Text Representation Using Multi-level Latent Dirichlet Allocation

Amir H. Razavi and Diana Inkpen

School of Electrical Engineering and Computer Science
University of Ottawa
{araza082,diana}@eecs.uottawa.ca

Abstract. We introduce a novel text representation method to be applied on corpora containing short / medium length textual documents. The method applies Latent Dirichlet Allocation (LDA) on a corpus to infer its major topics, which will be used for document representation. The representation that we propose has multiple levels (granularities) by using different numbers of topics. We postulate that interpreting data in a more general space, with fewer dimensions, can improve the representation quality. Experimental results support the informative power of our multi-level representation vectors. We show that choosing the correct granularity of representation is an important aspect of text classification. We propose a multi-level representation, at different topical granularities, rather than choosing one level. The documents are represented by topical relevancy weights, in a low-dimensional vector representation. Finally, the proposed representation is applied to a text classification task using several well-known classification algorithms. We show that it leads to very good classification performance. Another advantage is that, with a small compromise on accuracy, our low-dimensional representation can be fed into many supervised or unsupervised machine learning algorithms that empirically cannot be applied on the conventional high-dimensional text representation methods.

Keywords: Latent Dirichlet Allocation (LDA), Text representation, Topic extraction, Text mining, Multilevel representation.

1 Introduction

For many years, classification of text data has been regarded as a practical and effective text mining task. In order to improve the performance of such an important task, we always need an informative and expressive method to represent the texts [18] [17]. In this regard, if we consider the words as the smallest informative units of a text, there is a variety of well-known quantitative information measures that can be used to represent a text. Such methods have been used in a variety of information extraction projects, and in many cases have even outperformed some syntax-based methods. There are a variety of Vector Space Modeling (VSM) methods which have been well explained and compared, for example in [20]. However, these kinds of representations disregard valuable knowledge that could be inferred by considering the different types of relations between the words. These major relations are actually the essential components that, at a higher level, could express concepts or explain the main topic of a

M. Sokolova and P. van Beek (Eds.): Canadian AI 2014, LNAI 8436, pp. 215–226, 2014.

text. A representation method which could add some kind of relations and dependencies to the raw information items, and illustrate the characteristics of a text in a more extensive manner, could play an important role in knowledge extraction, concept analysis and sentiment analysis tasks.

In this paper, the main focus is on how we represent the topics of the texts. Thus, we first introduce a LDA topic-based representation method as the selected approach, and in the second stage, we build a multi-level topic representation based on the first step. In the third stage, we run machine learning algorithms on a representation that combines various topical representation levels, in order to explore the most discriminative representation for the task of text classification.

2 Background and Related Work

In most text classification tasks, the text are represented as a set of independent units like unigrams / bag of words (BOW), bigrams, and/or multi-grams which construct the feature space, and the text is normally represented only by the assigned value (binary, frequency, or TF-IDF[1]), which is explicitly about the existence of the features in the text [19]. In this case, since most lexical features occur only a few times in each context, if at all, the representation vector tends to be very sparse. This method has two disadvantages. First, very similar contexts may be represented by different features in the space. Second, in short texts, we will have too many zero features for machine learning algorithms, including supervised classification methods.

Capturing the right sense of a word in its context is a critical issue in the representation methods. When we review the literature in this area, we find some useful hypotheses, such as: "You shall know a word by the company it keeps" [8], and that the meanings of words are (largely) determined by their distributional patterns; this is known as the Distributional Hypothesis [10] [11], which state that words which occur in similar contexts tend to be similar. There are many works about semantic similarity based on the Distributional Hypothesis [14].

