The Number of Topics Optimization: Clustering Approach*

Fedor Krasnov[1][0000−0002−9881−7371] and Anastasiia Sen[2][0000−0002−1949−1642]

[1] Gazpromneft STC, 75-79 Moika River emb., Saint Petersburg, 190000, Russia
krasnov.fv@gazprom-neft.ru,
[2] Saint Petersburg State University, 7-9 Universitetskaya Emb.,
Saint Petersburg, 199034, Russia
anastasiia.sen@gmail.com

Abstract. Although topic models have been used to build clusters of documents for more than ten years, there is still a problem of choosing the optimal number of topics. The authors analyzed many fundamental studies undertaken on this subject in recent years. The main problem is the lack of a stable metric of the quality of topics obtained during the construction of the topic model. The authors analyzed the internal metrics of the topic model: Coherence, Contrast and Purity to determine the optimal number of topics and concluded that they are not applicable to solve this problem. The authors analyzed the approach to choosing the optimal number of topics based on the quality of the clusters. For this purpose, the authors considered the behavior of the cluster validation metrics: Davies Bouldin Index, Silhouette Coefficient, and Calinski-Harabaz.

The cornerstone of the proposed new method of determining the optimal number of topics based on the following principles:

- Setting up a topic model with additive regularization (ARTM) to separate noise topics;
- Using dense vector representation (GloVe, FastText, Word2Vec);
- Using a cosine measure for the distance in cluster metric that works better on vectors with large dimensions than Euclidean distance.

The methodology developed by the authors for obtaining the optimal number of topics was tested on the collection of scientific articles from the OnePetro library, selected by specific themes. The experiment showed that the method proposed by the authors allows assessing the optimal number of topics for the topic model built on a small collection of English-language documents.

Keywords: clustering & additive regularization topic model & validation metrics & Davies Bouldin Index & ARTM

1 Introduction

Topic models have been using successfully for clustering texts for many years. One of the most common approaches to topic modeling is the Latent Dirichlet Allocation (LDA) [10] which models a fixed number of topics selected as a parameter based on the Dirichlet distribution for words and documents. The result is a flat, soft probabilistic clustering of terms by topics and documents by topics. All the topics received are equal, they do not create any characteristic signs that could help the researcher to identify the most useful topics, that is, to choose a subset of topics that are best suited for human interpretation. The problem of finding the metric characterizing such interpretability is a subject of study by many researchers [44, 27, 41, 21].

The topic model is not able to read the insights of the researcher and therefore must have the settings for the task that the researcher is going to solve. According to studies [7, 1] topic models based on the LDA have the following parameters:

- α: the parameter of the prior Dirichlet distribution for "documents-topics";
- β: parameter of the prior Dirichlet distribution for "topics-words";
- tn: the number of topics;
- b: the number of discarded initial iterations according to Gibbs sampling;
- n: the number of samples;
- si: sampling interval.

In the recent study [1], published in 2018, an attempt was made to find the optimal values of the above parameters using the algorithm of *Differential Evolution* [42]. The authors chose a modified Jaccard Similarity metric as the cost-function. As a result, a new LDADE algorithm was created, in which free parameters from the Differential Evolution algorithm appeared and they also need to be optimized.

There is a difference between evaluating of a complete set of topics and evaluating individual topics to filter out unwanted information (noise). To evaluate a complete set of topics, researchers usually look at the *Perplexity* metric [2] for the corpus of documents.

This approach does not work very well according to the results of studies [45, 15] because the *Perplexity* does not have an absolute minimum, and with increasing of iterations it becomes asymptotic [25].

The most common use of *Perplexity* is to detect the "elbow effect", that is, when the pattern of growth in the orderliness of the model changes drastically. *Perplexity* depends on the power of the dictionary and the frequency distribution of words in the collection, hence we get its drawbacks:

– it cannot evaluate the quality of deletion of stop words and non-topic words;
– it cannot compare rarefying methods for dictionary;
– it cannot compare uni-gram and n-gram models.

The authors of the LDA made a study of the quality of topics using the Bayesian approach in [34]. It important to note that the Hierarchical Dirichlet

process (HDP) [43] solved the issue of the optimal number of topics, although it used not for documents, but for the whole collection.

