked Deadly Riots From meat Sent Wednesday September PM To Subject Meet The Right Wing Extremist Behind Anti Muslim Film That Sparked Deadly Riots htemaxbiamental.commeet the right wing extremist behind anti musiim thhn that sparked deadly riots Sent from my Verizon Wireless LTE DROID US Department of State Case No Doc No Date STATE DEPT PRODUCED TO HOUSE SELECT BENGHAZI COMM SUBJECT TO AGREEMENT ON SENSITIVE INFORMATION REDACTIONS NO FOIA WAIVER STATE CB |
| 5      | 8   | Friday March PM Huma Abedin Fw Latest How Syria is aiding Qaddafi and more Sid hrc memo Syria aiding libya docx Pis print |
| 6      | 9   | FYI                                                    |
| 7      | 10  | Wednesday September PM Fwd more on libya Libya sept docx Sending direct Just in Sent from my Verizon Wireless LTE DRUID |
| 8      | 11  | Fyi                                                    |
| 9      | 12  | Wednesday September PM Fwd more on libya Libya sept docx Sending direct Just in Sent from my Verizon Wireless LTE DRUID |
| 10     | 13  | Fyi                                                    |
5.2.2. LDA topic model and toolkit selection. LDA topic extraction is implemented by the Python language-based machine learning package gensim [17], and TF-IDF calculations are also implemented in Python programming.

5.3. Hardware Environment and Experimental Platform
The experimental environment is shown in Table 4.

| Category                  | Details          |
|---------------------------|------------------|
| CPU                       | Intel(R) Core(TM) i5-7200U |
| RAM                       | 4.00GB           |
| Programming language      | Python           |
| Version                   | Python 3.6.5     |
| IDE                       | Pycharm          |

5.4. Comparative Analysis of Experimental Results
The experiment sets the number of topics $A$ to [10, 100], and takes the step size of 2 for LDA topic extraction. The Perplexity indicator and the KM_SSVW indicator are calculated on the test set to determine the optimal number of topics.

5.4.1. Calculation of perplexity metrics. From the value of the confusion in Figure 2, when the number of topics $A = 65$, the confusion index of LDA is minimized, and the optimal number of topics is 65.

5.4.2. Calculation of the KM_SSVW indicator. In Figure 3, the keyword matching rate is about 80, the matching rate is the highest, and the number of classified topics is 30.
Figure 3. Mail preprocessing data.

Table 5 shows the 10 topics in the LDA topic classification. Through the subjective statistical value words, the highest ratio is 80%, and the number of classification topics is 30.

Table 5. Subject value words for manual statistics

| Topic | Subject word |
|-------|--------------|
| 1     | us, afghanistan, state, war, ok, military, government, new, would, people |
| 2     | taliban, labour, afghan, message, one, received, ashton, waldorf, that’s, try |
| 3     | us, would, time, new, one, people, policy, also, women, first |
| 4     | word, time, ill, woodward, hikers, hope, know, next, day, Tuesday |
| 5     | pm, re, senate, huma, sullivan, Sunday, abedin, jacob, may, office |
| 6     | pis, print, today, prepare, reach, thank, wing, fco, cdm, discuss |
| 7     | pls, print, clips, press, cameron, imagejpg, copies, doc, see, ireland |
| 8     | qaddr, talk, would, work, roger, traveling, business, emailed, unicef, anytime |
| 9     | call, tomorrow, send, get, also, note, email, letter, want, see |
| 10    | would, israel, obama, president, new, could, party, settlements, us, said |

According to the experimental results, the optimal number of topics calculated by the Perplexity indicator is 65. The optimal number of topics calculated by the KM_SSVV indicator is 30. The LDA is used to extract the subject of the mail data set and analyze the result. The results of some topic extractions are shown in Table 2 (only the top 10 topics are shown and the probability values of the keywords are omitted). Therefore, the KM_SSVV evaluation method can also find the optimal topic classification by modifying the number of subject classifications, which confirms the validity of KM_SSVVM.