In 2003, Blei, Ng and Jordan presented the Latent Dirichlet Allocation (LDA) model and a Variational Expectation-Maximization algorithm for training their model. These topic models are a kind of hierarchical Bayesian models of a corpus [2]. The model can unveil the main themes of a corpus, which can potentially be used to organize, search, and explore the documents of the corpus. In the LDA topic modeling, a "topic" is a distribution over the feature space of the corpus and each document can be represented by several topics with different weights. The number of topics and the proportion of vocabulary that create each topic are considered as two hidden variables of the model. The conditional distribution of these variables, given an observed set of documents, is regarded as the main challenge of the model.

Griffiths and Steyvers in 2004 applied a derivation of the Gibbs sampling algorithm for learning LDA models [9]. They showed that the extracted topics capture a meaningful structure of the data. The captured structure is consistent with the class labels assigned by the authors of the articles. The paper presents further applications

[1] Term frequency–inverse document frequency.

of this analysis, such as identifying "hot topics" by examining temporal dynamics and tagging some abstracts to help exploring the semantic content. Since then, the Gibbs sampling algorithm was shown as more efficient than other LDA training methods, e.g., variational EM and Expectation-Propagation [15]. This efficiency is attributed to a famous attribute of LDA namely, "the conjugacy between the Dirichlet distribution and the multinomial likelihood". This means that the conjugate prior is useful, since the posterior distribution is the same as the prior, and it makes inference feasible; therefore, when we are doing sampling, the posterior sampling becomes easier. Because of this, the Gibbs sampling algorithm was applied for inference in a variety of models which extend LDA [21], [7], [4], [3], [13].

Recently, Mimno et al. presented a hybrid algorithm for Bayesian topic modeling in which the main effort is to combine the efficiency of sparse Gibbs sampling with the scalability of online stochastic inference [16]. They used their algorithm to analyze a corpus that included 1.2 million books (33 billion words) with thousands of topics. They showed that their approach reduces the bias of variational inference and can be generalized by many Bayesian hidden-variable models.

3 Datasets

In order to have a proper evaluation on our multi-level LDA representation, we conducted experiments and evaluation on two well-known textual datasets which are publicly available and can be used and compared in the future. We needed to run experiments on topic/subject classified datasets. The main difference between the two selected datasets is the number of train / test data samples and the distribution of the topics in the data that let us to also compare the performance of the proposed method in two cases: in the first dataset we have a balanced distribution over the class labels, while in the second dataset the distribution is unbalanced over the same set of topic labels.

3.1 Reuters R8 Subset

The first dataset that we chose to run our experiments on was the well-known R8 subset of the Reuters-21578 collection (from UCI machine learning repository[2]), a typical text classification dataset benchmark. The source document collection was downloaded from the CSMining Group's datasets[3]. The data includes the 8 most frequent classes of Reuteres-21578; hence the topics that will be considered as class labels in our experiments are "acq, crude, earn, grain, interest, money, ship, and trade".

In order to follow the Sebastiani's convention [18], we also call these sets R8. In addition to R8, the R10 subset was used by some researchers and it contains 10 classes, as the name indicates). The only difference between R10 and R8 is that the classes "corn" and "wheat", which are intimately related to the class "grain" were removed. The distribution of documents per class in the R8 subset is shown in the Table 1.

---

[2] `http://archive.ics.uci.edu/ml/index.html`
[3] `http://csmining.org/index.php/data.html`

Table 1. Class distribution of training and test subsets for R8

Class	# of train documents	# of test documents	Total # of documents
Acq	1596	696	2292
Earn	2840	1083	3923
Grain	41	10	51
Interest	190	81	271
Money-fx	206	87	293
Ship	108	36	144
Trade	251	75	326
Crude	253	121	374
Total	5485	2189	7674

According to the numbers from Table 1, the baseline of any classification experiment over this dataset may be considered as 51%, for a trivial classifier that puts everything in the most frequent class, which is "earn".

