Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2020"

Moscow, June 17–20, 2020

STABILITY OF TOPIC MODELING VIA MODALITY REGULARIZATION

Derbanosov R. (derbanosov@gmail.com)

National Research University Higher School of Economics, Moscow, Russia

Bakhanova M. (marybakhanova@gmail.com)

Skolkovo Institute of Science and Technology;
National Research University Higher School of Economics, Moscow, Russia

Probabilistic topic modeling is a tool for statistical text analysis that can give us information about the inner structure of a large corpus of documents. The most popular models—Probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation—produce topics in a form of discrete distributions over the set of all words of the corpus. They build topics using an iterative algorithm that starts from some random initialization and optimizes a loss function. One of the main problems of topic modeling is sensitivity to random initialization that means producing significantly different solutions from different initial points.

Several studies showed that side information about documents may improve the overall quality of a topic model. In this paper, we consider the use of additional information in the context of the stability problem. We represent auxiliary information as an additional modality and use BigARTM library in order to perform experiments on several text collections. We show that using side information as an additional modality improves topics stability without significant quality loss of the model.

Key words: topic modeling, topic modeling stability, artm, topic models regularization

DOI: 10.28995/2075-7182-2020-19-198-210

ПОВЫШЕНИЕ СТАБИЛЬНОСТИ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ С ПОМОЩЬЮ ДОПОЛНИТЕЛЬНОЙ МОДАЛЬНОСТИ

Дербаносов Р. (derbanosov@gmail.com)
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

Баханова М. (marybakhanova@gmail.com)
Сколковский институт науки и технологий;
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

1. Introduction

Topic modeling is a statistical method for analyzing a corpus of documents. The result of the modeling is a set of topics. Each topic is usually represented as a discrete distribution over the set of all words in the corpus. Some applications of topic modeling are information search [30], [31], [13], analysis of text documents [3], [25], [28], [29], images and video data [8], [11], [20], audio data [32], problems of bioinformatics [23], [24].

The most popular algorithms for topic modeling solve the task of stochastic matrix factorization i. e. approximate representation of a stochastic matrix F as a product of two stochastic matrices $F \approx \Phi\Theta$. Matrix F is obtained from the collection of texts by assigning $F[i,j]$ to the number of occurrences of i-th word in j-th document and column normalization. Matrix $F \in \mathbb{R}^{|W|\times|D|}$ is usually called word-document matrix, where $|W|$ is a number of words and $|D|$ is a number of documents in the corpus. Matrices $\Phi \in \mathbb{R}^{|W|\times|T|}$ and $\Theta \in \mathbb{R}^{|T|\times|D|}$ are called word-topic matrix and topic-document matrix, where $|T|$ is a number of topics that is usually fixed before run of the algorithm. If we fix some stochastic matrix factorization $F \approx \Phi\Theta$ we may interpret distributions in columns of the matrix Φ as topics.

Two most popular approaches to the topic modeling are *Probabilistic Latent Semantic Analysis* (PLSA) [10] and *Latent Dirichlet Allocation* (LDA) [3]. The basic hypothesis of the PLSA model is the conditional independence hypothesis: the probability of a word occurrence in a document is conditionally independent of the document given a topic. LDA is a Bayesian version of PLSA. The main assumption of the LDA model is that ϕ_{wt} and θ_{td} are generated from the Dirichlet distribution. *Additive Regularization of Topic Models* (ARTM) [25], [28], [14] extends the formulation of PLSA by adding different regularizers to the loss function. Some of them are described in **2.2**.

Usually algorithms use random initialization and then converge to some local optimum. One of the main problems of topic modeling is instability i. e. convergence

to different solutions from different initializations. Mathematical origins of this issue were studied in [6], [9], [18], [5] where authors research the problem of uniqueness of Nonnegative Matrix Factorization (NMF). Another approach to the problem is customization of basic algorithms to achieve better stability.

In the paper [2] the authors proposed ensemble methods and compared their performance with standard LDA and NMF approaches. The idea of their K-Fold method is to train several base topic models, transform them into the intermediate representation and build the final topic model on the top of this representation. According to experiments performed on annotated text corpora, K-Fold ensemble strategy can produce more stable and accurate topic models.

