Computación y Sistemas
ISSN: 1405-5546
computacion-y-sistemas@cic.ipn.mx
Instituto Politécnico Nacional
México

Apishev, Murat; Koltcov, Sergei; Koltsova, Olessia; Nikolenko, Sergey; Vorontsov, Konstantin

Mining Ethnic Content Online with Additively Regularized Topic Models

Computación y Sistemas, vol. 20, núm. 3, 2016, pp. 387-403
Instituto Politécnico Nacional
Distrito Federal, México

Available in: http://www.redalyc.org/articulo.oa?id=61547469008

Mining Ethnic Content Online with Additively Regularized Topic Models

Murat Apishev[2,4], Sergei Koltcov[1], Olessia Koltsova[1], Sergey Nikolenko[1,3], Konstantin Vorontsov[4,5]

[1] National Research University Higher School of Economics, St. Petersburg, Russia

[2] Moscow State University, Moscow, Russia

[3] Steklov Institute of Mathematics at St. Petersburg, St. Petersburg, Russia

[4] Yandex, Moscow, Russia

[5] Moscow Institute of Physics and Technology, Moscow, Russia

great-mel@yandex.ru, kol-sergei@yandex.ru, ekoltsova@hse.ru, sergey@logic.pdmi.ras.ru, vokov@forecsys.ru

Abstract. Social studies of the Internet have adopted large-scale text mining for unsupervised discovery of topics related to specific subjects. A recently developed approach to topic modeling, additive regularization of topic models (ARTM), provides fast inference and more control over the topics with a wide variety of possible regularizers than developing LDA extensions. We apply ARTM to mining ethnic-related content from Russian-language blogosphere, introduce a new combined regularizer, and compare models derived from ARTM with LDA. We show with human evaluations that ARTM is better for mining topics on specific subjects, finding more relevant topics of higher or comparable quality.

Keywords. Topic modeling, additive regularization of topic models, computational social science.

1 Introduction

Topic models have become a common tool for unsupervised analysis of large text corpora, mining a corpus for latent topics expressed as distributions over words while at the same time inferring how documents are distributed among these topics. In essence, a topic model decomposes the sparse word-document matrix into a product of word-topic and topic-document matrices; this idea was first fleshed out in probabilistic latent semantic analysis (PLSA) [15], and now the topic model of choice is latent Dirichlet allocation (LDA), which is a Bayesian version of PLSA with Dirichlet priors assigned to word-topic and topic-document distributions [9, 14].

Over the years, LDA has received tremendous attention, with many extensions developed for many different purposes, but each of them has been a separate research project, with a new version of one of the two basic inference algorithms for LDA: either a variational approximation for the new posterior distribution or a new Gibbs sampling scheme. Hence, it is hardly practical to expect a researcher, especially in social sciences, to develop new LDA extensions for each new problem; even slight modifications of an existing extension may be hard both to develop and to implement in software.

Computación y Sistemas, Vol. 20, No. 3, 2016, pp. 387–403
ISSN 1405-5546
doi: 10.13053/CyS-20-3-2473

Hence, we use a recently developed approach called additive regularization for topic modeling (ARTM) [31] and the corresponding open-source implementation BigARTM [32]. ARTM extends the basic PLSA model with a general regularization mechanism that can directly express desired properties in the objective function, and the inference algorithm results automatically.

As a special case, ARTM with smoothing regularizers can mimic the LDA model [34], although such regularization results in the same model stability as already noted for LDA itself [24, 34]. Note that recent studies uncover deep problems with the basic LDA model, specifically its instability stemming from numerous local maxima of the objective function [1, 18, 19].