3.2 Reuters Transcribed Subset

The second dataset that we chose for our experiments is the Reuters Transcribed Subset. This is a selection of 20 files from each of the 10 largest classes in the Reuters-21578 collection. The data includes 10 directories labeled by topic name, each containing 20 files of transcriptions in that class (except for the 'trade' directory, which has 21 files). The topics that will be considered as class labels in our experiments are "acq, crude, earn, grain, interest, money, ship and trade" (the "corn" and "wheat" classes were removed for consistency with our first dataset). Since the 8 class labels (topics) are distributed evenly over the Reuters Transcribed Subset data, the baseline for any classification experiment over this dataset may be considered as 12.5%.

4 Method

4.1 Preprocessing

In the preprocessing stage, initially, internet addresses and email addresses were filtered out. Then all the delimiters such as spaces, tabs or newline characters, in addition to some characters like: "\ \r : ( ) ` 1 2 3 4 5 6 7 8 9 0 \ ' , ; = \ [ ] ; / < > { } | ~ @ # $ \ % ^ & * _ + " have been removed from documents, whereas the expressive characters like: " - . ' ' ! ? " were kept. Punctuations (such as quotes, " ") could be useful for determining the scope of speaker's message. This step considerably reduces the size of feature space and prevents the system from dealing with a large number of unreal tokens[4] as features for our classifiers and LDA estimation/inferences.

[4] For example: "aaaaaaa", "buzzzzzzzzzzzzzzz" and "--------a".

Two types of stop-words removal were performed: static stop words removal and corpus based dynamically stop words removal. For the first one, we tokenized the documents individually to be passed to the static stop-word removal step that is based on an extensive list of stop-words which has been already collected specifically for the Reuter corpus. In the second one, additional stop words were determined based on their frequency, distribution and the tokenization strategy over the corpus (i.e., unigrams, bigrams, 3 or 4 grams). We removed tokens with very high frequency relative to the corpus size where those appear in every topical class (i.e., those are almost useless for the topic identification task).

The output of this stage passed to the stemming process through the Snowball stemming algorithm. The output of this stage was formatted for two different purposes; first, an ".arff" file to be used as our training/testing datasets for the classification task; second, a standard text file format ".txt" to be fed to the LDA topical estimation / inference modeling[5].

4.2 LDA Multi-level Topic Modeling

For our goal of topic extraction from the two Reuters subsets, we developed a method based on the original version of LDA presented in [2]. LDA is a generative probabilistic model of a corpus. The basic idea is that the documents are represented as a weighted relevancy vector over latent topics, where a topic is characterized by a distribution over words. We applied and modified the code originally written by Heinrich [12] based on the theoretical description of Gibbs Sampling. A remarkable attribute of the chosen method is that lets a word to participate in more than one topical subset, based on its different senses / usages in its context.

The R8 subset that we used for the LDA topical representation was already passed through the preparation and filtration processes (the pre-processing). In this way, each document is represented by a number of topics in which each topic contains a small number of words inside (i.e., each topic consists in a cluster of words); and each word can be assigned to more than one topic across the entire input data (e.g., polysemous words can be in more than one topic). Therefore, the LDA method assigns some clusters of words as topics, with different weights, for each document.

For example; the following is one topical cluster: {"investment", "success", "plan", "company", "organization", "rate", "market", "sale", "contract", "profit"} extracted by the LDA model estimation process. The number of topics and the number of words inside each topic are two parameters of the method that can be adjusted as needed. In this research, the number of words in each topic has been set to maximum 10 words in each cluster. We observed that increasing the number of words inside each topic decreases the consistency of the topical clusters and make them noisy. These topical clusters will be regarded as dimensions / features of a new vector space, to represent the corpus in a lower-dimensional space. We may have more than one cluster, in which each feature (word) in the feature space belongs to, with some degree of

[5] The details are included to help the research become replicable.

membership. Since the number of clusters is another parameter of the LDA algorithm, in the first level, we initially choose N = 256 as the number of topical clusters. Then, in order to avoid aggressive topical cluster merging, which may cause the loss of meaningful topics, we set the number of clusters to N = N / 2 in order to obtain more general clusters / topics, and we continued this process at the next levels, for more generalization, until the number of topic clusters reaches the number of expected classes.

