Question Tags or Text for Topic Modeling: Which is better

Abdel Nasser H. Zaied
Vice-dean for Education and Students Affairs,
College of Computers and Informatics,
Zagazig University, Egypt
nasserhr@zu.edu.eg, nasserhr@gmail.com

Abstract

Topic modelling is a probabilistic based statistical model used to find the latent topics that best depicts the content of the documents. Community Question Answering websites such as Quora, Stack Overflow and Yahoo! Answers have been prevalently in use, performs topic modeling as lot of queries pour in on daily basis which make it challenging to understand, summarize and synthesize the main topic of discussions. On these websites there are basically two sources of information that are available to analyze the key latent topics: questions text and tags. Questions are in textual format and tags are the keywords or tokens that are related to the question being asked which describes the content of the question. In past studies, most of the researchers have used question text for the purpose of topic modeling. It is still unclear why tag is not being considered for topic modeling. To combat this issue, this paper performs topic modeling using both question tags and text. The topic modeling based on tags has been compared with text based on two metrics namely coherence and perplexity. Experiment has been conducted on three real time datasets namely Artificial intelligence, Software Engineering and quantum computing from Stack exchange website. At high level tag-based topic modelling looked promising but closer observation revealed the opposite. It has been found that topic modeling using question text is preferable as topic modelling using tags collapses after a certain number of topics.

Introduction

Popularity of the internet has given rise to many online social communities question-answering websites. These community websites are very popular among the various categories of users. Community Question Answering websites such as Quora, Stack Overflow and Yahoo! Answers provide a common platform to all types of users to sharetheir knowledge, challenges, latest innovations and trends and discuss their queries or issues. There has been a significant growth in the user's participation in these websites which has resulted in exponential rise in data. It has become challenging to understand, summarize and synthesize the main topic of discussions. Analysis of this text can help in finding the hidden information about the nature of the user, developments in technology and trending topics invarious domains

Topic modelling is a very popular technique to identify the hidden topics in the text corpus [1]. Topic modelling describes each document as a probabilistic collection of topics and each topic as a probabilistic collection of different words. Topic modelling has been widely used in many application areas including research trends in software engineering [2][3], biomedical science [4], research papers [5], scientific literature [6], scientific publications [7], COVID research [8], citation recommendations [9], social networks, bioinformatics, mining sentiments [10]. Latent Dirichlet Allocation (LDA) is popularly used technique to perform topic modelling.

Relational Topic Models (RTM), "Pachinko Allocation model" (PAM), Biterm Topic Modelling (BTM) and Topic over Time (TOT) are other topic modelling techniques used in different context [11][24][25][26][27]. To perform topic modelling on these community Q&A websites, either questions which is an unstructured text or tags can be considered. Tags provide key information about the questions by specifying the domain, language and technology. In past studies, researchers have mostly used question text for analyzing the contents of the Q&A forum. It is still unclearwhy questions tags cannot be used for extraction of the latent topics. To combat this issue this paper has performed topic modelling on stack exchange, a popular Q&A website using both questions tags and text. It has been observed that topic modelling using tags approach collapses after a certain number of topics. The main research contributions of the paper have been summarized as follows:

Performed topic modelling using question tags and text.
LDA based topic modelling has been employed on three stack exchange websites namely Artificial Intelligence, Software Engineering and Quantum Computing.
Perplexity and Coherence score have been used to evaluate the topic modelling results.

The remaining paper is organized as follows: Section 2 details the related work. Section 3 presents the proposed methodology to perform topic modelling. Section 4 presents experimental results and discussion. The paper is concluded in Section 5.