Let us pay attention to the difference between the LDA, HDP, and hierarchical Latent Dirichlet Allocation (hLDA) [8, 9], since these are different topic models. LDA creates a flat, soft probabilistic clustering of terms by topic and documents by topic. In the HDP model, instead of a fixed number of topics for a document, the Dirichlet process generates the number of topics, which leads to the fact that the number of topics is also a random variable. The "hierarchical" part of the name belongs to another level added by the Dirichlet process, which creates several topics, and the topics themselves are still flat clusters. The hLDA model is an adaptation of the LDA, which models the topics as the distribution of a new, predetermined number of topics taken from the Dirichlet distribution. The hLDA model still considers the number of topics as a hyper parameter, that is, regardless of the data. The difference is that clustering is now hierarchical: the hLDA model studies the clustering of the first set of topics, providing more general abstract relationships between topics (and, therefore, words and documents). Note that all three models described (LDA, HDP, hLDA) add a new set of parameters that require optimization, as is noted in the study [13].

One of the main requirements for topic models is human interpretability [39]. In other words, whether the topics contain words that, according to a person's subjective judgments, are representative of a single coherent concept. In [35], Newman showed that the human assessment of interpretability well correlates with an automated quality measure called coherence.

The research [24] of 2018, proposed to minimize the Rényi and Tsallis entropies to find the optimal number of topics in the topic modeling. In this study, topic models derived from large collections of texts are considered as non-equilibrium complex systems, where the number of topics is considered as the equivalent of temperature. This allows us to calculate the free energy of such systems — the value through which the Renyi and Tsallis entropies are easily expressed. The metrics obtained based on entropy make it possible to find a minimum depending on the number of topics for large collections, but in practice we rarely find small collections of documents.

A study [6], published in 2018, proposed a matrix approach to improving the accuracy of determining topics without using optimization. On the other hand, the study [31] noted that increasing the accuracy of the model is contrary to human interpretability. In particular, the study [20], completed in 2018, created the VisArgue framework designed to visualize the model's learning process to determine the most explainable topics.

The use of the statistical measure *TF–IDF* as a metric for quantifying the quality of topics was studied in [37]. There is also a series of studies combining the advantages of topic models and dense representations of word-vectors [3, 30, 17, 36].

The motivation of the research conducted by the authors of this paper was the fact that the study of a stable metric for the quality of topics continues. Moreover, the use of cluster analysis is one of the tools for analyzing the stability

of topics [32] and the optimal number of topics [33], but it does not consider the benefits of the special training capabilities of the topic model with sequential regularization and dense representation of word-vectors.

To validate the quality of clusters, quite a lot of metrics have been developed. For example, Partition Coefficient [4], Dunn Index [19], as well as DPI [18] and its modifications [23, 47], Silhouette [40], which are involved in clustering algorithms. Nevertheless, in the case of a topic model, we already get clusters of topics and do not need a clustering algorithm, but only to evaluate the clusters obtained. For validation of clusters it is necessary to consider them in space with concepts of proximity and distance. For words, such a space is a vector representation of words. Significant results in this direction were obtained in researches [38, 11, 46]. Words presented in the form of dense vectors, reflect the semantic representation and have the properties of proximity and distance. Therefore, presenting the topics in the form of dense vectors, the authors created a new variation of the *DPI* metric for the topics, which the authors called *cDPI*.

The remainder of the paper is described as follows: the proposed methodology and research hypothesis are presented in Section 2; the results of testing a new quality metric are explained in Section 3. We conclude our paper in Section 4.

2 Research methodology

Consider ways to build a topic model for a specific collection of documents. Collection is homogeneous if it contains documents of the same type. For example, a collection of scientific articles from one conference, created on a single template, is homogeneous. In the case of a homogeneous collection of scientific articles, each document has a similar structure, postulated by a conference template. All scientific articles consist of introduction, presentation of research results and conclusion. Thus, it is possible to present a document in the form of a distribution of the main topic and auxiliary topics: introduction and conclusion.

Of course, the main topics in different documents may be different. However, the collection of scientific articles may be limited to the choice of certain headings from the thematic rubrics of the conference. Then the number of topics we will know. Figure 1 shows matrix distribution of topics on the documents.