The authors of [15] proposed a modification of the standard latent Dirichlet allocation (LDA) model called granulated LDA (GLDA). The method is based on local density regularization that assigns the same topic with high probability to the words that meet together in the context. As for evaluation of model stability, the authors used Jaccard similarity and the number of stable topics based on Kullback-Leibler distance. The study shows that GLDA seems to reduce instability while yielding the same topic quality as classical topic models.

There are several studies [12], [33], [21] that show positive influence of additional information about documents in the collection on topic modeling performance. In this paper, we propose a method of increasing the stability that uses multimodal topic modeling. We use words as a first modality and different types of tags as additional modalities. We show that even using a partially labeled corpus (5% or 20% of the whole corpus) may increase the stability of PLSA model without significant loss of model quality.

2. Background

2.1. PLSA

Let D be a collection of documents, and let W be its vocabulary. The idea of probabilistic topic modeling is to describe how a collection of documents D is generated by a finite set of topics T. According to PLSA [10], the term distribution in each document $d \in D$ can be decomposed as a mixture of term probabilities for topics and topic probabilities for documents:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d), d \in D, w \in W, \tag{1}$$

where $\phi_{wt} = p(w|t)$ is the distribution of words in topics and $\theta_{td} = p(t|d)$ is the distribution of topics in documents. The parameters ϕ_{wt} and θ_{td} form stochastic matrices Φ and Θ. The problem of finding these matrices can be considered as an approximate matrix factorization task $F \approx \Phi\Theta, F = (\hat{p}_{wd})_{|W| \times |D|}$, where $\hat{p}_{wd} = n_{wd}/n_d$ is a frequency estimate of the conditional probability $p(w|d)$, n_d is the length of the document d, n_{wd} is the number of occurrences of the word w in the document d.

Parameters of the PLSA model are estimated via maximizing log-likelihood function with linear constraints:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in W} n_{wd} \log \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta} \tag{2}$$

$$\sum_{w \in W} \phi_{wt} = 1, \quad \phi_{wt} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geq 0. \tag{3}$$

The process of solving this problem consists of random initialization of the matrix Φ and application of EM algorithm.

Most Bayesian approaches, such as LDA, use a prior Dirichlet distribution as the main regularizer, thus complicate the combination with other regularizers. ARTM is a modern extension of PLSA model proposed in [25] that is free from excess probabilistic assumptions. It does not require parameters to be generated from Dirichlet distribution and allows to use different regularizers that may have no probabilistic interpretation at all. Suppose $R_i(\Phi, \Theta)$, $i = 1, 2, \ldots, n$ are n regularizers that we want to maximize along with the likelihood $L(\Phi, \Theta)$. In ARTM, we solve multi-criteria problem via maximization of the linear combination of L and R_i with some nonnegative regularization coefficients τ_i:

$$\begin{gathered} R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta), \\ L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}. \end{gathered} \tag{4}$$

Matrices Φ and Θ are estimated using EM algorithm, which can be described by two iteratively repeated steps.

At the E-step, we estimate the condition probability $p(t|d, w)$ for all words in documents (d, w) using Bayes formula:

$$p(t|d, w) = \frac{\phi_{wt} \theta_{td}}{\sum_{s \in T} \phi_{ws} \theta_{sd}}. \tag{5}$$

These probabilities are used to calculate parameters n_{wt}—the number of occurrences of the word w in the collection D with relation to the topic t and n_{td}—the number of words in the document d with relation to the topic t.

$$n_{wt} = \sum_{d \in D} n_{wd} p(t|d, w), \quad n_{td} = \sum_{w \in W} n_{wd} p(t|d, w). \tag{6}$$

At the M-step, we calculate parameters ϕ_{wt} and θ_{td} as frequency estimates of the corresponding conditional probabilities:

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt} \frac{\partial R(\Phi, \Theta)}{\partial \phi_{wt}} \right)_+, \quad \theta_{td} \propto \left(n_{td} + \theta_{td} \frac{\partial R(\Phi, \Theta)}{\partial \theta_{td}} \right)_+, \tag{7}$$

where the sign $\propto$ means that the distribution on the left is obtained after the normalization of the right expression, and $(x)_+ = \max\{x, 0\}$. Thus, we can add different regularizers to set necessary constrains to the topic model. In this work, we will use the following regularizers: smoothing, sparsing, decorrelation and modality.

