Article

A Method for Constructing Supervised Topic Model Based on Term Frequency-Inverse Topic Frequency

Zhinan Gou ^{1,*} , Zheng Huo ¹, Yuanzhen Liu ² and Yi Yang ²

¹ College of Information Technology, Hebei University of Economics and Business, Shijiazhuang 050061, China; huozheng@heuet.edu.cn

² College of Computer and Information, Hohai University, Nanjing 211100, China; liuyz_2000@163.com (Y.L.); yiyang2009@chnu.edu.cn (Y.Y.)

* Correspondence: gouzhinan@heuet.edu.cn

Received: 24 October 2019; Accepted: 2 December 2019; Published: 6 December 2019

Abstract: Supervised topic modeling has been successfully applied in the fields of document classification and tag recommendation in recent years. However, most existing models neglect the fact that topic terms have the ability to distinguish topics. In this paper, we propose a term frequency-inverse topic frequency (TF-ITF) method for constructing a supervised topic model, in which the weight of each topic term indicates the ability to distinguish topics. We conduct a series of experiments with not only the symmetric Dirichlet prior parameters but also the asymmetric Dirichlet prior parameters. Experimental results demonstrate that the result of introducing TF-ITF into a supervised topic model outperforms several state-of-the-art supervised topic models.

Keywords: topic model; supervised; term frequency; inverse topic frequency

1. Introduction

A topic model is a probabilistic generation model to find the representative topic of a document, with latent semantic analysis (LSA) [1], probabilistic latent semantic analysis (PLSA) [2] and latent Dirichlet allocation (LDA) [3] examples of classical topic models. However, these models are unsupervised, making the number and the content of topics difficult to control. As a result, a supervised topic model is proposed.

Supervised topic models have been successfully applied in the fields of tag recommendation [4,5] and document classification [6,7] in recent years. Furthermore, supervised topic models have been applied to various document-related tasks, and more successfully than unsupervised models. To the best of our knowledge, a supervised LDA was the prototype of the supervised topic model [8], which used variational methods to handle intractable posterior expectations. Since then, several supervised topic models have been proposed, e.g., discriminative LDA [9], a discriminative variation of LDA in which a class-dependent linear transformation is introduced in the topic mixture proportions. Maximum entropy discrimination LDA [10] utilizes the max-margin principle to train supervised topic models and estimate predictive topic representations that are arguably more suitable for prediction. The above methods train a single topic for documents.

It is well-known that a document with only one topic is inappropriate; for example, a document on sport medicine includes both sport and medical topics. This paper studies supervised topic models for multiple topics. Labeled LDA (L-LDA) [11] is a classical supervised topic model that matches the topics to the labels in the document. The number of topics is determined by the metadata of the document (such as labels), and topic terms have a better way to interpret topics. Partially labeled LDA [12] makes use of the learning machinery of topic models to discover the hidden topics within each label, as well as unlabeled, corpus-wide latent topics. Nonparametric labeled LDA [13] uses a

mixture of random measures as a base distribution of the hierarchical Dirichlet process framework with good effect. Dependency-LDA [14] further considers the label frequency and label dependency observed in the training data for constructing the supervised topic model.

Although supervised topic models have been studied extensively, most neglect the fact that topic terms have the ability to distinguish topics. Zou et al. [15] proved that a term weight with a more discriminative knowledge of topics is important.

In this paper, we propose a term frequency-inverse topic frequency (TF-ITF) method for constructing a supervised topic model, in which the weight of each topic term has the ability to distinguish topics.

The rest of the paper is organized as follows: the background knowledge of the research is reviewed in Section 2. Section 3 describes a term frequency-inverse topic frequency (TF-ITF) method for constructing a supervised topic model in detail. The experiments of our proposed method are conducted in Section 4. This paper is concluded in Section 5.

In order to make the proposed method easier to describe and understand, we summarize the notation needed in our formulation in Table 1.

Table 1. Notations description.