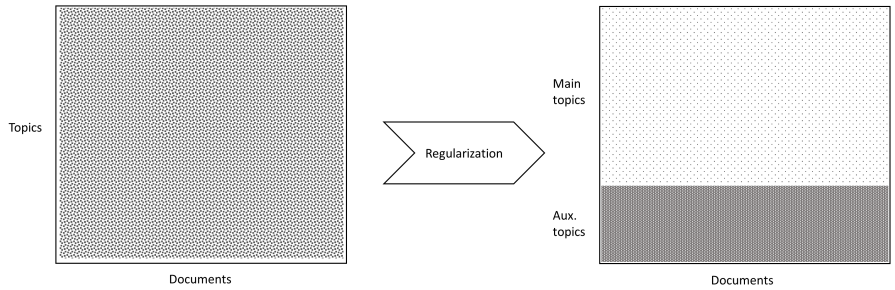

Fig. 1. "Topics-documents" scheme.

As we see on the left side of Figure 1, topic model leads to the emphasis of topics and their distribution homogeneously over the documents. Such a picture of the probabilities of the "topics-documents" matrix can be obtained using, e.g., models based on the LDA algorithm [10]. In addition, the right side of Figure 1 shows the result of the model with sequential ARTM [44]. The main and auxiliary topics are highlighted through the management of the learning process of the model. The principle of classifying a topic as auxiliary may be formulated as the existence of such a topic in the overwhelming number of documents. That is, the probabilities of the auxiliary topics will be distributed uniformly and tightly across the documents. Furthermore, the main topic will be a sparse vector for each document, since each document is characterized by one main topic.

We show that the existing internal metrics of the topic model are not suitable for determining the optimal number of topics. To do this, consider the internal automated metrics of the quality of topics. We introduce the concept of core topics:

$$W_t = \{w \in W \,|\, p\,(t|w) \geq threshold\}\,.$$

The following quality metrics of the topic model can be calculated based on the topics kernel:

- Purity of the topics : $Purity = \sum_{w \in W_t} p(w|t)$
- Size of the topic kernel : $|W_t|$
- Contrast of the topics: $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$
- Coherence of the topics: $Coh_t = \frac{2}{k(k-1)} \sum\limits_{i=1}^{k-1} \sum\limits_{j=1}^{k} PMI(w_i, w_j)$, where k is the interval in which the combined use of words is calculated, point-wise mutual information $PMI(w_i, w_j) = \log \frac{N \cdot N_{w_i w_j}}{N_{w_i} \cdot N_{w_j}}$, $N_{w_i w_j}$ – the number of documents in which words w_i and w_j appear in interval k at least once. N_{w_j} – the number of documents in which the word w_i appear at least once, and N is the number of words in the dictionary.

As can be seen from the formulas for the internal metrics of the topic model, each of these metrics can be measured for a different number of topics (tn). Consider the behavior of the metric *Kernel size* depending on the number of topics. With an increase in the number of topics, the core size will decrease, since the normalization conditions must be satisfied when constructing the matrices "topics-words" and "documents-topics": the sum of the probabilities must be equal to one. For metrics, the *Purity of topics* and the *Contrast of topics*, the nature of changes with an increase in the number of topics will also be monotonously decreasing, since the sum of the probabilities of the topics included in the core will decrease. On the other hand, for the metric, *Coherence to topics*, behavior with an increase in the number of topics will be monotonously increasing, as the contribution from PMI will grow. The specific nature of the changes in the metrics examined may vary; therefore it is advisable to try to find the extreme point using numerical methods, if it is possible.

The quality of the topics of short messages from the point of view of clusters was reviewed in [5] using NMF (Non-negative Matrix Factorization) and metrics reflecting the entropy of clusters. The matrix approach (LSI + SVD) to the selection of clusters of topics from the program code was investigated in [29] with a modified vector proximity metric. The research of the topic model's quality [33] use metric Silhouette Coefficient [40] with Euclidean distance for sparse subject vectors. Consequently, in these works, clusters in the space of dense vectors–words constituting topics and non-Euclidean distances in metrics remain unexplored.

In [16, 25, 22], the instability of topics with respect to the order of processed documents was discovered and investigated. Therefore, to calculate the quality metrics of the topics, it is necessary to perform calculations for the corpus of documents with a random order to eliminate the dependence on the order of documents. The possibility of stabilizing the topic model with the help of regularization was shown in [26]. Based on the analysis, the authors formulated a methodological framework, depicted as a diagram in Figure 2.

