Iterative Improvement of a Topic Model with User Feedback

A Preprint

Alex Gorbulev
gorbulev.ai@phystech.edu

Vasiliy Alexeev
vasiliy.alekseyev@phystech.edu

Konstantin Vorontsov
vokov@forecsys.ru

Abstract

We introduce the method of topic modeling using user feedback. The user marks a topic as relevant, irrelevant, or "garbage". The main problem is to improve the base model preserving relevant topics. The number of "garbage" topics should decrease. We provide the solution using topic modeling algorithms and regularizers for sparsing and decorrelation. We run the experiment on Lenta.ru news dataset.

Keywords Topic Modeling · ARTM · Natural language processing

1 Introduction

Topic modeling is the actually developing [1] method of text analysis, which is used in sociological studies [4]. Topic model estimates a probability of belonging to one of the topics. Topic modeling was used in studies of Croatian news during COVID-19 pandemic [2] and Lithuanian media representation of climate change [6].

However, learning a topic model is an ill-posed problem with an infinite set of solutions [8]. To bound the number of solutions, we add regularization penalty terms. For example, decolleration improves topic coherence [7], sparsing regularization increases the number of zero elements of matrices.

During the study period, only a part of topics could be relevant. At the same time, irrelevant topics, which duplicate relevant topics, and "garbage" topics, which are not related to the research, could contain relevant documents. To improve the quality of the research, we should reveal as more relevant texts as possible.

Our goal is to build an interpretable renewable topic model using user feedback. The user markup consists of topic distribution by category: relevant, irrelevant and "garbage" topics. The improvement of the model preserves previously found relevant topics and determines new relevant topics in place of "garbage" topics.

To solve the problem, we use additive regularization of topic models (ARTM). Open source libraries BigARTM and TopicNet [3] implement the methods of ARTM and include regularizers for decorrelation and smoothing. In ARTM, we optimize the topic model by a sum of criteria [9]. We use Lenta.ru news dataset. The dataset consists of 16449 articles in Russian language from May, 2008 to August, 2008.

2 Problem Statement

Let D denote a set of texts and W denote a set of all terms from these texts. Each term can represent a single word or a key phrase [7]. Each document $d \in D$ is a sequence of n_d terms $(w_1, \ldots, w_{n_d})$ from the set W [9]. The set of topics T is finite. Text collection D is considered to be a sample drawn independently from a discrete distribution $p(w, d, t)$ over a finite space $W \times D \times T$ [7]. According to the law of total probability and the assumption of conditional independence [8]:

$$p(w \mid d) = \sum_{t \in T} p(w \mid t, d) p(t \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d) = \sum_{t \in T} \varphi_{wt} \theta_{td} \tag{1}$$

The problem of topic modeling is to find parameters φ_{wt} and θ_{td} by text collection D that approximate frequency estimates for the conditional probabilities of $\widehat{p}(w \mid d)$. In most cases, the number of topics $|T|$ is

much smaller than the vocabulary size $|W|$ and the collection size $|D|$. This problem is equivalent to finding an low-rank stochastic matrix decomposition [7]

$$F \approx \Phi\Theta \tag{2}$$

where $F = (\widehat{p}_{wd})_{|W|\times|D|}$ is the matrix of term frequencies in documents, $\Phi = (\varphi_{wt})_{|W|\times|T|}$ is the matrix of terms of the topics, $\Theta = (\theta_{td})_{|T|\times|D|}$ is the matrix of topics of the documents.