2.2. Additive regularization of topic models

Smoothing regularizer. If we want ϕ_{wt} and θ_{td} to be close to some discrete distributions β_w and α_d in terms of Kullback–Leibler divergence we can use a smoothing regularizer:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max_{\Phi, \Theta}, \tag{8}$$

where β_0 and α_0 are regularization coefficients. Hence, the M-th step of the algorithm gives equations:

$$\phi_{wt} \propto (n_{wt} + \beta_0 \beta_w)_+, \quad \theta_{td} \propto (n_{td} + \alpha_0 \alpha_t)_+ . \tag{9}$$

It is recommended to use a prior Dirichlet distributions or Bayesian inference for distributions β_w and α_t. The effect of this regularizer is an increase in small values of ϕ_{wt} and θ_{td} due to a slight decrease in their large values. As a result, generated topics may include general vocabulary words, stop words and rare words that are usually excluded from topics.

Sparsing regularizer. Usually we assume that each word and each document relate to a small number of topics. It means that matrices Φ and Θ should be sparse. We can achieve it using a sparsing regularizer. One can notice that sparsing is an inverse procedure to smoothing. Therefore, sparsing and smoothing differ only in the sign of parameters β_w and α_t.

Decorrelation regularizer. Decorrelation regularizer formalises the requirement that topics have to differ from each other. It can be satisfied via minimizing the sum of covariances between distributions ϕ_{wt} and ϕ_{ws} for all pairs of topics t, s:

$$R(\Phi, \Theta) = -\tau \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max_{\Phi, \Theta}, \tag{10}$$

where τ is a regularization coefficient. In this case, the formula for the regularized M-step takes the form:

$$\phi_{wt} \propto (n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws})_+. \tag{11}$$

2.3. Multimodal topic modeling

Usually documents can be described not only by words but also by terms of other *modalities* [27]. For example, textual modalities are tags, n-grams, named entities and natural language words. The last one is what we used to deal with in topic modeling. Pictures and web-sites are non-textual modalities. We can consider documents as a set of tokens taken from different modalities. The diverse meta-data represented by modalities can be helpful for determining topics, and, vice-versa, topics may be used to predict missing meta-data.

Multimodal topic modeling occurred to be an effective approach for solving different problems. For example, for a given parallel collection of text translation we can model topics and then use them for the cross-language search. In this case, each language is considered as a modality. Experiments showed that the combination of parallel documents and bilingual dictionaries improves the quality of cross-language

search in comparison with models using only bilingual dictionaries [7]. Also, multimodal topic model can be applied for constructing recommendations [27]. In this study, the authors focused on the article recommendation in the online-platform, they used different modalities, such as words from text, user's feedback, tags, authors and user-specified categories. According to the results, the combination of modalities reasonably improves recommendation ranking.

Multimodal topic model and the regularized EM-algorithm for this case were firstly introduced in [27].

Let M be a set of modalities, and let W_m, $m \in M$ be a vocabulary of modality m. These vocabularies do not intersect and can be united into the set $W = \bigsqcup_{m \in M} W_m$ containing terms of all modalities. A model of $p(w|d)$ is introduced for each modality $W_m, m \in M$:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d), \ d \in D, \ w \in W_m. \tag{12}$$

The main concept of such modeling is that topics $p(t|d)$ are the same for all modalities. As for the distribution of words in topics, the matrices $\Phi_m = (\phi_{wt})_{|Wm| \times |T|}$ are normalized separately and stacked vertically into the matrix $\Phi = (\phi_{wt})_{|W| \times |T|}$.

If we consider the log-likelihood of each modality as a regularizer with coefficient τ_m, then the optimization problem has the following form:

$$\sum_{m \in M} \sum_{d \in D} \sum_{w \in W_m} \tau_m n_{wd} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}, \tag{13}$$

$$\sum_{w \in W_m} \phi_{wt} = 1, \quad \phi_{wt} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geq 0. \tag{14}$$

3. Metrics

3.1. Quality of topic modeling

There are several metrics for measuring quality of topic model. Most previous works have exlusively focused on perplexity measure that describes the speed and level of convergence of the model. Perplexity can be represented as an inverse function of the likelihood of model parameters. One of the drawbacks of this metric is that it depends on the data size, therefore, it is hard to compare results of this measure obtained from models trained on different datasets.