Notation	Description
w	Term
c	Topic
d	Document d
n	Number of terms in document d
N_i	The i th term of document d
$p(w)$	The term w probability in corpus
β	The Dirichlet prior parameter
η	The scale coefficient
γ	The smoothing factor
C	Number of topics in corpus
C_w	Number of topics for Term w
SN	The sampling number matrix of topic-term
Φ	The probability matrix of topic-term
Φ^{weight}	The weight matrix of topic-term
Φ'	The final generated probability matrix of topic-term

2. Background

In this section, we review some representative works on supervised topic models in detail, including L-LDA [11], Dependency-LDA [14] and CF-weight-LDA [15]. Then we analyze the limitations of the related works in the generation process of topic terms.

2.1. L-LDA (Labeled Latent Dirichlet Allocation)

L-LDA is an extension of the unsupervised LDA model. In order to contain the supervision, it applies a 1:1 correspondence between topics and labels. In addition to labels, keywords of scientific papers and categories of news are also considered as topics. L-LDA assumes that each document d is restricted to a multinomial distribution θ_d over labels which are part of the corpus. Each label is represented as a topic c that is a multinomial distribution Φ_c over the terms. The generative process for the algorithm is shown in Table 2.

It should be noted that it first samples a document topic distribution θ_d from a symmetric Dirichlet prior α_d of each document d .

The whole process of L-LDA contains topic constraints, which is different from LDA. More detailed descriptions of L-LDA are presented in the literature [11].

Table 2. Generative process of labeled latent Dirichlet allocation (L-LDA).

Steps	Description
1	For each topic $c \in C$:
2	Generate the multinomial topic distributions over vocabulary $\Phi_c \sim \text{Dirichlet}(\beta_c)$
3	End for
4	For each document $d \in D$:
5	Generate the multinomial topic distributions over document $\theta_d \sim \text{Dirichlet}(\alpha_d)$ according to the label set of each document d
6	For each term $W_{d,n} \in d$:
7	Generate topic $Z_{d,n} \sim \text{Multinomial}(\theta_d)$
8	Generate term $W_{d,n} \sim \text{Multinomial}(\Phi_{Z_{d,n}})$
9	End for
10	End for

2.2. Dependency-LDA (Dependency Latent Dirichlet Allocation)

Dependency-LDA further considers the label frequency and label dependency observed in the training data using an asymmetric document-label Dirichlet prior [15], as defined by Equation (1):

$$\alpha = \left[\eta_d \frac{N_{d,1}}{M_d} + \alpha_d, \eta_d \frac{N_{d,2}}{M_d} + \alpha_d, \dots, \eta_d \frac{N_{d,i}}{M_d} + \alpha_d, \dots, \eta_d \frac{N_{d,C}}{M_d} + \alpha_d \right] \quad (1)$$

The generative process for the algorithm is shown in Table 3.

Table 3. Generative process of Dependency-LDA.

Steps	Description
1	for each topic $c \in C$:
2	Generate the multinomial topic distributions over label $\Phi_c \sim \text{Dirichlet}(\beta')$
3	End for
4	for each label $l \in L$:
5	Generate the multinomial topic distributions over vocabulary $\varphi_l \sim \text{Dirichlet}(\beta)$
6	End for
7	for each document $d \in D$:
8	Generate the multinomial topic distributions over document $\theta_d' \sim \text{Dirichlet}(\alpha')$
9	for each label $i \in \{1, 2, \dots, M_d\}$:
10	Generate the topic $Z'_{d,i} \sim \text{Multinomial}(\theta_d')$
11	Generate the label $c'_{d,i} \sim \text{Multinomial}(\Phi_{Z'_{d,i}})$ according to the label set of each document d
12	End for
13	Compute the asymmetric Dirichlet prior α using Equation (1)
14	Generate the multinomial topic distributions over document $\theta_d \sim \text{Dirichlet}(\alpha)$
15	For each term $W_{d,n} \in d$:
16	Generate topic $Z_{d,n} \sim \text{Multinomial}(\theta_d)$
17	Generate term $W_{d,n} \sim \text{Multinomial}(\varphi_{Z_{d,n}})$
18	End for
19	End for

More detailed descriptions of Dependency-LDA are presented in the literature [14].