By running the LDA topic estimation algorithm, we have a topical cluster membership distribution vector for each document in the corpus. This can be considered as a new representation of our documents in a space with N dimensions at each level. We applied exactly the same process to add another N / 2 dimensions (the number of topical clusters) for each document, and keep adding dimensions for the next levels, as long as N is greater than or equal the number of corpus classes (8 in our case). According to this procedure, the total number of topic-based representation levels is equal to six on our dataset (six levels).

The final step is to integrate all the extracted features in the six levels as one integrated topical representation of the corpus. This representation then will be compared with the initial BOW representation, and we will also combine the two representations (BOW and multi-level LDA features), in order to increase the discrimination power of the features for text classification task.

4.3 Text Topic Classification

As mentioned in Section 3, the first classification dataset consists of 5485 training and 2189 testing short / medium length documents listed in 8 categorical topics. For the two datasets, we initially applied the TF-IDF method which is a classic method that gives higher weights to terms that are frequent in a document, but rare in the whole corpus. For this representation, we also applied the Snowball stemming algorithm (in order to reduce the feature space). After removing stop-words and stemming, we obtained 17387 words as the BOW feature set of the R8 subset for the general topic classification task. For evaluation of our representation over small number of train/ test data (versus a large number of training and testing sets of the R8 dataset), a set of stop-words removed form of the TF-IDF based representation of Reuters Transcribed Subset was selected. For the BOW representation of the second subset we also applied Snowball stemming algorithm on the feature space which includes 5480 words.

As the second and complementary representation of our two datasets, we used the integrated topical representation vector of the documents calculated using the LDA technique, which produced 504 topical features for the 6 levels (256 + 128 + 64 + 32 + 16 + 8 = 504).

To conduct our empirical performance evaluation of a supervised machine learning algorithm, it is good to have two disjoint subsets: training and testing. Partitioned training and testing datasets can provide reliable results only when we have enough samples to split into large enough subsets for the training and testing processes. This was the case for the R8 subset, as shown in Table 1; and it was not the case for the Reuters transcribed subset. When we do not have enough instances to split into large enough training and testing sets, we evaluate our classification process based on stratified 10-fold cross-validations. This means that we split the entire dataset into 10 almost equal size and class distribution folds, then train a classifier 10 times on a different 9 fold integration of the entire 10-folds, and test it on the 10th one. We did this for both datasets.

Before we integrated the 6 levels of topical representations, we used them individually for our text classification task and noticed that the discriminative power of the individual levels are about 10 to 30 percentage points less than the corresponding BOW representation. In other words, replacing any level of topical representation decreases the classification accuracy, comparing to the BOW representation. We found that the level with 32 topical dimensions was the most discriminative level, but still about 10 percentage points lower than the BOW representation.

We will see in section 5 that the LDA multi-level topical representation solely in many cases is able to outperform that BOW representation. Note that the topical representation of the corpus is a relatively low-dimensional representation of the corpus compared to the BOW high-dimensional representation. This allows more machine learning algorithms to be used in real-world settings with about the same performance. However, the integration of our low-dimensional representation vectors with the conventional BOW representation can boost the classification accuracy in the high dimensional space.

We evaluated the BOW representation and the multi-level LDA representation separately, and then we integrated the two representations. When we integrated them into one representation, we obtain 17891 features (17387 words plus 504 topics) for the first subset and 5984 features (5480 words plus 504 topics) for the second subset.

As part of the supervised machine learning core of the system, we trained a variety of classifiers, in order to evaluate the benefits of the text representation models. As classifiers for our experiments, we chose Support Vector Machines (SVM) because of the usually high performance, Multinomial Naïve Bayes (NB), because of the good performance on text data, and Decision Trees (DT), since the learning model is in a human- comprehensible form.