Let T^i denote a set of topics on the iteration $i \in \mathbb{N}$, $T^i_+ \subset T^i$ denote a subset of relevant topics, $T^i_0 \subset T^i$ denote a subset of irrelevant topics, $T^i_- \subset T^i$ denote a subset of "garbage" topics, M_i denote a model on the iteration i, where $T^i = T^i_+ \sqcup T^i_0 \sqcup T^i_-$. The iterative improvement of the model M_i is the building of the model M_{i+1}, where the set of topics T_{i+1} satisfies the following requirements:

$$T^i_+ \subset T^{i+1}_+, \ \left|T^{i+1}_-\right| \leq \left|T^{i+1}_-\right|$$

3 Method

We assume to learn the new model M_{i+1} using additive regularization of topic models (ARTM) by following steps:

1. Set the alternative value of a random seed;
2. Fix the columns of the matrix Φ corresponding to relevant topics T^i_+ by sparsing regularizer [10]

$$R(\Phi, \Theta) = \beta_0 \sum_{t\in T_+} \sum_{w\in W} \beta_{wt} \widetilde{\varphi}_{wt} \ln \varphi_{wt} \tag{3}$$

   The matrix $\widetilde{\Phi}$ corresponds to the matrix of the model M_i. To fix the topics, we should use this regularizer with sufficiently large coefficient. During the experiment, we use the topic fixing regularizer with a coefficient $\beta_0 = 10^9$.
3. Use the regularizer of decorrelation to reveal new relevant topics:

$$R(\Phi) = -\tau \sum_{t\in T_-\cup T_0} \sum_{s\in T_-} \sum_{w\in W} \varphi_{wt}\widetilde{\varphi}_{ws} \to \max \tag{4}$$

$$R(\Phi) = -\tau \sum_{t\in T_-\cup T_0 s\in T_-} \langle \varphi_t, \widetilde{\varphi}_s \rangle \to \max \tag{5}$$

$$R(\Phi) = -\tau \sum_{t\in T_-\cup T_0} \left\langle \varphi_t, \sum_{s\in T_-} \widetilde{\varphi}_s \right\rangle \to \max \tag{6}$$

$$\frac{\partial R}{\partial \varphi_{wt}} = -\tau[t \in T_- \cup T_0] \sum_{s\in T_-} \widetilde{\varphi}_{ws} \tag{7}$$

$$\varphi_{wt} = norm_{w\in W} \left(n_{wt} - \tau\varphi_{wt}[t \in T_- \cup T_0] \sum_{s\in T_-} \widetilde{\varphi}_{ws} \right) \tag{8}$$

As external criterion, we use the number of relevant topics $|T_+|$ and the number of "garbage" topics $|T_-|$. The more $|T_+|$ and less $|T_-|$, the better.

We assume to use the following metrics as internal criteria:

1. Perplexity [8]

$$\mathcal{P}_m(D; p) = \exp\left(-\frac{1}{n_m} \sum_{d\in D} \sum_{w\in W^m} n_{dw} \ln p(w \mid d) \right) \tag{9}$$

2. Sparsity of the matrix Φ

3. Average topic contrast [8], where we define the contrast of the topic as

$$con_t = \frac{1}{|W_t|} \sum_{w \in W_t} p(t \mid w) \tag{10}$$

$$W_t = \{w \in W \mid \varphi_{wt} > \frac{1}{|W|}\} \tag{11}$$

4 Experiment

The code of the experiment is written in Python. We use the open source libraries BigARTM and TopicNet.

4.1 Dataset Description

To learn topic models, we use Lenta.ru news dataset. The dataset contains 16449 news articles in Russian from May 1, 2008 to August 31, 2008. Each article belongs to one of the 11 categories:

1. «Бывший СССР» (English: The Former USSR)
2. «Дом» (English: Home)
3. «Из жизни» (English: From the Life)
4. «Интернет и СМИ» (English: The Internet and the Media)
5. «Культура» (English: Culture)
6. «Мир» (English: World)
7. «Наука и техника» (English: Science and Tech)
8. «Россия» (English: Russia)
9. «Силовые структуры» (English: Security Forces)
10. «Спорт» (English: Sport)
11. «Экономика» (English: Economics)

The user could mark up different news articles from one category as belonging to relevant topic or "garbage" topic. To provide more precise clusterization, we assume to set the number of topics $|T|$ at 50.