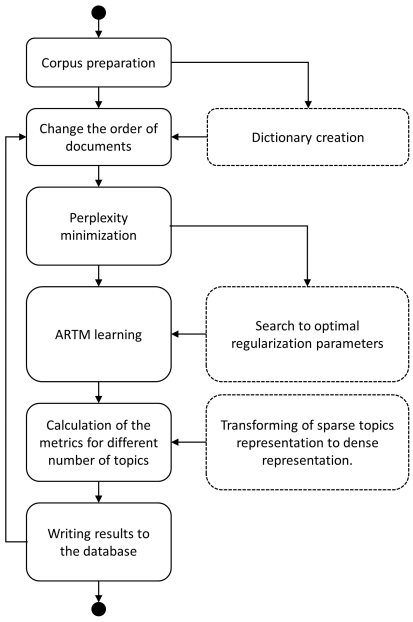

Fig. 2. Research framework.

Figure 2 shows the sequence of actions repeated for one corpus of documents a significant number of times, in order comparable to the number of documents in the corpus. On the right, actions that are performed only once are displayed: the formation of a dictionary, the adjustment of the regularization parameters of the topic model, and the transformation of the sparse presentation space of topics into a dense representation. Based on this methodological framework, digital experiments were developed and carried out as described in the next section.

3 Experiment

For the experiment was used corpus of scientific and technical articles on topics related to the development of oil and gas fields. In total, 1695 articles in English were selected in 10 areas of research according to the rubrics. The creation of a dictionary for the selected corpus is described in detail in the previous study by the authors [28]. To build a topic model, the BigARTM library was used, which allows for customization of the topic model by sequential regularization. The choice and adjustment of the regularization parameters of the topic model were made by the authors in a previous study [28]. To transform the sparse space of the vectors-words that make up the topics, the GloVe library was chosen [38]. To obtain a visual representation of the form of a dense representation of topics, a projection was made on a two-dimensional space with the distances preserved

using the MDS library [12]. Figure 3 presents the view of obtained clusters of topics.

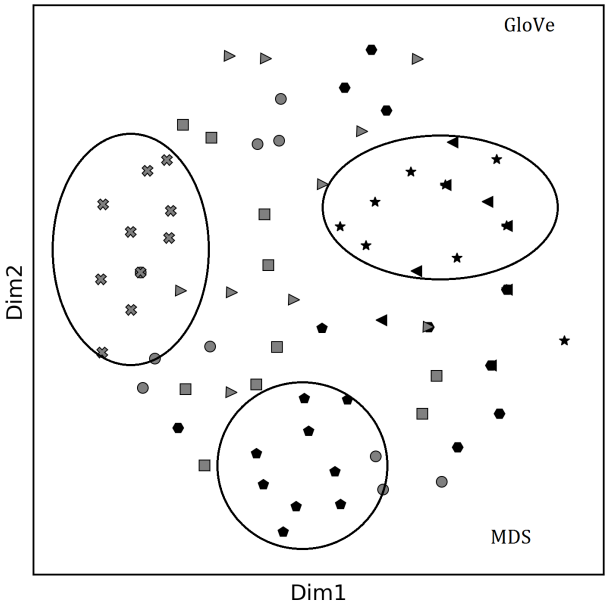

Fig. 3. Projection of a dense presentation of topics with preservation of distances.

In Figure 3, two-dimensional projections of words from topics are highlighted with different markers. Ovals emphasize precise visual grouping of words in the topics.

Figure 4 presents the preliminary calculations of main metrics behavior of the topic model, set up in accordance with the methodology proposed by the authors, depending on the number of topics.

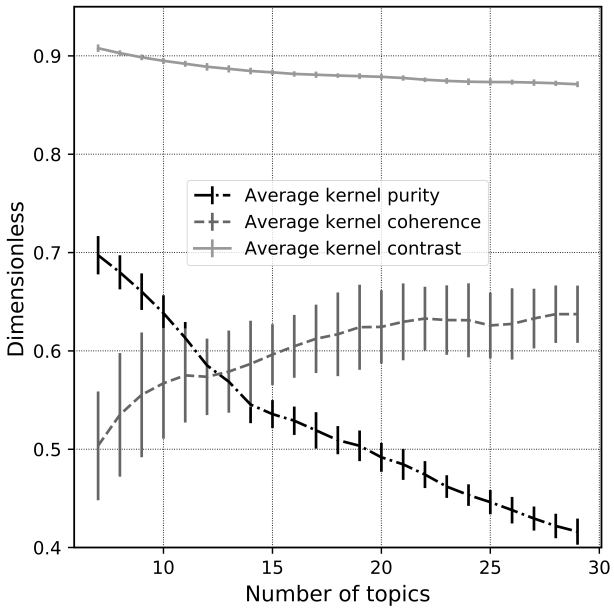

Fig. 4. Dependencies of the main internal metrics of the quality of the topic model on the number of topics.