2.3. CF-Weight LDA (CF-Weight Latent Dirichlet Allocation)

Zou et al. [15] proved that a term weight with a more discriminative knowledge of topics is important. Each term is weighted by the corresponding CF-weight, which is based on the class frequency as follows:

$$\text{CF-Weight}_v = \eta \log\left(\frac{C}{CF_v}\right) + \gamma \quad (2)$$

where CF_v is the topic frequency of term v , i.e., the number of labels where term v has occurred in the training data; C is the total number of labels; η is the scale coefficient; and γ is the smoothing factor. A term with a higher (lower) class frequency will correspond to a smaller (larger) CF-weight; according to Equation (2), the model of CF-weight LDA is shown in Figure 1.

Figure 1. Model of CF-weight LDA.

In Figure 1, the two observable parts of the grey filled areas, c and $CF - W$, are explicit variables that indicate labels and weighted terms, respectively. Blank nodes are implicit variables and unobservable. The goal of CF-weight LDA model training is to estimate all C label-word distributions Φ . Given observed data $CF - W$, this is achieved by approximating the posterior distribution of latent variables $P(\Phi, \theta, z | CF - W, \alpha, \beta)$.

In the Gibbs sampler algorithm, the probability of term $w_{d,n}$ in the topic c is given by Equation (3):

$$P(Z_{d,n} = c | w_{d,n}, Z_d^{-n}, \alpha, \beta) \propto \frac{SN_{d,c}^{-n} + \alpha}{SN_d^{-n} + |y_d|\alpha} \cdot \frac{SN_{c,w_{d,n}}^{-n} + \beta}{SN_c^{-n} + V\beta} \quad (3)$$

where $SN_{d,c}$ and SN_d are the number of terms with CF-weights assigned to topic c and the total number of terms with CF-weights in document d , respectively; $SN_{c,w_{d,n}}$ and SN_c are the number of terms $w_{d,n}$ with CF-weights assigned to topic c and the total number of terms with CF-weights assigned to topic c , respectively; the superscript $-n$ is a quantity except for the term token in position n ; α and β are Dirichlet prior parameters, and y_d is the total number of labels of document d .

More detailed descriptions of CF-weight LDA are presented in the literature [15].

Although accomplishments have been achieved in existing research works for supervised topic models, unfortunately, two limitations exist:

- (1) The L-LDA and Dependency-LDA supervised topic models achieve correspondence between topics of document and multiple labels. The topic of each document is not the only one. However, they neglect the weight of terms by the topic frequency knowledge.
- (2) The CF-weight LDA supervised topic model uses the weight of terms with topic frequency knowledge. Nevertheless, the term frequency in the topic is neglected, and the calculating formula for the weight of terms reflects the poor ability of distinguishing topics without certain judgment conditions, for example, the perplexity value of the topic model.

To address the problems mentioned above, we propose a method for constructing a supervised topic model based on term frequency-inverse topic frequency to achieve a better topic model.

3. Proposed Method

The Gibbs sampler, a Markov chain Monte Carlo (MCMC) algorithm, which can approximately obtain a sequence of observations with a specified multivariate probability distribution, is used in this paper, since direct sampling is difficult [16]. The topic distribution of terms in each loop is updated in Gibbs sampling [17], which is more flexible compared with the maximum likelihood estimation method.

Firstly, each document is preprocessed through tokenization, lemmatization and deletion of unsuitable terms; secondly, the sampling number matrix of topic-term SN is generated by the Gibbs sampler. The principle of Gibbs sampling is simple, effective and easy to implement quickly, which is why Gibbs sampling is used to construct the topic model in this step.

When the perplexity value of the topic model tends to be stable, the sampling number matrix of topic-term SN is calculated and output. Based on SN , we can obtain the probability matrix of topic-term Φ . The probability of term v of topic c is $\Phi_{c,v}$:

$$\Phi_{c,v} = \frac{SN_{c,v} + \beta}{SN_c + V\beta} \quad (4)$$

where $SN_{c,v}$ is the number of terms v assigned to topic c ; SN_c is the total number of all terms assigned to topic c ; V is the total number of terms in the corpus; and β is the initial parameter.