Related work

Topic modelling has been an area of active research for programmers from many years. It has been successfully applied to various domains. Even though many topic models are available but still developers are using only limited number of the topic models to mine the trends in the software repositories. LDA is one of the most popular modelling methods used by the developers. Developers don't explore the values of different variable and without going into the details they prefer to work with the default parameter values for these models. Even though there is a probability that the exploration of parameter values can result in the better modelling results [2]. Relationship between various documents can be studied using the probabilistic topic model. This approach is helpful for topic modelling in large number of documents with significant data size [12]. Topic modelling has been successfully applied to the text corpus generated from the online news. Online news corpus is very massive and it can be analyzed to find the to track the latest news and as well as to observe the change in topics over time. The number of topics (number of news per day) are to be found out explicitly as number of topics changes every day. It is really helpful in observing the change in topic trend over time and to check the topic of the day [13].
Topic modelling can also be applied to find the connection between the collection of documents based on multiple factors. Different documents can be connected with each other based on these factors and can be evaluated using pseudo closure function on topics obtained from topic modelling [16]. Sometimes topic terms obtained from the topic modelling are not much related with each other. Quality of such topics can be enhanced by using the semantic information between the terms.
Normally the topic generated by topic modelling are not very cohesive in nature. Sometime topic also contains the words which are not semantically related with each other. Certain approaches have been suggested to improve the semantic relevance between the words in the topic. Topics cohesiveness can be enhanced by analyzing the output of
Non-Negative Matrix Factorization (NMF) to find the hidden factors which can be merged to enhance cohesiveness in topic. Semantic Topic Combination approach is basically a context sensitive approach which can be used even in cases where context is not provided [17]. Use of the fisher vector instead of simple fixed length vectors and Advanced Semantic Topic Combination (ASToC) also produces semantically enhances topics [18].
Topic modelling can also be applied to find the topics on the small text like tweets, headlines, news etc. Short text documents don't have the enough contextual knowledge to find out the topic. [14] proposed an approach based on semantic relation between the content of the short text to find the latent topic. SeaNMF is based on the semantic information and non-negative matrix factorization (NMF). Semantic information is discovered using skip-gram view. Anchor based topic modelling Archetypal LDA (A-LDA) can also be applied to the short text. Hashtags and tags provide key information about the context and the semantics. These words are used as a seed in the SeededLDA to find the topics in small texts. The increase in seed leads to coherent topics [15]. Initially it was assumed that short text being small in size contains only one topic, but now it is proved that a few numbers of topics can be related to the small text. GPU-DMM and GPU-PDMM models are based on Dirichlet Multinomial Mixture model.
It has been found from the past studies that researchers are frequently using text for the purpose of analyzing the topics of discussion in the Q&A forums. Questions tags also provides important information about the posted questions. It describes the content of the posted question. It is still unclear why question tags are not being considered for analyzing the latent topics. To combat this issue, we have analyzed and compared the LDA based topic modellingon question text and tags data of the questions posted on the Stack exchange website.

Proposed methodology

Figure 1 shows the framework used to perform topic modelling using questions tags and text of the questions posted on Q&A websites. The proposed framework consists of three major steps: Data Extraction, Data Preprocessing and Perform Topic Modelling.

Step 1: Data Extraction: Questions posted on Stack exchange websites are extracted. The three websites considered for experiment are Software Engineering (SE), Artificial Intelligence (AI) and Quantum Computing (QC). Questions body was separated from their tags. The question body contains the unstructured text which was further processed using step 2.