As we can see from Figure 4, the nature of the dependencies is monotonous and does not allow to determine the optimal number of topics. Measurements of the main internal metrics are made for 1000 different random orders of documents. The y –axis represents the value of one standard deviation. Evidently that for the metric the *Contrast of the core*, the deviations are minimal. For metrics, *Purity* and *Coherence* of the core the greater values characterize the best quality of the topic model.

A characteristic point can be considered the number of topics equal to 12, when the curves of changes in the metric *Purity* and *Coherence* of topics intersect. Consider the dependencies of the following metrics: *Calinski-Harabaz Index* [14], *Silhouette Coefficient* [40], used to validate the number of clusters.

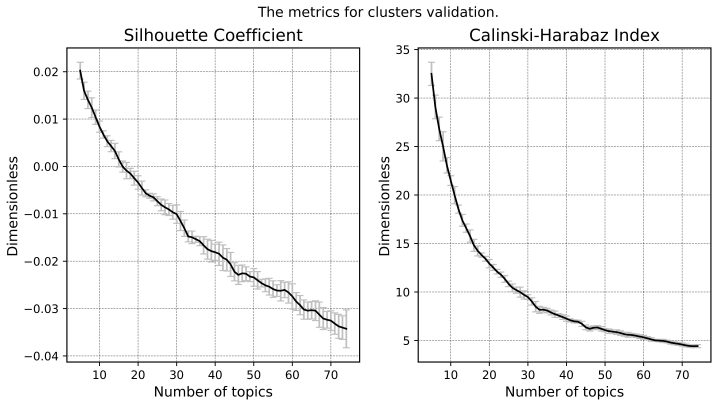

Fig. 5. Cluster Validation Metrics.

According to Figure 5, *the Calinski-Harabaz Index* and *Silhouette Coefficient* metrics do not make it possible to determine the optimal number of topics. As the number of topics increases, the values of these metrics decrease, which means that clusters become worse from the point of view of these metrics. The *cDBI* metric developed by the authors and shown in Figure 6 behaves differently depending on the number of topics.

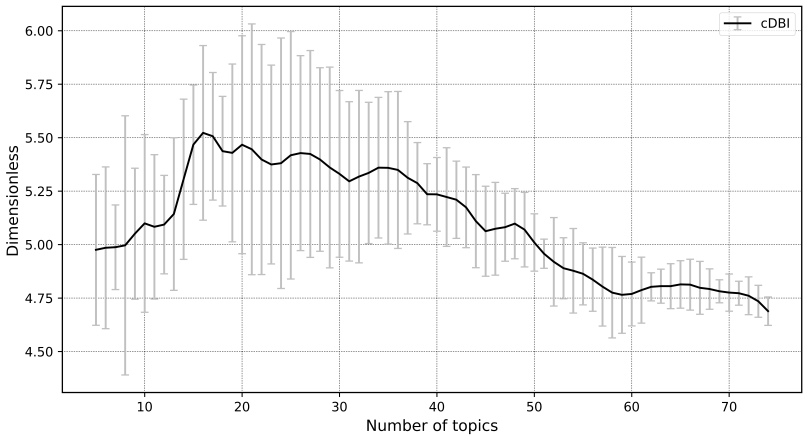

Fig. 6. *cDBI* metric.

In Figure 6 maximum clearly expressed with the number of topics equal to 16. The algorithm for calculating the *cDBI* metric is based on the ideology of the Davies Bouldin Index metric proposed in [18] and modified in [23, 47].