Flexibility is a big advantage of ARTM in practice, especially for digital humanities where one often has but a feeling of what one is looking for. Having trained a trial model in the form of regular LDA or ARTM without regularizers, a researcher can formulate what is lacking and what is desired of the resulting topic model. In most cases, BigARTM lets a researcher combine regularizers from a built-in library in order to meet a set of requirements to the model quickly and efficiently. To achieve all these results, a social scientist has only to learn how to create regularizes and set their parameters; this can be done easily by editing a few lines of code; with no mathematical inference and no coding. Having trained a trial model in the form of regular LDA or ARTM without regularizers, a researcher can formulate what is lacking and what is desired of the resulting topic model. The BigARTM framework also lets one quickly develop and test new regularizers tailored specifically for one's problem. In most cases, BigARTM lets a researcher combine regularizers from a built-in library in order to meet a set of requirements to the model quickly and efficiently.

In this work, we show one such application of the ARTM approach for the problem of mining a large corpus/stream of user-generated texts (in our case, blog posts) for specific topics of discourse (in our case, ethnic-related topics) defined with a fixed dictionary of subject terms (ethnonyms). To achieve a good topic model, we split the entire set of topics into subject-related and background topics, develop a new regularizer that deals with this predefined dictionary of subject terms, and build a combination of regularizers to make topics more interpretable, sparse, and diversified. The ARTM framework lets us do all of these things seamlessly, without complicated inference and developing new algorithms.

We present an extensive evaluation of our results, concentrating on interpretability evaluation produced by a team of human assessors; this is both needed in our case study (where it is important for topics to be interpretable for non-specialists) and generally represents a gold standard for the quality of a topic model. Experimental results suggest that while the basic (unregularized or weakly regularized) ARTM model is no better than regular LDA, new regularizers significantly improve both number and quality of relevant topics.

The paper is organized as follows. In Section 2, we introduce the basic PLSA model, its Bayesian counterpart LDA, and the general setting of the ARTM approach. In Section 3, we review regularizers used in this work and comment on their effect on the resulting topic model. Section 4 lists the specific models we have trained and covers the results of our case study in finding ethnic-related texts in a large dataset of blog posts. Section 5 concludes the paper.

A preliminary version of this work has appeared in the Proceedings of the 15th Mexican International Conference on Artificial Intelligence (MICAI 2016) [4]; compared to the conference version, we have conducted a novel study of topic modeling on a reduced collection filtered with respect to the top words of relevant topics (see Section 4.5) and extended the survey part of the paper (Section 2).

2 Topic Modeling and Related Work

Let D denote a finite set (collection) of documents (texts) and let W denote a finite set (vocabulary) of all terms from these documents. Each term can represent a single word or a key phrase. Following the "bag of words" model, we represent each document d from D as a subset of terms from the vocabulary W. Assume that each term occurrence in each document refers to some latent

topic from a finite set of topics T. A text collection is considered as a sequence of triples (d_i, w_i, t_i), $i = 1, \ldots, n$, drawn independently from a discrete distribution $p(d, w, t)$ over the finite probability space $D \times W \times T$. Terms w_i and documents d_i are observable variables, while topics t_i are latent variables.

A *probabilistic topic model* represents the probabilities $p(w \mid d)$ of terms occurring in documents as mixtures of term distributions in topics $\phi_{wt} = p(w \mid t)$ and topic distributions in documents $\theta_{td} = p(t \mid d)$:

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

This mixture also directly corresponds to the generative process for a document d: for each term position i, sample topic index t_i from distribution $p(t \mid d)$ and then sample the word w_i from distribution $p(w \mid t_i)$.

Parameters of a probabilistic topic model are usually represented as matrices

$$\Phi = \big(\phi_{wt}\big)_{W \times T}, \quad \Theta = \big(\theta_{td}\big)_{T \times D},$$

with non-negative and normalized columns ϕ_t and θ_d representing multinomial word-topic and topic-document distributions respectively.