The perplexity is a measurement of how well a probability distribution or probability model predicts a sample. It may be used to compare probability models. A low perplexity indicates the probability distribution is good at predicting the sample [18]. For a data set D , perplexity is defined by the following equation:

$$Perplexity(D) = \exp\left(-\frac{\sum_{d=1}^M \log p(w_d)}{\sum_{d=1}^M N_d}\right) \quad (5)$$

where M denotes the number of documents in the data set, w_d is a term, and N_d is the number of terms in a given document d [19].

When the perplexity value decreases and gradually becomes stable, the training of topic model is completed. The probability matrix of topic-term Φ can be calculated by Equation (4).

Next, we calculate the weight matrix of the topic-term Φ^{weight} . The calculation method is enlightened by TF-IDF (term frequency-inverse document frequency), a numerical statistic which can reflect how important a term is to a document in a corpus from information retrieval fields [20]. Similarly, in topic model fields, the importance of a term in a topic can be expressed by its weight. In this paper, we propose a term frequency-inverse topic frequency (TF-ITF) method for calculating the weight matrix of the topic-term Φ^{weight} . The calculation of weights is mainly in two parts:

- (1) Term frequency: Term frequency is the number of times the term has appeared in the topic. Higher values of the term frequency indicate that the term can represent the topic better.
- (2) Inverse topic frequency: Inverse topic frequency is a measure of the importance of a term. The less number of topics in which the term has occurred indicates the term can distinguish topics better.

Based on term frequency and inverse topic frequency, the equation of term frequency-inverse topic frequency is as follows:

$$\Phi_{c,v}^{weight} = \eta * \left[\frac{SN_{c,v}}{SN_c} * \log\left(\frac{C}{C_v}\right) \right] + \gamma \quad (6)$$

where $\Phi_{c,v}^{weight}$ is the weight of the term v in topic c ; $SN_{c,v}$ denotes the number of terms v occurring in the topic c ; SN_c is the total number of terms in topic c ; $SN_{c,v}/SN_c$ denotes the term frequency; C_v is the number of topics in which term v has occurred in the training data; C is the total number of topics; η is

the scale coefficient, where $\eta \geq 1$ generally and γ is the smoothing factor, where $\gamma \leq 0.01$ to avoid the situation that the weight is equal to 0.

The elements of Φ^{weight} and Φ are multiplied to generate the final probability matrix of topic-term Φ' as follows:

$$\Phi' = \Phi^{weight} \cdot \Phi \quad (7)$$

To sum up, the training steps of the proposed method are as follows:

- Step 1. Set the appropriate Dirichlet prior parameters α and β , and set the number of topics T according to the label set of the corpus.
- Step 2. Choose a topic at random for each term in the corpus.
- Step 3. Loop through each term in the corpus, and use Equation (3) to calculate the probability of each term in all topics.
- Step 4. Repeat step 3 until the Gibbs sampling algorithm converges, which means the perplexity value of documents is stable. Calculate the probability matrix of topic-term Φ by Equation (4).
- Step 5. Calculate the weight matrix of topic-term Φ^{weight} by Equation (6).
- Step 6. Generate the final probability matrix of topic-term Φ' by Equation (7).

4. Experimental Results and Analysis

In this section, we conduct experiments to verify the efficiency of the proposed method on a real dataset. Performance improvement of our approach is compared with state-of-the-art supervised topic models.

4.1. Experiment Environment

The experiments were executed on a personal computer with an AMD FX-Series CPU FX-8350 @4.0 GHz, the processor having eight physical computing cores, 16 GB DDR3 RAM memory @1600 MHz. The machine was running Windows7 on 64 bits, TensorFlow 1.4.0 with CPU support only and Python 3.6.

4.2. Dataset

To demonstrate the efficiency of our proposed method, we use a benchmark dataset: the Reuters Corpus-10,788 dataset is based on a large number of Reuters news reports and is widely used in research of natural language processing, recommendation systems, and information retrieval. The Reuters Corpus-10,788 dataset is obtained readily by the Python natural language processing toolkit. The dataset has 10,788 documents that contain about 1,300,000 terms and 90 topics. Based on a randomized approach, experimental data consists of 3000 documents from the Reuters Corpus-10,788 dataset.