4.2 Preprocessing

We split the title and the text of each article into tokens. Each token represents a single word. Next, we use the morph analyzer pymorphy2 [5] to lemmatize tokens. At the same time, we exclude stop-words from the tokens set. We create bigrams from two consequent lemmatized tokens. The next step is to choose top-10000 bigrams by pointwise mutual information (PMI). We represent each document as a sequence of the selected bigrams. To work with TopicNet, we convert our token data to Vowpal Wabbit format.

The total elapsed time is 11 minutes 14 seconds.

4.3 Base Model

As the base model M_0, we use the topic model without regularizers corresponding to $|T|$ specific topics. The model M_0 has regularizers for sparsing the columns of the matrices Φ and Θ related to a single background topic. The corresponding coefficients are equal to 0. We use two modalities:

1. Lemmatized tokens (@lemmatized), the coefficient is 1.0;
2. Bigrams (@bigram), the coefficient is 0.8.

The value of the parameter seed is equal to 21. We learn the model M_0 by 50 iterations.

Model	$\lvert T_+ \rvert$	$\lvert T_0 \rvert$	$\lvert T_- \rvert$
M_0	5	1	44
M_1	7	1	42
M_2	8	1	41
M_3	8	1	41

Table 1: User markup data

4.4 Creating the New Model

Let we know the user markup based on the topic model M_{i-1} and the parameters of M_{i-1}. We create the model M_i by following steps::

1. We increase the value of the parameter seed by 21;
2. Using regularizer (3), fix the topics from T_+^{i-1};
3. Create the decorrelation regularizer (4) for topics from T_0^{i-1} and T_-^{i-1};
4. Save another parameters of the model the same.

Inexistience of the determined solution [8] is one of the disadvantages of topic modeling. Topic models are dependent on the randomness. To reduce dependence, we vary the value of the parameter seed. Each value is unique for each version of the model.

4.5 Results

After the training of the model M_0, the user identified 5 relevant topics:

1. 12 (Dmitry Medvedev, international relationships)
2. 21 (the conflict in Georgia, August 2008)
3. 29 (the politics of Ukraine)
4. 30 (2008 United States presidential election)
5. 35 (Putin as Prime Minister of Russia)

The user identified the topic 31 as irrelevant for duplication the topic 21. Next, we created the model M_1. The regularizer (3) preserved previously found relevant topics. Moreover, the user identified 2 new relevant topics:

1. 14 (sanctions against Iran)
2. 15 (the politics of Zimbabwe, elections)

Next, we created the model M_2. The sparsing regularizer preserved each topic from T_+^1 as relevant. The user identified the topic 46 (Hague Tribunal) as new relevant topic. After training the new model M_3, the user markup remained the same.

By each modality, the models M_1 и M_2 have bigger perplexity than the model M_0. In most cases, topic models without regularizers have less perplexity than topic models with regularizers [7]. At the same time, after 50 iterations the perplexity of the model M_3 is similar to the perplexity of the model M_0.

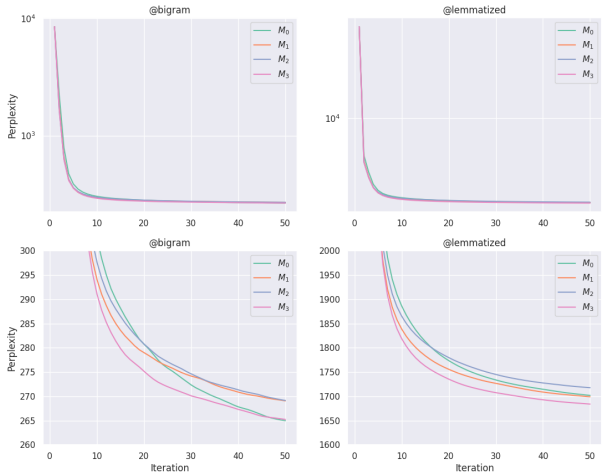

Figure 1: Perplexity

The sparsity of the matrix Φ varies slightly from model to the model.