5 Results and Discussion

We initially ran our selected classification algorithms on the three representations (BOW, topical based and the integrated one) over the R8 dataset. We found the Precision, Recall, F-measure, Accuracy, True Positive (TP) and False Positive (FP) rates (the most common and declarative evaluation measures recently used in most machine learning papers), and calculated their weighted average (Wtd. Avg.) value for our experiments. For example, the weighted average value of “Recall” is calculated by averaging the recall of each class value, weighted by the percentage of that value in the test-set. We conducted the classification evaluation by training on the training set and testing on the separate test set.

Table 2. Comparison of the classification evaluation measures for different representation methods on the split R8 data (5485 training and 2189 testing documents)

Evaluation measure → / Representation/ Classifier used ↓	TP Rate Avg.	FP Rate Avg.	Precision Avg.	Recall Avg.	F1-Measure Avg.	Accuracy %
BOW / SVM	0.933	0.028	0.93	0.933	0.929	93.33
LDA Topics / SVM	0.959	0.016	0.96	0.959	0.959	95.89
LDA+BOW / SVM	0.970	0.011	0.970	0.970	0.970	**97.03**
BOW / NB	0.952	0.013	0.956	0.952	0.952	95.20
LDA Topics / NB	0.946	0.017	0.944	0.946	0.944	94.61
LDA+BOW / NB	0.955	0.01	0.957	0.955	0.956	**95.52**
BOW / DT	0.915	0.037	0.914	0.915	0.915	91.54
LDA Topics / DT	0.918	0.031	0.92	0.918	0.918	91.78
LDA+BOW / DT	0.921	0.032	0.921	0.921	0.92	**92.10**

As a second scenario, we also trained and test the same set of classifiers using 10-fold cross-validation on the whole dataset, to check the stability of the results when training and testing sets are rotationally changed by stratified 10-fold cross-validation (this means that the classifier is trained on nine parts of the data and tested on the remaining part, then this is repeated 10 times for different splits, and the results are averaged over the 10 folds; this is repeated 10 times).

We performed experiments with the three classification algorithms (SVM, Multinomial NB, and DT), for each of the three representations, to check the stability the results. We changed the "Seed", which is a random parameter of the 10-fold cross-validation in order to avoid the accidental "over-fitting".

The evaluation measures calculated over the three representations for the R8 data set are shown in Tables 2 and 3. We report the rate of true positives, the rate of false positives, the precisions, recalls and F-measure averaged over the 8 classes, and the accuracy of the classification task.

According to the best of our knowledge, the accuracy of our integrated representation method on the Reuters R8 dataset is higher than any simple and combinatory representation method from related work, which reports accuracies of 88%-95% [6], [1], [22], while 96% was reached with SVM on a complex representation method based on kernel functions and Latent Semantic Indexing [5].

For the second subset, since the dataset only consisted of 161 short / medium length documents labeled with the 8 classes, we performed our evaluation process using the 10-fold cross-validation method. The calculated values for each of the evaluation measures are shown in the table 4.

We recall that in this dataset the class values are almost evenly distributed in all training and testing subsets (12.5% baseline). Similarly to the results in tables 1 and 2, each of the evaluation measures that appear in table 4 is a macro average of 8 class label values (one class label vs. the other) multiplied by the 10 folds of the 10-fold-cross-validation. For example, the Recall measure is the average of the Recalls values

Table 3. Comparison of the classification evaluation measures for different representation methods on entire R8 data, using 10 fold cross-validation