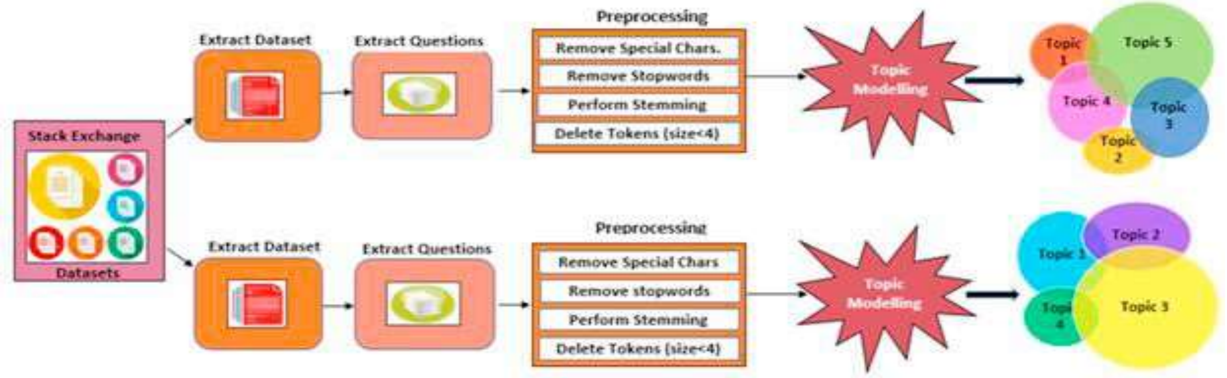

Fig. 1. Overview of Tags vs. topic modelling approach.

Step 2: Data Preprocessing: This step is initiated by cleaning the questions in textual format. Six pre-processing steps are applied to clean the initial questions text. English stop words like 'is', 'are', 'where' etc. which occurs frequently in the text and don't contribute any useful information are removed. The stop word list is provided by the Python nltk tool. Question text consists of various HTML tags and the URL of the websites. HTML tags are removed from the question. Only the words with length greater than three were kept in the corpus. Next, Lemmatization has been performed on the text to reduce the words to their root words using NTLK library wordnet lemmatizer. Thus, dictionary is created using these root words known as tokens. Dictionary contain unique words with key value pairs. From this dictionary, words with frequency less than three are deleted.

A bag of word is created from the dictionary. A bag of word indicates the words present in a specific document. A high frequency of a word is not a clear indication of the importance of the word in the dictionary. So, the importance of the words is evaluated using the TF_IDF approach.

In TF-IDF term frequency specifies the frequency of the word in a document. But high value of TF does not indicate more importance of the word in the document. For example, some words like "is", "the etc. may have high frequency in the document, but they will not contribute any valuable information to the topic. Value of IDF indicates if it is a common word or infrequent word. Common occurring words have low value of IDF while the infrequent words havehigher value of IDF.

Words having higher values of TFIDF will be more relevant to the document. They make topic more interpretable.

Once pre-processing gets completed, topic modelling is done using LDA.

Step 3: Topic Modelling on Text and tag: Topic modelling is performed separately on the body and tag part of the question. Topic modelling is performed by using Latent Dirichlet Model from the Genism's library. Number of topics in LDA are varied from 2 to 54 to get the most appropriate number of topics. Number of iterations for LDA are kept at 400 and alpha and beta are assigned default values.

LDA discovers the topics hidden in the documents using the generative process. LDA is a generative probabilistic model. Each document can be expressed as a probabilistic distribution of the topics identified from the corpus and each topic can be expressed as a collection of the different terms with different probability. Same term can be a part of different topics with different probabilities. LDA exploits the Bag of Word approach and the ordering of the document in the corpus and ordering of the terms in the topic is not important. Unlike the PLSA, LDA can predict the topics even for a new data that, was not a part of the corpus [1]. LDA discovers the latent topics. LDA represents the relationship between the hidden and observed variable. Documents of a corpus are the observed variable while the structure of the topic, distribution of words in a topic and distribution of a topics in a document is a hidden parameter. Let us assume a corpus having D number of documents with K number of topics and N be number of words in each document. Various notations used in algorithm are

Nomenclature

α	Probability distribution of topics present in each document
β	Probability distribution of words for each topic
θ	Probability distribution of topic
θd	Probability distribution of topic in document d
φk	Multinomial word Distribution for topic k
K	Number of latent topics
W	word in the topic
Dirk	Dirichlet Distribution for topic k

LDA Algorithm:

```
For each topic k $\epsilon$ K draw multinomial distribution $\varphi_k$ , for topic k ~ Dir ($\beta$)
For each document d $\epsilon$ [1, D] do
    Draw multinomial topic distribution $\theta$, for document d~Dir ($\alpha$)
    Draw document length N ~ Pois($\xi$)
    For each word w $\epsilon$ [1, N] in document d
        Draw topic $Z_{d,w}$ ~ Multinom($\theta_d$)
        Draw a word $word_{d,w}$ ~ Multinom($\varphi_z$)  p($word_{d, w}$ | $z_{d, w}$ , $\beta$)
```

1. Experimental results:

This section describes the dataset used, evaluation parameters and experimental results and its analysis.