Some recent studies [15], [1] evaluate their models by pairwise information based metric called *coherence* [22]. The practical meaning of coherence follows a simple idea: if we describe the topic as a set of words then these words are likely to meet together in the context. In addition, coherence seems to reflect well the interpretability of topics [1]. Let k be an adjustable parameter meaning the number of top words in the topic $t \in T$, and let $W^t = \{w_1, \ldots, w_k\}$ be the corresponding set of top words. Then coherence formula for topic t is defined as follows:

$$C_t = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \left(\log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \right)_+, \tag{15}$$

where probabilities can be estimated by frequencies:

$$p(u, v) = \frac{n(u, v)}{n}, \quad p(u) = \frac{n(u)}{n}, \tag{16}$$

$$n(u) = \sum_{w \in W} n(u, w), \quad n = \sum_{w \in W} n(w). \tag{17}$$

There are different types of calculating co-occurrencies $n(u, v)$. In this paper, we calculate in how many documents the pair (u, v) occurred at least once:

$$n(u, v) = \sum_{d=1}^{|D|} [\exists\ i, j : w_{id} = u, w_{jd} = v]. \tag{18}$$

We have described above how to compute coherency only for one topic. To obtain the coherence score for the topic model we simply average coherencies for all topics in the model. The higher coherency is, the better.

3.2. Stability of topic modeling

Let's denote by $\{M_1, M_2, \ldots, M_r\}$ the set of topic models generated as a result of r runs of the algorithm on the same data. Assume that these models are similar if their topics are similar. To measure similarity between two topics represented by t top words, we propose to calculate the measure that we call *Stable Words* (SW), and describe it by the following formula:

$$SW(R_i, R_j) = \frac{|R_i \cap R_j|}{t}, \tag{19}$$

where R_i is a list of top t tokens of topic i. SW takes values in $[0, 1]$, and the value 1 corresponds to the identical top words. SW can be interpreted as a modified Jaccard Index, these two metrics differ only by the denominator: in Jaccard Index we divide by the set union size $|R_i \cup R_j|$. We consider SW as more interpretable measure in terms of topic stability because it is simply the share of stable words.

We should find topic correspondence between two sets of topics in order to compute similarity of these two sets. The best topic matching between two models with $|T|$ topics can be found using SW defined in Eq. 20. We construct a matrix S, where the element s_{ij} represents similarity between the i-th topic of the first model and the j-th topic of the second model. Then we find the optimal matching P by solving the minimal weight bipartite matching problem applying the Hungarian algorithm [16].

To obtain the score of similarity between the set of r models we compute *Average Stable Words* (ASW):

$$ASW = \frac{2}{r(r-1)} \sum_{i \le j, i \ne j} \frac{1}{|T|} \sum_{s=1}^{|T|} SW(R_{is}, R_{j\pi(s)}), \tag{20}$$

where $\pi(s)$ is a topic of the model j matched to the topic s of the model i.

4. Experiments

We performed experiments on five texts collections: 20NewsGroups, Reuters52, Cade, WebKB and Habr. 20NewsGroups [17] is a set of documents classified in 20 newsgroups. Reuters52 [19] is a collection of articles of 1987 year from Reuters that was manually classified by Reuters Ltd. The documents in WebKB[1] are webpages collected by the World Wide Knowledge Base project of the CMU text learning group. Cade is a subset of web pages extracted from the CADÊ Web Directory which points to Brazilian webpages labeled by human experts [4]. Habr is a dataset of articles from IT blogging platform http://habrahabr.ru [27] with 5 modalities: text of the blogpost, author, users that leave comments at the blogpost, hub that is a site section, tags that are generated by the author. We used preprocessing from [4] for 20NewsGroups, Reuters52, Cade and WebKB datasets. All datasets were splited by train and test sets in a ratio of 60 to 40. Coherence was measured on hold-out test dataset.

We take text labels as an additional modality for 20NewsGroups, Reuters52, Cade and WebKB datasets. Each document contains one token of such a modality. We use different percentage of labeled documents: 5%, 20%, 50% and 100% in order to simulate partially labeled collection. We tested tags, habs modalities and the combination of four modalities: authors, tags, hubs and users on Habr dataset.

All experiments were performed using an open source library for topic modeling BigARTM [26]. Models were trained until convergence on the train part of a dataset.