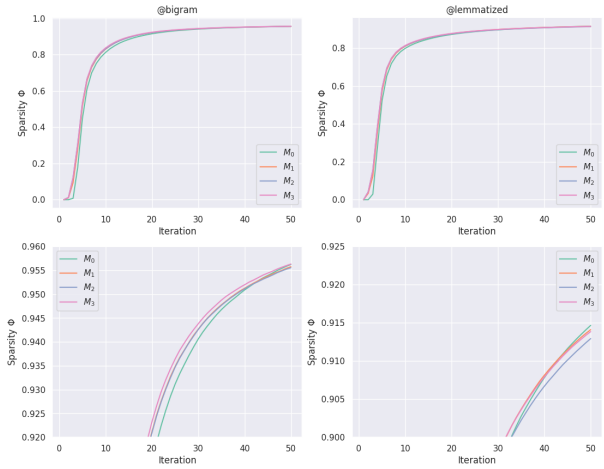

Figure 2: Sparsity of the Φ

By average topic contrast, the models M_0 and M_3 show the similar results. By lemmatized tokens modality, the model M_0 has the biggest average topic contrast. At the same time, the differencies are quite small.

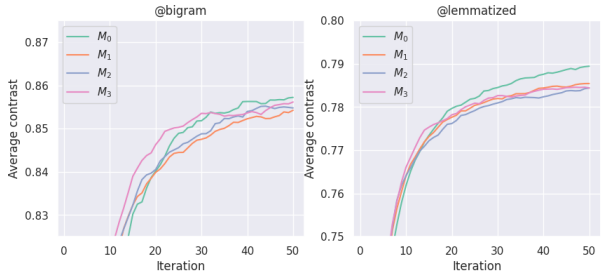

Figure 3: Average topic contrast

5 Conclusion

In this paper, we propose the method of the iterative improvement of a topic model. To fix relevant topics, we use sparsing regularizer. To reveal new relevant topics, we use decorrelation regularizer. In this experiment, the sparsing regularizer preserves relevant topics, and the number of relevant topic increases after using regularizer for the decorrelation. We plan to conduct new experiments on specialized datasets to test the universality of the proposed method.

References

[1] Jedidiah Aqui and Michael Hosein. Mobile ad-hoc networks topic modelling and dataset querying. In 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), pages 1–6, 2022.

[2] Petar Kristijan Bogović, Ana Meštrović, Slobodan Beliga, and Sanda Martinčić-Ipšić. Topic modelling of croatian news during covid-19 pandemic. In 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), pages 1044–1051, 2021.

[3] Victor Bulatov, Vasiliy Alekseev, Konstantin Vorontsov, Darya Polyudova, Eugenia Veselova, Alexey Goncharov, and Evgeny Egorov. TopicNet: Making additive regularisation for topic modelling accessible. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 6745–6752, Marseille, France, May 2020. European Language Resources Association.

[4] Paul DiMaggio, Manish Nag, and David Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. Poetics, 41(6):570–606, 2013. Topic Models and the Cultural Sciences.

[5] Mikhail Korobov. Morphological analyzer and generator for russian and ukrainian languages. In Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry I. Ignatov, and Valeri G. Labunets, editors, Analysis of Images, Social Networks and Texts, volume 542 of Communications in Computer and Information Science, pages 320–332. Springer International Publishing, 2015.

[6] Florian Rabitz, Audronė Telešienė, and Eimantė Zolubienė. Topic modelling the news media representation of climate change. Environmental Sociology, 7(3):214–224, 2021.

[7] Konstantin Vorontsov and Anna Potapenko. Additive regularization of topic models. Machine Learning, 101:1–21, 12 2014.

[8] К. В. Воронцов. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект bigartm.

[9] К. В. Воронцов. Аддитивная регуляризация тематических моделей коллекций текстовых документов. In Доклады РАН, volume 456, pages 268–271, 2014.

[10] К. В. Сухарева А. В., Воронцов. Построение полного набора тем вероятностных тематических моделей. In Интеллектуальные системы. Теория и приложения, volume 23, pages 7–23, 2019.