1.1. Dataset details

Datasets of Software Engineering (SE), Artificial Intelligence (AI) and Quantum computing (QC) websites from the Stack exchange to conduct the experiment. Question text and the tag part has been extracted and separated to perform topic modelling. Table 1 presents the total number of tags, number of unique tags, total number of questions, number of unique tokens in questions dataset and total number of tokens in question data. The data for 5, 3 and 8 years has been considered for AI, QC and SE website for experimentation.

1.2. Experimental results

Experiment was conducted to evaluate the performance of topic modelling using question text and tag. Number of topics in the LDA are varied from two to fifty-four and the value of the coherence and perplexity is observed for all three datasets for question dataset and tags dataset. Figure 2 depicts the coherence score for AI, SE and QC datasets respectively. Tags data coherence score shows a gradual increase or decrease in values for all datasets but after a certain point there is abrupt increase in coherence values. Spike in coherence score is observed approximately after topic number thirty-five. Questions datasets shows poor coherence score as compared to tags data for all topic numbers except a few. It is clearly visible from figure 2 that for almost all values topics numbers except one, tags coherence score surpasses the coherence score of the questions. In AI maximum value of coherence score for questions is 0.53 while for the tags it is 0.75. In the coherence graph (Figure 2b) for Software engineering dataset maximum coherence score for questions is 0.57 and 0.84 for tags. But for majority of topic numbers coherence score for tags exceeds the coherence score for questions. For quantum computing dataset coherence score for tags is always greater than coherence score for text. For QC maximum coherence score for questions is 0.47 and 0.51 for tags.

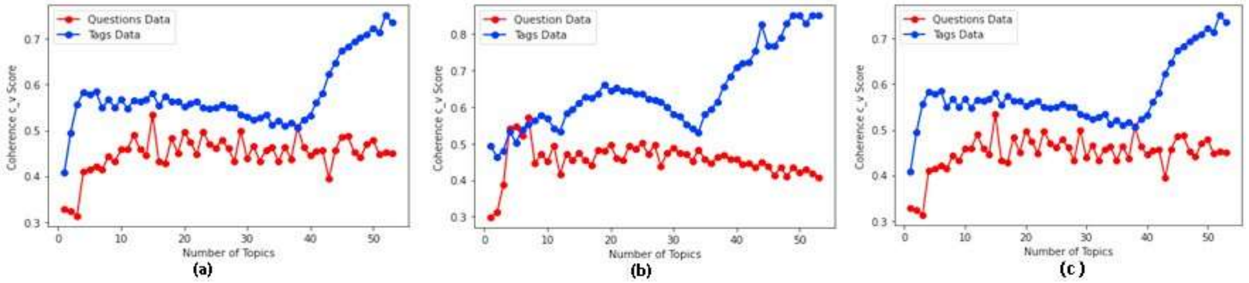

Fig. 2. Coherence c_v Score (a) AI; (b) SE; (c) QC.

Even though Coherence value graphs are giving an illusion that coherence using tags must be better. While examining the resultant topic patterns using both text and tags it has been observed that in case of tags only single topic was displayed for the topic number with the highest score. This pattern was observed in all the datasets. After a certain value of topics, all the topics will contain the same terms. For example, in case of AI dataset highest coherence using tags is obtained for topic number 43, same terms are observed for all topic numbers greater than 43. The model fails to produce different terms for topics 44 onwards. It indicates that the topic model based on tags fails after a certain number of topics.