Each topic can be described as a set of the most frequent words of this topic. Several descriptions in terms of top words for 20NewsGroups are presented in **Table 1**.

Table 1: Example of 5 topics on 20NewsGroups with number of topics $|T|=10$

% of labels	Top 10 words
0	topic 1: game, team, plai, player, win, season, hockei, last, score, leagu topic 2: space, nasa, research, univers, gov, orbit, launch, program, center, system topic 3: car, price, sale, bui, want, mail, sell, speed, apr, engin topic 4: gun, state, israel, law, isra, govern, weapon, american, right, arab topic 5: kei, encrypt, chip, govern, secur, clipper, system, presid, public, work
50	topic 1: game, team, plai, player, win, season, hockei, last, leagu, score topic 2: space, nasa, scsi, system, control, orbit, work, card, launch, data topic 3: car, wire, ground, engin, power, work, water, back, want, light topic 4: gun, state, israel, isra, bike, weapon, kill, apr, law, arab topic 5: kei, govern, encrypt, system, secur, chip, presid, clipper, public, program

[1] http://www.cs.cmu.edu/~webkb/

% of labels	Top 10 words
100	topic 1: window, game, team, plai, win, hockei, file, player, season, nhl topic 2: space, nasa, work, presid, govern, orbit, state, launch, system, program topic 3: car, engin, want, come, back, work, speed, price, start, auto topic 4: gun, weapon, law, state, firearm, fire, govern, crime, control, arm topic 5: kei, armenian, encrypt, chip, govern, israel, secur, isra, system, turkish

To measure stability of the set of models generated over $r = 100$ runs, we used ASW (Eq. 21) with top $t = 10$ tokens for each topic. The estimation of the quality of the models was conducted using coherence score (Eq. 16) based on top $t = 10$ terms for each topic. We tried several set of hyperparameters for each model and present results with the best coherence. We performed several experiments with usual decorrelation, sparcing and smoothing regularizations.

Table 2: Topic stability and quality on 20NewsGroups with number of topics $|T|=10$

Modality	Regularizer	ASW	Coherence
Words, 100% of labels	Labels modality	0.86±0.01	0.16±0.01
Words, 50% of labels	Labels modality	0.86±0.01	0.18±0.01
Words, 20% of labels	Labels modality	0.83±0.01	0.21±0.01
Words, 5% of labels	Labels modality	0.79±0.01	0.24±0.01
Words	—	0.78±0.01	0.26±0.02
Words	Decorrelation Φ	0.77±0.01	0.26±0.02
Words	Sparcing Θ	0.75±0.01	0.28±0.02
Words	Smoothing Φ	0.53±0.05	0.90±0.14
Words	Sparcing Φ	0.53±0.01	0.40±0.02

Table 3: Topic stability and quality on 20NewsGroups with number of topics $|T|=60$

Modality	Regularizer	ASW	Coherence
Words, 100% of labels	Labels modality	0.76±0.00	0.35±0.01
Words, 50% of labels	Labels modality	0.70±0.00	0.46±0.01
Words, 20% of labels	Labels modality	0.63±0.01	0.55±0.01
Words, 5% of labels	Labels modality	0.60±0.01	0.62±0.01
Words	—	0.61±0.01	0.62±0.02
Words	Decorrelation Φ	0.60±0.01	0.62±0.02
Words	Sparcing Θ	0.12±0.00	0.10±0.02
Words	Smoothing Φ	0.43±0.08	0.68±0.58
Words	Sparcing Φ	0.25±0.00	0.74±0.01

Table 4: Topic stability and quality on Reuters52 with number of topics $|T|=60$

Modality	Regularizer	ASW	Coherence
Words, 100% of labels	Labels modality	0.65±0.01	0.72±0.01
Words, 50% of labels	Labels modality	0.60±0.01	0.81±0.01
Words, 20% of labels	Labels modality	0.56±0.00	0.87±0.01
Words, 5% of labels	Labels modality	0.54±0.01	0.87±0.01
Words	—	0.53±0.01	0.90±0.01
Words	Decorrelation Φ	0.52±0.01	0.91±0.01
Words	Sparcing Θ	0.08±0.00	0.07±0.01
Words	Smoothing Φ	0.68±0.05	0.53±0.19
Words	Sparcing Φ	0.20±0.00	0.83±0.02