Evaluation measure → / Representation/ Classifier used ↓	TP Rate Avg.	FP Rate Avg.	Precision Avg.	Recall Avg.	F1-Measure Avg.	Accuracy %
BOW / SVM	0.947	0.021	0.947	0.947	0.946	94.67
LDA Topics / SVM	0.959	0.016	0.960	0.959	0.959	95.89
LDA+BOW / SVM	0.973	0.01	0.973	0.973	0.973	**97.29**
BOW / NB	0.949	0.015	0.951	0.949	0.950	94.91
LDA Topics / NB	0.926	0.035	0.926	0.926	0.922	92.57
LDA+BOW / NB	0.946	0.015	0.947	0.946	0.965	**94.59**
BOW / DT	0.904	0.04	0.904	0.904	0.947	90.40
LDA Topics / DT	0.917	0.034	0.918	0.917	0.917	91.73
LDA+BOW / DT	0.919	0.032	0.919	0.919	0.919	**91.88**

Table 4. Comparison of the classification evaluation measures for different representation methods on the Reuters Transcribed Subset, using 10 fold cross-validation

Evaluation measure → / Representation/ Classifier used ↓	TP Rate Avg.	FP Rate Avg.	Precision Avg.	Recall Avg.	F1-Measure Avg.	Accuracy %
BOW / SVM	0.580	0.043	0.572	0.582	0.562	58.11
LDA Topics / SVM	0.577	0.045	0.598	0.567	0.579	57.72
LDA+BOW / SVM	0.647	0.037	0.644	0.647	0.643	**64.65**
BOW / NB	0.562	0.049	0.552	0.562	0.542	56.21
LDA Topics / NB	0.538	0.051	0.568	0.537	0.559	54.31
LDA+BOW / NB	0.627	0.041	0.624	0.627	0.623	**62.68**
BOW / DT	0.516	0.054	0.536	0.514	0.517	51.74
LDA Topics / DT	0.565	0.042	0.558	0.551	0.545	56.72
LDA+BOW / DT	0.617	0.041	0.614	0.617	0.613	**61.68**

of the 8 classes, while for each class the value is the macro average of 10 Recall values calculated for the 10 runs of cross-validation. Since the distribution is balanced, averaging over 80 runs, the numbers tend to stay at the mean value of their range.

The best results were obtained with the SVM classifier for the integrated representation BOW and LDA Topics, achieving an accuracy of 97% (51% baseline) for the first data set and 65% (12.5% baseline) for the second subset. Although the small number of documents in each fold may increase the variance of the results from fold to fold, our results on the second subset also confirm the applicability of the presented method even for corpora with a small number of documents.

The improvement over the BOW representation is statistically significant, according to a paired t-test. It is known that the BOW representation is difficult to outperform in topic classification tasks. The fact that our integrated representation succeeded shows that the features that we added bring valuable semantic information.

6 Conclusions and Future Work

We designed and implemented a multi-level text representation method that we tested on the Reuters R8 dataset and on the Reuters Transcribed dataset. Our system applied LDA topical modeling estimation / inference for the topic-based representation purpose. The method was evaluated using Multinomial Naïve Base, SVM and Decision Tree classification algorithms.

We showed that the proposed method is not only useful for dimensionality reduction of the usual high-dimensional representations of the textual datasets (e.g., BOW) without compromising the performance, but also the performance of the classifiers reveals that integrating the proposed representation with the conventional BOW representation can improve the overall discrimination power of the classifiers. However, the quality of the topic-based representation potentially can even be boosted by using a larger textual background resource collected in the same domain in order to build the LDA models.

Our text classification method has the several advantages. In the LDA representation each document is represented by the LDA weighted membership distribution of the topical word clusters, with a classification performance almost similar to that of the BOW representation; hence any other high dimensional vector representation of any collection of documents can be also replaced by its LDA weighted membership distribution, in order to reduce the dimensionality and consequently to deal with the curse of dimensionality without compromising the classification performance. The lower-dimensional representation can be used for any supervised / unsupervised machine learning algorithm that cannot be applied on high-dimensional data.

The performance of the topical-based representation method via the LDA algorithm can simply be improved by adding a source of background data in the same domain.

One limitation of our method is that the current design is based on case insensitive text. The method could be developed based on case sensitive texts for more precise presentation treatment of named entities.