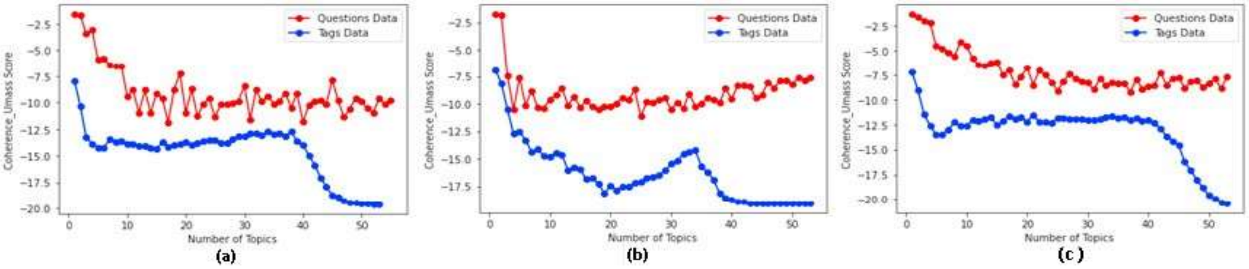

Fig. 3. Coherence Umass Score (a) AI; (b) SE; (c) QC.

For Umass coherence, a coherence value near to zero is desirable. Figure 3 (a, b, c) shows the Umass coherence graph for text and tags dataset for AI, SE and QC datasets. Lower coherence values are obtained for the tag data in all three data sets. Coherence value drops sharply after a certain number of topics. The crashing point for the c_v coherence and umass coherence is almost at the same topic number. A close observation of c_v graphs and umass graphs indicates that the crash in both graphs occurs almost at the same value of topic number. The AI topic model crashed after topic number 39. While questions data for all the datasets does not show such a drop in the values for umass coherence score. The similar results in c_v and umass confirms that the number of topics beyond this point does not correctly classify the dataset.

Graphically results indicates that tag-based topic modelling is better than text-based topic modelling. But abrupt depletion or abrupt jump was observed in c_v and umass coherence after certain topic number indicating that a further investigation is required. The reason behind the sudden depletion or rise in values of c_v and umass coherence were investigated by executing our datasets for k (topic number) beyond the crashing point. Experiment was conducted for various values of k beyond crashing point. It was observed that top ten topics from the tag have same term repeated in multiple topics. This occurrence of same term in multiple topics is leading to rise in value of coherence score. Because now almost all terms are same so it is giving very high coherence value. The top ten topics from the tag and text mining are shared in Table 2.

It is clearly visible from the topics obtained from the tag modelling that same term is repeated in multiple topics. Same term present in different topics is highlighted the using the same color. Perplexity is applied as the next metric to investigate the results.

Table 2. Top 10 Topics of Tag vs. text mining for k=45 for Artificial intelligence dataset.

Top 10 topics from Text modelling	Top 10 topics from Tag modelling
code, self, conv, model, activation, layer, loss, quot, import, dense	deeplearning, neuralnetworks, lossfunctions, machinelearning, regression, aidesign, classification, datasets, comparison, hopfieldnetwork
quot, depth, code, head, reference, model, project, strong, please,blog	
span, container, math, class, reward, frac, action, state, function,agent	
batch, mathbf, options, quot, class, span, weight, gradient, number, variable	
image, entropy, feature, discriminator, dataset, imgur, stack, cross,https, take	
policy, strong, move, theta, beta, action, state, game, target,alphazero	
latent, representations, fitness, autoencoders, accept, encoder, better , paper, reward, return	
pixel, smooth, production, image, classifications, visual, tackle,card, setup, dont	
service, collections, destination, cloud, present, past, minimum,progressive, days, systematic	
semantic, lookup, category, text, individual, entry, multiple, exclude , resource, single	

machinelearning, gradientdescent, neuralnetworks, backpropagation,deeplearning, theory, train, gameai, aidesign, gametheory