4.3. Evaluation Metrics

The final generated probability matrix of topic-term Φ' is an important experimental result that is estimated by predicting the topics of test documents. For a test document, firstly, we construct the term frequency vector of the document, and then calculate the similarity between the test document and each topic of the final generated probability matrix of topic-term Φ' using a cosine measure in sequence. Finally, the output is sorted in descending order according to the similarity value. The hit ratio of the topics for a test document is the final prediction accuracy. The measurement is shown as follows.

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

where *Precision* is the precision of topics for a test document, *TP* is the number of true topics, and *FP* is the number of false topics.

For example, the topics of test document “test/18,609” are “grain, sugar, rice”, while our method predicted the first three topics of test document “test/18,609” as “sugar, money-fx, rice”, with a hit ratio of 66.7%. The final result is the average precision of topics in test documents, as defined by Equation (9):

$$Precision_{avg} = \frac{\sum_1^N Precision}{N} \quad (9)$$

where $Precision_{avg}$ is the average precision of topics for test documents, and N is the number of test documents.

It is important to note that the purpose of our method is to evaluate the effect of the supervised topic models by predicting the precision of the topics for the documents, so as to verify the explanatory ability of the supervised topic models.

4.4. Baseline Methods

To assess the efficiency of our proposed approach, three state-of-the-art methods of constructing supervised topic models were compared with our method. A classic supervised topic model L-LDA is proposed in [11], denoted by L-LDA₂₀₀₉, that uses a symmetric document-label Dirichlet prior. Dependency-LDA uses an asymmetric document-label Dirichlet prior in [14], denoted by D-LDA₂₀₁₂. On the basis of these two methods, we propose a term frequency-inverse topic frequency method for constructing supervised topic models, denoted by L-LDA_{TF-ITF} and D-LDA_{TF-ITF}. In addition, the importance of a term weight with more discriminative knowledge of the topic was proved for another method [15], based on L-LDA and denoted by CFW-LDA₂₀₁₈. All methods trained topic models using Gibbs sampling.

4.5. Comparison with Baseline Methods

In the process of experiment, this paper sets Dirichlet prior parameters α as 0.001 and β as 0.001 for all methods. The parameters of CF-weight LDA are the same as with our method: η is set to 50 and γ is set to 0.001 according to the existing research. Table 4 shows the experimental results in the Reuters Corpus-10,788 dataset.

Table 4. The experimental results.

Methods	Precision
L-LDA ₂₀₀₉	48.61%
D-LDA ₂₀₁₂	47.49%
CFW-LDA ₂₀₁₈	52.46%
L-LDA _{TF-ITF}	53.76%
D-LDA _{TF-ITF}	47.93%

As shown in Table 4, for the baseline method of supervised topic model with symmetric Dirichlet prior parameters, L-LDA₂₀₀₉, the corresponding improved method, L-LDA_{TF-ITF}, performs better than L-LDA₂₀₀₉ by 5.15%. At the same time, for the baseline method of supervised topic model with asymmetric Dirichlet prior parameters, D-LDA₂₀₁₂, the corresponding improvement method, D-LDA_{TF-ITF}, performs better than D-LDA₂₀₁₂ by 0.44%. The accuracy benefits from the weight of the topic term.

In addition, L-LDA_{TF-ITF} performs better than CFW-LDA₂₀₁₈ by 1.3% under the same conditions. CFW-LDA₂₀₁₈ uses the weight of terms by the topic frequency knowledge. Nevertheless, the term frequency of the topic is neglected, and the calculating formula for weight of terms reflects the poor ability to distinguish topics without certain judgment conditions. To summarize, it is verified that the method proposed in this paper trains a more effective supervised topic model that has better topic interpretation.

5. Conclusions

In this paper, we have highlighted that most supervised topic models neglect the use of terms with the significant ability to distinguish topics. In order to overcome this limitation, we introduce a term frequency-inverse topic frequency (TF-ITF) method for constructing supervised topic models, so that every topic term has a weight, thereby providing the ability to distinguish between topics. Experimental results demonstrate that the method for constructing supervised topic models based on TF-ITF outperforms several state-of-the-art supervised topic models.