Table 5: Topic stability and quality on Cade with number of topics $|T|=10$

Modality	Regularizer	ASW	Coherence
Words, 100% of labels	Labels modality	0.75±0.02	1.30±0.03
Words, 50% of labels	Labels modality	0.72±0.02	1.27±0.03
Words, 20% of labels	Labels modality	0.71±0.02	1.31±0.02
Words, 5% of labels	Labels modality	0.69±0.02	1.32±0.03
Words	—	0.69±0.02	1.33±0.02
Words	Decorrelation Φ	0.69±0.02	1.33±0.02
Words	Sparcing Θ	0.71±0.02	1.37±0.02
Words	Smoothing Φ	0.45±0.02	1.57±0.11
Words	Sparcing Φ	0.50±0.01	1.43±0.04

Table 6: Topic stability and quality on WebKB with number of topics $|T|=10$

Modality	Regularizer	ASW	Coherence
Words, 100% of labels	Labels modality	0.70±0.02	0.37±0.02
Words, 50% of labels	Labels modality	0.65±0.02	0.43±0.02
Words, 20% of labels	Labels modality	0.66±0.02	0.44±0.01
Words, 5% of labels	Labels modality	0.64±0.02	0.47±0.01
Words	—	0.64±0.02	0.49±0.02
Words	Decorrelation Φ	0.64±0.02	0.49±0.02
Words	Sparcing Θ	0.54±0.02	0.46±0.03
Words	Smoothing Φ	0.68±0.04	0.31±0.04
Words	Sparcing Φ	0.40±0.01	0.53±0.02

Table 7: Topic stability and quality on Habr with number of topics $|T|$=60

Modality	Regularizer	ASW	Coherence
Words, authors, users, tags, hubs	Combination of modalities	0.73±0.00	0.40±0.01
Words, tags	Tags modality	0.63±0.00	0.56±0.01
Words, hubs	Hubs modality	0.54±0.01	0.73±0.02
Words	Smoothing Φ	0.51±0.06	0.27±0.09
Words	Decorrelation Φ	0.51±0.01	0.77±0.02
Words	—	0.51±0.01	0.77±0.02

The results of topic modeling on 20NewsGroups, Reuters52, Cade and WebKB datasets (Tables 2–6) indicate that increase in the percentage of labels leads to stability growth. Moreover, models with regularizers, such as sparcing and smoothing, yield very low values of ASW compared to models with labels modality. Even 5% or 20% of labels may be enough to significantly increase model stability. However, we observe a drop in coherence score, especially in the models with high percentage of labels. Note, that models with labels modality trained on Reuters52 produce comparable and even higher coherence than models with other regularizers.

Experiments on Habr dataset show that the model combination of all five modalities outperforms all other models in terms of stability measure (Table 7). We see that the use of one additional modality—hubs or tags—increases ASW score but results in a slight dercease of quality in comparison with the use of other regularizers.

Overall, we conclude that models with different modalities, such as labels and additional meta-data, produce more stable topics. At the same time, the model with labels modality may yield low coherence score if the percentage of labels is high.

5. Conclusion

Modern topic modeling approaches suffer from instability of their results even with fixed dataset and hyperparameters. We have demonstrated that stability of topic modeling algorithm may be improved with the help of side information. Evaluation on several text corpora shows that regularization of the PLSA model with additional modalities leads to less impact of random initialization and thus more stable modeling even if side information was provided only for some subset of documents.

While our experiments were conducted on five significantly different datasets, it is still an open question what combination of additional information is the best choice for improving stability with the smallest degradation of metrics of a model. The topic for further research is to find a combination of various regularizers with the best balance between modeling stability and the quality of topics.