python, classification, rewardfunctions, neuralnetworks, deeplearning, keras, machinelearning, train, qlearning, aibasics

qlearning, neuralnetworks, relu, deeplearning, machinelearning, math,generativemodel, backpropagation, paper, python]

objectdetection, definitions, deeplearning, machinelearning, symbolicai,calculus, neuralnetworks, theory, math, python

train, datasets, neuralnetworks, gradientdescent, plan, bert, languagemodel,capacity, graphtheory, onlinelearning

comparison, machinelearning, hiddenlayers, neuralnetworks, topology,humanbrain, plan, languagemodel, graphtheory, hypothesisclass

tensorflow, deeplearning, neuralnetworks, keras, hillclimbing,backpropagation, localsearch, search, algorithm, gettingstarted

comparison, machinelearning, hiddenlayers, neuralnetworks, topology,humanbrain, plan, languagemodel, graphtheory, hypothesisclass comparison, machinelearning, hiddenlayers, neuralnetworks, topology,humanbrain, plan, languagemodel, graphtheory, hypothesisclass

Tag dataset is much smaller in size as compared to the text datasets and it contains only limited number of tokens. The limited count of number of tokens is giving rise to repetition of same term in multiple topics.

Perplexity value indicates how well a given probability distribution(model) can predict the given instance. A lower value indicates the goodness of a given model's prediction capability. Results of perplexity score for various datasets is shown in figure 4. Results of perplexity score are similar to the results of c_v and u_mass coherence. Till a specific number of topic numbers perplexity of tag dataset is better but after that a sudden drop in values is observed. For example. AI tag dataset perplexity value shows a sudden drop after topic value of 39 (Figure 4a). SE dataset (Figure 4b) and QC dataset (Figure 4c) again supports this phenomenon. The results of perplexity verified that fact that text-based topic modelling is giving stable results for all topic numbers, while tag results deteriorate after certain number of topics.

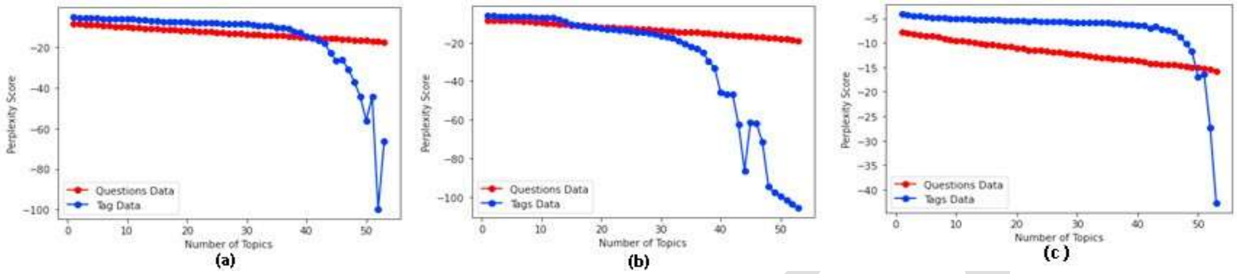

Fig. 4. Perplexity score (a) AI; (b) SE; (c) QC.

2. Conclusion

Community Q&A websites gets immense queries on daily basis which makes imperative to examine the prime themes and topics of discussion. LDA based topic modeling can help in discovering the coherent topics on these forums. This paper has investigated which is best source for topic modeling: Question tags or its text. Question text contains the real posted content whereas tags provide the key information about the content of the document. LDA model has been used to implement the topic modelling approach on three datasets. The results have been evaluated using three metrics: C_V and umass coherence score and perplexity. It has been observed that tag-based topic modelling approach performed better till a specific number of topics only, after which a sudden spike was observed in the coherence value and perplexity score. A closer look at the results revealed the flaw in tag-based modelling. Tag based dataset is very limited in size since it contains only tokens from tags of the questions, while the questions dataset contains the tokens from content of questions. Since tags dataset is limited in size so it is producing similar terms for topic numbers beyond a limit. On the other hand, question dataset contains considerable number of tokens and hence it has more choices for deciding the terms of a topic. Thus, it can be concluded that text-based topic modelling is better approach than tag-based topic modelling.