6. Summary and Expectations

At present, in the era of big data, the analysis of data is very important, so the demand for algorithms for processing data is constantly increasing. Starting from the characteristics of LDA, this paper proposes that the application of LDA in the subject classification of emails must pay attention to the effects of theme extraction and the number of topics. Combining the TF-IDF algorithm and the subjective statistical value word idea, the method of determining the optimal number of topics is proposed. It is proved that in the mail data knowledge mining, this method can effectively determine the number of topics to obtain better subject extraction results, helping workers to A significant topic is extracted from the massive data. Although the validity of the method is proved in the subject classification of the mail, there is no verification of the method for other types of data sets, such as
microblog short text, user comments, WeChat public platform articles. In addition, in the case of big data, there are many subjective statistical uncertainties, so improving subjective statistics is the next step.

7. Acknowledgment
This work was partially supported by Joint Funding Project of Beijing Municipal Commission of Education and Beijing Natural Science Fund Committee (KZ201710015010), Project of National Scientific Found (No.61370188), Project of Beijing Municipal College Improvement Plan (PXM2017_014223_000063) and New Project of Green Printing and Publishing Technology by Cooperative Creating Center (PXM_014223_000025) and BIGC Project (Ec201803 Ed201802 Ea201806)

8. References
[1] Blei D M, Ng A Y, 2003 Jordan M I. Latent Dirichlet Allocation (Journal of Machine Learning Research) p 993-1022.
[2] Ma Jianhong, Fan Yuexiang 2018 Classification of technical video text based on improved Labeled LDA model (Computer Engineering) p 274-279.
[3] Zhao Le, Zhang Xingwang 2018 Research Progress and Trend of Text Classification for LDA Topic Model (Computer Systems) p 10-18.
[4] Yao Li, Zhang Yuhuang 2017 Application of Improved Random Forest Algorithm Based on Topic Model in Text Classification (Computer Applications and Software) p 173-178, 212.
[5] Sun Yanfeng 2017 Text Classification Based on Machine Learning (Beijing: Beijing University of Technology).
[6] Jiang Quan 2018 Mining and forecasting of Weibo theme based on topic model (Changchun: Changchun University of Technology).
[7] Wang Jingru, Chen Zhen 2018 A Comparative Study of Text Subject Extraction Based on Implicit Dirichlet Distribution (Information Science) p 102-107.
[8] Wang Ping 2011 Literature Knowledge Mining Based on Probabilistic Topic Model (Journal of the China Society for Scientific and Technical Information) p 583-590.
[9] Hassan S U, Haddawy P 2015 Analyzing Knowledge Flows of Scientific Literature Through Semantic Links: A Case Study in the Field of Energy (Scientometrics) p 33-46.
[10] Liang H, Fang L 2012 Topic Discovery and Trend Analysis in Scientific Literature Based on Topic Model (Journal of Chinese Information Processing) p 109-115.
[11] Gerrish S, Blei D M 2010 A Language-based Approach to Measuring Scholarly Impact (In: Proceedings of the 27th International Conference on Machine Learning)
[12] D. M. Blei, J. D. Lafferty 2006 Correlated topic models (Advances in Neural Information Processing Systems) p 147-154.
[13] Li Wenbo,Sun Le, Huang Ruihong et al 2008 A new text classification algorithm based on Labeled-LDA model (Computer Science) p 620-627.
[14] Yuan Bo qiu, Zhou Yi min, Li Lin 2009 LDA based feature selection for spam filter (Computer Engineering and Applications) p 121-124.
[15] Grossman D A 2004 Information Retrieval: Algorithms and Heuristics (Springer Science & Business Media).
[16] Salton G, Buckley C 1987 Term-weighting approaches in automatic text retrieval (Information Processing & Management) p 513-523.
[17] RehurekR. gensim 0.10.2 [EB/OL]. [2014-12-11]. https://pypi.python.org/pypi/gensim.
[18] Python3 split () method | novice tutorial [EB/OL]. http://www.runoob.com/python/att-string-split.html.