References

1. *Alekseev, V. et al.:* Intra-text coherence as a measure of topic models' interpretability. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference 'Dialogue 2018'. 100–108 (2018).
2. *Belford, M. et al.:* Stability of topic modeling via matrix factorization. Expert Syst. Appl. 91, 159–169 (2017).
3. *Blei, D. et al.:* Latent dirichlet allocation. Journal of Machine Learning Research. 3, 993–1022 (2003).
4. *Cardoso-Cachopo, A.:* Improving Methods for Single-label Text Categorization. PdD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa (2007).
5. *Derbanosov, R., Irkhin, I.:* Issues of stability and uniqueness of stochastic matrix factorization. Computational Mathematics and Mathematical Physics. 3, (2020).
6. *Donoho, D., Stodden, V.:* When does non-negative matrix factorization give a correct decomposition into parts? In: Advances in neural information processing systems 16. pp. 1141–1148 MIT Press (2004).
7. *Dudarenko, M.:* Regularization of multilingual topic models. Vychisl. Metody Programm. 16, 26–38 (2015).
8. *Feng, Y., Lapata, M.:* Topic models for image annotation and text illustration. In: Human language technologies: The 2010 annual conference of the north American chapter of the association for computational linguistics. pp. 831–839 Association for Computational Linguistics, Los Angeles, California (2010).
9. *Gillis, N.:* Sparse and unique nonnegative matrix factorization through data preprocessing. The Journal of Machine Learning Research. 13, 3349–3386 (2012).
10. *Hofmann, T.:* Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international aCM sIGIR conference on research and development in information retrieval. pp. 50–57 ACM, New York, NY, USA (1999).
11. *Hospedales, T.M. et al.:* Video behaviour mining using a dynamic topic model. International Journal of Computer Vision. 98, 303–323 (2011).
12. *Hu, C. et al.:* Non-negative matrix factorization for discrete data with hierarchical side-information. (2016).
13. *Ianina, A. et al.:* Multi-objective topic modeling for exploratory search in tech news. In: Artificial intelligence and natural language. pp. 181–193 Springer International Publishing, Cham (2018).
14. *Kochedykov, D. et al.:* Fast and modular regularized topic modelling. In: 2017 21st conference of open innovations association (FRUCT). IEEE (2017).
15. *Koltcov, S. et al.:* Stable topic modeling with local density regularization. In: Internet science. pp. 176–188 Springer International Publishing, Cham (2016).
16. *Kuhn, H.:* The hungarian method for the assignment problem. In: Naval Research Logistics Quaterly. pp. 83–97 (1955).
17. *Lang, K.:* Newsweeder: Learning to filter netnews. In: Proceedings of the twelfth international conference on machine learning. pp. 331–339 (1995).
18. *Laurberg, H. et al.:* Theorems on positive data: On the uniqueness of NMF. Computational Intelligence and Neuroscience. 2008, 1–9 (2008).
19. *Lewis, D.D. et al.:* RCV1: A new benchmark collection for text categorization research. J. Mach. Learn. Res. 5, 361–397 (2004).

20. *LI, X.-x. et al.:* Simultaneous image classification and annotation based on probabilistic model. The Journal of China Universities of Posts and Telecommunications. 19, 107–115 (2012).
21. *Mehrotra, R. et al.:* Improving lDA topic models for microblogs via tweet pooling and automatic labeling. Presented at the (2013).
22. *Newman, D. et al.:* Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL. pp. 100–108 (2010).
23. *Pritchard, J. et al.:* Inference of population structure using multilocus genotype data. Genetics. 155, 945–959 (2000).
24. *Shivashankar, S. et al.:* Multi-view methods for protein structure comparison using latent dirichlet allocation. Bioinformatics. 27, i61–i68 (2011).
25. *Vorontsov, K.:* Additive regularization for topic models of text collections. Doklady Mathematics. 89, 301–304 (2014).
26. *Vorontsov, K. et al.:* BigARTM: Open source library for regularized multimodal topic modeling of large collections. In: AIST. pp. 370–384 (2015).
27. *Vorontsov, K. et al.:* Non-bayesian additive regularization for multimodal topic modeling of large collections. In: TM '15. (2015).
28. *Vorontsov, K., Potapenko, A.:* Additive regularization of topic models. Machine Learning. 101, 303–323 (2014).
29. *Vorontsov, K., Potapenko, A.:* Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. 29–46 (2014).
30. *Vulic, I. et al.:* Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. Information Retrieval. 16, 331–368 (2012).
31. *Vulic, I. et al.:* Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. Inf. Process. Manage. 51, 111–147 (2015).
32. *Wang, W.:* Instantaneous versus convolutive non-negative matrix factorization. Machine Audition: Principles, Algorithms and Systems. 353–370 (2011).
33. *Zhao, H. et al.:* Leveraging external information in topic modelling. Knowledge and Information Systems. 61, 661–693 (2018).