As future work, we plan to at least further explore the following two directions. Although the Gibbs sampling technique is fast to implement and easy to understand, we intend to obtain a better topic model by using a variational autoencoder, which is a powerful technique for learning latent representations. We will also apply our constructed supervised topic model in information retrieval, recommender systems, text classification, and other fields, to achieve better performance.

Author Contributions: Conceptualization, Z.G. and Z.H.; Methodology, Z.G. and Y.L.; Writing—Original Draft Preparation, Z.G.; Writing—Review and Editing, Z.G. and Z.H.; Visualization, Y.Y.; Project Administration, Z.G. and Y.Y.

Funding: This research was partially funded by the Scientific Research Project of Hebei Education Department under Grant No.ZD2,018,205.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* **1990**, *41*, 391–407. [\[CrossRef\]](#)
2. Thomas, H. Probabilistic Latent Semantic Analysis. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 15–19 August 1999; pp. 50–57.
3. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
4. Wu, Y.; Xi, S.; Yao, Y.; Xu, F.; Tong, H.; Lu, J. Guiding Supervised Topic Modeling for Content based Tag Recommendation. *Neurocomputing* **2018**, *314*, 479–489. [\[CrossRef\]](#)
5. Ou, W.; Huynh, V. Joint Aspect Discovery, Sentiment Classification, Aspect-Level Ratings and Weights Approximation for Recommender Systems by Rating Supervised Latent Topic Model. *J. Adv. Comput. Intell. Intell. Inform.* **2018**, *22*, 17–26. [\[CrossRef\]](#)
6. Zhang, Y.; Ma, J.; Wang, Z.; Chen, B. LF-LDA: A Supervised Topic Model for Multi-label Documents Classification. *Int. J. Data Warehous.* **2018**, *14*, 18–36. [\[CrossRef\]](#)
7. Soleimani, H.; Miller, D.J. Semi-supervised Multi-label Topic Models for Document Classification and Sentence Labeling. In Proceedings of the 25th ACM International Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–28 October 2016; pp. 105–114.
8. Blei, D.M.; McAuliffe, J.D. Supervised Topic Models. In Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 3–6 December 2007; pp. 121–128.
9. Lacoste-Julien, S.; Sha, F.; Jordan, M.I. DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification. In Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–11 December 2008; pp. 897–904.
10. Zhu, J.; Ahmed, A.; Xing, E.P. MedLDA: Maximum Margin Supervised Topic Models. *J. Mach. Learn. Res.* **2012**, *13*, 2237–2278.
11. Ramage, D.; Hall, D.; Nallapati, R.; Manning, C.D. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multilabeled Corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009; pp. 248–256.
12. Ramage, D.; Manning, C.D.; Dumais, S. Partially Labeled Topic Models for Interpretable Text Mining. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 457–465.

13. Kim, D.; Kim, S.; Oh, A. Dirichlet Process with Mixed Random Measures: A Nonparametric Topic Model for Labeled Data. In Proceedings of the 29th International Conference on Machine Learning, Edinburgh, UK, 26 June–1 July 2012; pp. 675–682.
14. Rubin, T.N.; Chambers, A.; Smyth, P.; Steyvers, M. Statistical Topic Models for Multi-label Document Classification. *Mach. Learn.* **2012**, *88*, 157–208. [CrossRef]
15. Zou, Y.; OuYang, J.; Li, X. Supervised Topic Models with Weighted Words: Multi-label Document Classification. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 513–523. [CrossRef]
16. Gibbs Sampling-Wikipedia. Available online: https://en.wikipedia.org/wiki/Gibbs_sampling (accessed on 23 September 2019).
17. Griffiths, T.L.; Steyvers, M. Finding Scientific Topics. In Proceedings of the National Academy of Sciences, Irvine, CA, USA, 9–11 May 2004; pp. 5228–5235.
18. Perplexity-Wikipedia. Available online: <https://en.wikipedia.org/wiki/Perplexity> (accessed on 24 September 2019).
19. Miha, P.; Vili, P. Text Classification Method based on Self-Training and LDA Topic Models. *Expert Syst. Appl.* **2017**, *80*, 83–93.
20. Tf-idf-Wikipedia. Available online: <https://en.wikipedia.org/wiki/Tf-idf> (accessed on 29 August 2019).

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).