Other directions of future work are to use some resources such as “Wordnet Domain” in our method in order to improve the quality of topical groups extracted via LDA, and to compare the performance of the proposed classification method in informal / unstructured and formal / structured corpora.

References

1. Aggarwal, C.C., Zhao, P.: Towards graphical models for text processing. Knowledge Information Systems (2012), doi:10.1007/ s10115-012-0552-3
2. Blei, D.M., Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B.: Hierarchical topic models and the nested Chinese restaurant process. In: Proceedings of the Conference on Neutral Processing Information Systems, NIPS 2003 (2003)

3. Blei, D.M., McAulie, J.: Supervised topic models. In: Proceedings of the Conference on Neutral Processing Information Systems, NIPS 2007 (2007)
4. Blei, D.M., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
5. Cardoso-Cachopo, A., Oliveira, A.L.: Combining LSI with other Classifiers to Improve Accuracy of Single-label Text Categorization. In: Proceedings of the First European Workshop on Latent Semantic Analysis in Technology Enhanced Learning, EWLSATEL 2007 (2007)
6. Chen, Y.-L., Yu, T.-L.: News Classification based on experts' work knowledge. In: Proceedings of the 2nd International Conference on Networking and Information Technology IPCSIT 2011, vol. 17. IACSIT Press, Singapore (2011)
7. McCallum, A., Wang, X.: Topic and role discovery in social networks. In: Proceedings of IJCAI 2005 (2005)
8. Firth, J.R., et al.: Studies in Linguistic Analysis. A synopsis of linguistic theory, 1930-1955. Special volume of the Philological Society. Blackwell, Oxford (1957)
9. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences 101(suppl. 1), 5228–5235 (2004)
10. Harris, Z.: Distributional structure. In: Katz, J.J., Fodor, J.A. (eds.) The Philosophy of Linguistics. Oxford University Press, New York (1964)
11. Harris, Z.: Distributional structure. In: Katz, J.J. (ed.) The Philosophy of Linguistics, pp. 26–47. Oxford University Press (1985)
12. Heinrich, G.: Parameter estimation for text analysis. Technical Report (2004), For further information please refer to JGibbLDA at the following link: `http://jgibblda.sourceforge.net/`
13. Li, W., McCallum, A.: Pachinko allocation: Dag-structured mixture models of topic correlations. In: Proceedings of ICML 2006 (2006)
14. McDonald, S., Ramscar, M.: Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In: Proceedings of the 23rd Annual Conference of the Cognitive Science Society 2001 (2001)
15. Minka, T., Lafferty, J.: Expectation propagation for the generative aspect model. In: Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence, UAI 2002 (2002), `https://research.microsoft.com/minka/papers/aspect/minka-aspect.pdf`
16. Mimno, D., Hoffman, M., Blei, D.M.: Sparse stochastic inference for latent Dirichlet allocation. In: Proceedings of International Conference on Machine Learning, ICML 2012 (2012)
17. Pan, X., Assal, H.: Providing context for free text interpretation. In: Proceedings of Natural Language Processing and Knowledge Engineering, pp. 704–709 (2003)
18. Sebastiani, F.: Classification of text, automatic. In: Brown, K. (ed.) The Encyclopedia of Language and Linguistics, 2nd edn., vol. 14, pp. 457–462. Elsevier Science Publishers, Amsterdam (2006)
19. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28(1), 11–21 (1972)
20. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research (JAIR) 37, 141–188 (2010)

21. Wang, X., McCallum, A.: Topics over time: A non-Markov continuous-time model of topical trends. In: Proceedings of ACM SIGKDD conference on Knowledge Discovery and Data Mining, KDD 2006 (2006)
22. Yuan, M., Ouyang, Y.X., Xiong, Z.: A Text Categorization Method using Extended Vector Space Model by Frequent Term Sets. Journal of Information Science and Engineering 29, 99–114 (2013)