Result: $cDBI$
$V := GloVe(ARTM(tn, \mu, (corpus\,of\,texts))$
for $t \in W$: **do**
  $C_t := \sum_{i \in t} V_t^{(i)}$
  $D_t := \frac{1}{\dim t} \sum_{i \in t} \frac{C_t \cdot V_t^{(i)}}{|C_t| \cdot |V_t^{(i)}|}$
end
$cDBI := \frac{1}{\dim W} \sum_{t \in T} \frac{D_t}{C_t}$

Algorithm 1: Calculation of $cDBI$ metrics.

In the above Algorithm 1 T denotes the number of selected, μ – this regularizing coefficients. Thus, using the $cDBI$ metric, it is possible to find the optimal number of topics for a collection of documents.

4 Conclusions

The authors investigated the question of choosing the optimal number of topics for building a topic model for a given corpus of texts. The result of this study was a technique that allows you to determine the optimal number of topics for corpus of texts.

It should be said that the proposed method was experimentally confirmed under the following conditions:

- A small collection of documents;
- English language of documents (monolingual);
- Thematic uniformity.

An important methodological trick of the authors is the preparation of a topic model using sequential regularization. In previous studies of this collection of documents [28], the authors obtained numerical estimates of the coefficients for the regularizing components of the topic model (μ).

When forming a collection of texts, conditions were set that limited the number of topics of scientific articles according to the topic rubrics to 10. The essence of the experiment was to confirm the selected number of topics using an optimization approach based on the quality metric developed by the authors of the topic model — $cDBI$.

As a result, the experiment showed that the maximum value of the $cDBI$ metric for test corpus is achieved with the average number of topics equal to 16 with standard deviation 2. The result was obtained with a large number of model training to eliminate the influence of the order of documents in the collection.

In conclusion, it is important to emphasize that this study can serve as a methodological groundwork for the creation of software frameworks and proposes support for solving one of the fundamental problems of semantic text processing: determining the sense of a text fragment (article).

References

1. Agrawal, A., Fu, W., Menzies, T.: What is wrong with topic modeling? and how to fix it using search-based software engineering. Information and Software Technology **98**, 74–88 (jun 2018). https://doi.org/10.1016/j.infsof.2018.02.005, https://doi.org/10.1016/j.infsof.2018.02.005
2. Asuncion, A., Welling, M., Smyth, P., Teh, Y.W.: On smoothing and inference for topic models. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. pp. 27–34. UAI '09, AUAI Press, Arlington, Virginia, United States (2009), http://dl.acm.org/citation.cfm?id=1795114.1795118
3. Batmanghelich, K., Saeedi, A., Narasimhan, K., Gershman, S.: Nonparametric spherical topic modeling with word embeddings. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/p16-2087, https://doi.org/10.18653/v1/p16-2087
4. Bezdek, J.C.: Cluster validity with fuzzy sets. Journal of Cybernetics **3**(3), 58–73 (1973). https://doi.org/10.1080/01969727308546047, https://doi.org/10.1080/01969727308546047
5. Bicalho, P.V., d. O. Cunha, T., Mourao, F.H.J., Pappa, G.L., Meira, W.: Generating cohesive semantic topics from latent factors. In: 2014 Brazilian Conference on Intelligent Systems. pp. 271–276. IEEE (Oct 2014). https://doi.org/10.1109/bracis.2014.56, https://doi.org/10.1109/bracis.2014.56
6. Bing, X., Bunea, F., Wegkamp, M.H.: A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. CoRR **abs/1805.06837** (05 2018)
7. Binkley, D., Heinz, D., Lawrie, D., Overfelt, J.: Understanding lda in source code analysis. In: Proceedings of the 22Nd International Conference on Program Comprehension. pp. 26–36. ICPC 2014, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2597008.2597150, http://doi.acm.org/10.1145/2597008.2597150
8. Blei, D.M., Griffiths, T.L., Jordan, M.I.: The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. J. ACM **57**(2), 7:1–7:30 (feb 2010). https://doi.org/10.1145/1667053.1667056, http://doi.acm.org/10.1145/1667053.1667056
9. Blei, D.M., Jordan, M.I., Griffiths, T.L., Tenenbaum, J.B.: Hierarchical topic models and the nested chinese restaurant process. In: Proceedings of the 16th International Conference on Neural Information Processing Systems. pp. 17–24. NIPS'03, MIT Press, Cambridge, MA, USA (2003), http://dl.acm.org/citation.cfm?id=2981345.2981348
10. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (mar 2003)
11. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017), http://aclweb.org/anthology/Q17-1010
12. Borg, I., Groenen, P.: Modern multidimensional scaling: Theory and applications. Journal of Educational Measurement **40**(3), 277–280 (sep 2003). https://doi.org/10.1111/j.1745-3984.2003.tb01108.x, https://doi.org/10.1111/j.1745-3984.2003.tb01108.x
13. Bryant, M., Sudderth, E.B.: Truly nonparametric online variational inference for hierarchical dirichlet processes. In: Proceedings of the 25th In-