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003) "Latent Dirichlet Allocation." *Journal of machine Learning research 3*: 993–1022.

Panichella, Annibale, Bogdan Dit, Rocco Oliveto, Massimilano Di Penta, Denys Poshynanyk, and Andrea De Lucia (2013) "How to effectively use topic models for software engineering tasks? An approach based on Genetic Algorithms." *35th International conference onsoftware engineering (ICSE)*: 522-531.

Silva, Camila Costa, Matthias Galster, and Fabian Gilson (2021) "Topic modeling in software engineering research." *Empirical SoftwareEngineering* 26 (6):1- 62.

Lin, Xiaoguang, Mingxuan Liu, and Ju Zhang (2020) "A Top-Down Binary Hierarchical Topic Model for Biomedical Literature." *IEEEAccess* 8: 59870–59882.

Asmussen, Claus Boye, and Charles Møller (2019) "Smart literature review: a practical topic modelling approach to exploratory literaturereview." *Journal of Big Data* 6 (1): 1-18.

Chen, Hongshu, Ximeng Wang, Shirui Pan, and Fei Xiong (2019) "Identify topic relations in scientific literature using topic modeling."
IEEE Transactions on Engineering Management 68 (5): 1232-1244.

Älgå, Andreas, Oskar Eriksson, and Martin Nordberg (2020) "Analysis of scientific publications during the early phase of the COVID-19pandemic: Topic modeling study." *Journal of medical Internet research,* 22 (11): e21559.

Liu, Jiaying, Hansong Nie, Shihao Li, Xiangtai Chen, Huazhu Cao, Jing Ren, Ivan Lee, and Feng Xia (2021) "Tracing the Pace of COVID-19 Research: Topic Modeling and Evolution." *Big Data Research* 25: 100236.

Färber, Michael, and Adam Jatowt. (2020) "Citation recommendation: approaches and datasets." *International Journal on Digital Libraries*
21 (4): 375-405.

Sharma, Govind and M. Narasimha Murty (2011) "Mining sentiments from songs using latent dirichlet allocation." *International*
Symposium on Intelligent Data Analysis , Springer, Berlin, Heidelberg.

Chang, Jonathan, and David Blei (2009) "Relational Topic Models for Document Networks." *Artificial intelligence and statistics*, PMLR.

Hua, Ting, Chang-Tien Lu, Jaegul Choo, and Chandan K. Reddy (2020) "Probabilistic topic modeling for comparative analysis of document collections." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 14 (2): 1-2.

Xu, Guixian, Yueting Meng, Zhan Chen, Xiaoyu Qiu, Changzhi Wang, and Haishen Yao (2019) "Research on Topic Detection and Tracking for Online News Texts." *IEEE access* 7: 58407-58418.

Shi, Tian, Kyeongpil Kang, Jaegul Choo, and Chandan K. Reddy (2018) "Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations." *Proceedings of the 2018 World Wide Web Conference* :1105-1114.

Steuber, Florian, Mirco Schoenfeld, and Gabi Dreo Rodosek (2020) "Topic Modeling of Short Texts Using Anchor Words." *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics*: 210-219.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015) "Exploring the space of topic coherence measures." *Proceedings of the eighth ACM international conference on Web search and data mining*: 399-408.

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum (2011) "Optimizing semantic coherence in topic
models." *Proceedings of the 2011 conference on empirical methods in natural language processing*: 262-272.

Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin (2010) "Automatic evaluation of topic coherence." *The 2010 annual conference of the North American chapter of the association for computational linguistics* : 100-108.

Chen, Stanley F., Douglas Beeferman, and Roni Rosenfeld (1998) "Evaluation metrics for language models.", *Carnegie Mellon University. Journal contribution.*