ternational Conference on Neural Information Processing Systems - Volume 2. pp. 2699–2707. NIPS'12, Curran Associates Inc., USA (2012), http://dl.acm.org/citation.cfm?id=2999325.2999436

14. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics **3**(1), 1–27 (1974). https://doi.org/10.1080/03610927408827101, https://www.tandfonline.com/doi/abs/10.1080/03610927408827101
15. Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Proceedings of the 22Nd International Conference on Neural Information Processing Systems. pp. 288–296. NIPS'09, Curran Associates Inc., USA (2009), http://dl.acm.org/citation.cfm?id=2984093.2984126
16. Chuang, J., Roberts, M.E., Stewart, B.M., Weiss, R., Tingley, D., Grimmer, J., Heer, J.: Topiccheck: Interactive alignment for assessing topic model stability. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 175–184. Association for Computational Linguistics (2015). https://doi.org/10.3115/v1/N15-1018, http://aclweb.org/anthology/N15-1018
17. Das, R., Zaheer, M., Dyer, C.: Gaussian lda for topic models with word embeddings. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 795–804. Association for Computational Linguistics (2015). https://doi.org/10.3115/v1/P15-1077, http://aclweb.org/anthology/P15-1077
18. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **1**(2), 224–227 (feb 1979). https://doi.org/10.1109/TPAMI.1979.4766909, http://dx.doi.org/10.1109/TPAMI.1979.4766909
19. Dunn†, J.C.: Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics **4**(1), 95–104 (jan 1974). https://doi.org/10.1080/01969727408546059, https://doi.org/10.1080/01969727408546059
20. El-Assady, M., Sevastjanova, R., Sperrle, F., Keim, D., Collins, C.: Progressive learning of topic modeling parameters: A visual analytics framework. IEEE Transactions on Visualization and Computer Graphics **24**(1), 382–391 (Jan 2018). https://doi.org/10.1109/TVCG.2017.2745080
21. Fang, D., Yang, H., Gao, B., Li, X.: Discovering research topics from library electronic references using latent dirichlet allocation. Library Hi Tech **36**(3), 400–410 (02 2018). https://doi.org/10.1108/LHT-06-2017-0132, https://app.dimensions.ai/details/publication/pub.1101114990
22. Greene, D., O'Callaghan, D., Cunningham, P.: How many topics? stability analysis for topic models. In: Machine Learning and Knowledge Discovery in Databases, pp. 498–513. Springer: Berlin, Heidelberg (2014)
23. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Clustering validity checking methods: Part ii. SIGMOD Rec. **31**(3), 19–27 (sep 2002). https://doi.org/10.1145/601858.601862, http://doi.acm.org/10.1145/601858.601862
24. Koltcov, S.: Application of rényi and tsallis entropies to topic modeling optimization. Physica A: Statistical Mechanics and its Applications **512**, 1192–1204 (dec 2018). https://doi.org/10.1016/j.physa.2018.08.050, https://doi.org/10.1016/j.physa.2018.08.050

25. Koltcov, S., Koltsova, O., Nikolenko, S.: Latent dirichlet allocation: Stability and applications to studies of user-generated content. In: Proceedings of the 2014 ACM Conference on Web Science. pp. 161–165. WebSci '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2615569.2615680, http://doi.acm.org/10.1145/2615569.2615680
26. Koltcov, S., Nikolenko, S.I., Koltsova, O., Filippov, V., Bodrunova, S.: Stable topic modeling with local density regularization. In: Internet Science, pp. 176–188. Springer International Publishing (2016)
27. Koltsov, S., Pashakhin, S., Dokuka, S.: A full-cycle methodology for news topic modeling and user feedback research. In: Staab, S., Koltsova, O., Ignatov, D.I. (eds.) Social Informatics. pp. 308–321. Springer International Publishing, Cham (2018)
28. Krasnov, F., Ushmaev, O.: Exploration of hidden research directions in oil and gas industry via full text analysis of onepetro digital library. International Journal of Open Information Technologies **6**(5), 7–14 (2018)
29. Kuhn, A., Ducasse, S., Gîrba, T.: Semantic clustering: Identifying topics in source code. Information and Software Technology **49**(3), 230–243 (mar 2007). https://doi.org/10.1016/j.infsof.2006.10.017, https://doi.org/10.1016/j.infsof.2006.10.017
30. Law, J., Zhuo, H.H., He, J., Rong, E.: LTSG: Latent topical skip-gram for mutually improving topic model and vector representations. In: Pattern Recognition and Computer Vision, pp. 375–387. Springer International Publishing (2018)
31. Lipton, Z.C.: The mythos of model interpretability. Queue **16**(3), 30:31–30:57 (jun 2018). https://doi.org/10.1145/3236386.3241340, http://doi.acm.org/10.1145/3236386.3241340
32. Mantyla, M.V., Claes, M., Farooq, U.: Measuring lda topic stability from clusters of replicated runs. In: Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. pp. 49:1–49:4. ESEM '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3239235.3267435, http://doi.acm.org/10.1145/3239235.3267435
33. Mehta, V., Caceres, R.S., Carter, K.M.: Evaluating topic quality using model clustering. In: 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). pp. 178–185 (Dec 2014). https://doi.org/10.1109/cidm.2014.7008665, https://doi.org/10.1109/cidm.2014.7008665
34. Mimno, D., Blei, D.: Bayesian checking for topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 227–237. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2145432.2145459
35. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 100–108. HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1857999.1858011
36. Nguyen, D.Q., Billingsley, R., Du, L., Johnson, M.: Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics **3**, 299–313 (2015), http://aclweb.org/anthology/Q15-1022
37. Nikolenko, S.I., Koltcov, S., Koltsova, O.: Topic modelling for qualitative studies. Journal of Information Science **43**(1), 88–102 (jul 2016). https://doi.org/10.1177/0165551515617393, https://doi.org/10.1177/0165551515617393

38. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics (2014). https://doi.org/10.3115/v1/D14-1162, http://aclweb.org/anthology/D14-1162
39. Rossetti, M., Stella, F., Zanker, M.: Towards explaining latent factors with topic models in collaborative recommender systems. In: 2013 24th International Workshop on Database and Expert Systems Applications. IEEE (09 2013). https://doi.org/10.1109/DEXA.2013.26, https://doi.org/10.1109/dexa.2013.26
40. Rousseeuw, P.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**(1), 53–65 (nov 1987). https://doi.org/10.1016/0377-0427(87)90125-7, http://dx.doi.org/10.1016/0377-0427(87)90125-7
41. Seroussi, Y., Bohnert, F., Zukerman, I.: Authorship attribution with author-aware topic models. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2. pp. 264–269. ACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012)
42. Storn, R., Price, K.: Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. J. of Global Optimization **11**(4), 341–359 (dec 1997). https://doi.org/10.1023/A:1008202821328, https://doi.org/10.1023/A:1008202821328
43. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Sharing clusters among related groups: Hierarchical dirichlet processes. In: Proceedings of the 17th International Conference on Neural Information Processing Systems. pp. 1385–1392. NIPS'04, MIT Press, Cambridge, MA, USA (2004), http://dl.acm.org/citation.cfm?id=2976040.2976214
44. Vorontsov, K., Potapenko, A., Plavin, A.: Additive regularization of topic models for topic selection and sparse factorization. In: Statistical Learning and Data Sciences, pp. 193–202. Springer International Publishing (2015)
45. Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D.: Evaluation methods for topic models. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 1105–1112. ICML '09, ACM, New York, NY, USA (2009). https://doi.org/10.1145/1553374.1553515, http://doi.acm.org/10.1145/1553374.1553515
46. Wu, L.Y., Fisch, A., Chopra, S., Adams, K., Bordes, A., Weston, J.: Starspace: Embed all the things! In: AAAI (2018)
47. Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. **13**(8), 841–847 (aug 1991). https://doi.org/10.1109/34.85677, http://dx.doi.org/10.1109/34.85677