2.1 PLSA

In *Probabilistic Latent Semantic Analysis* (PLSA) [15, 16] the topic model (1) is trained by log-likelihood maximization with linear constraints of nonnegativity and normalization:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \;\to\; \max_{\Phi, \Theta} \quad (2)$$

under constraints

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

where n_{dw} is the number of occurrences of the term w in the document d. The solution of this optimization problem satisfies the following Karush–Kuhn–Tucker conditions with auxiliary variables p_{tdw}, n_{wt}, n_{td}:

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\big(\phi_{wt} \theta_{td}\big), \quad (4)$$

$$\phi_{wt} = \underset{w \in W}{\operatorname{norm}}(n_{wt}), \quad \theta_{td} = \underset{t \in T}{\operatorname{norm}}(n_{td}), \quad (5)$$

where $n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}$, $n_{td} = \sum_{w \in d} n_{dw} p_{tdw}$, and the "norm" operator transforms a vector $(x_t)_{t \in T}$ into $(\tilde{x}_t)_{t \in T}$ representing a discrete distribution:

$$\tilde{x}_t = \underset{t \in T}{\operatorname{norm}}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}.$$

This system follows from (2)-(3) and can be solved numerically. The simple-iteration method for this system of equations is equivalent to the EM algorithm and is typically used in practice. It repeats two steps in a loop according to the equations above.

The E-step (4) can be understood as the Bayes rule for the probability distribution of topics $p_{tdw} = p(t \mid d, w)$ for each term w in each document d. Auxiliary variable n_{wt} estimates how many times the term w is associated with the topic t over all documents; n_{td} estimates how many terms from document d are associated with the topic t. The M-step (5) can be interpreted as frequency estimation for conditional probabilities ϕ_{wt} and θ_{td}. The iterative process begins with a random initialization of Φ and Θ.

2.2 LDA

The latent Dirichlet allocation (LDA) model [9, 14] introduces prior Dirichlet distributions for the vectors of term probabilities in topics $\phi_t \sim \mathrm{Dir}(\beta)$ as well as for the vectors of topic probabilities in documents $\theta_d \sim \mathrm{Dir}(\alpha)$ with vector parameters $\beta = (\beta_w)_{w \in W}$ and $\alpha = (\alpha_t)_{t \in T}$ correspondingly.

Inference in LDA is usually done via either variational approximations or Gibbs sampling, In the basic LDA model, with the latter reducing to the so-called *collapsed Gibbs sampling*, where θ and ϕ variables are integrated out, and topic t_i for each word position (d_i, w_i) is iteratively resampled from $p(t \mid d, w)$ distribution estimated according to the same formula (4), similar to PLSA, but

with smoothed Bayesian estimates of conditional probabilities:

$$\phi_{wt} = \underset{w \in W}{\text{norm}}(n_{wt} + \beta_w), \quad \theta_{td} = \underset{t \in T}{\text{norm}}(n_{td} + \alpha_t),$$

where n_{wt} is the number of times term w has been generated from topic t and n_{td} is the number of terms in document d generated from topic t except the current triple (d_i, w_i, t_i).

Over the recent years, the basic LDA model has been subject to many extensions, each presenting either a variational of a Gibbs sampling algorithm for a model that extends LDA to incorporate some additional information or presumed dependencies.

Extensions that add new dependencies include *correlated topic models* (CTM) that exploit the fact that some topics are more or less similar to each other and may share words with each other, using logistic normal distribution instead of Dirichlet to model correlations between topics [6], *Markov topic models* use Markov random fields to model the interactions between topics in different parts of the dataset (different text corpora), connecting a number of different hyperparameters β_i in a Markov random field expressing prior constraints [20], *relational topic models* construct a hierarchical model reflecting the structure of a document network as a graph [11], and so on.

Extensions that use additional external information include various time-related extensions such as *Topics over Time* [36] or *dynamic topic models* [7, 35], that apply when documents have timestamps (e.g., news articles or blog posts) and represent topic evolution in time; *supervised LDA* that assigns each document with an additional observed response variable [8], an approach that can be extended further to, e.g., recommender systems [23]; sentiment-related extensions add sentiment variables to the basic topic model and train both topics and sentiment variables in various contexts [21, 30, 38], and so on. In particular, a lot of work has been done on nonparametric LDA variants based on Dirichlet processes that can determine the optimal number of topics automatically [13, 28, 37].

For our present purpose of mining and analyzing documents related to a specific user-defined topic, the LDA extensions that appear to be most relevant are the *Topic-in-Set knowledge* model and its extension with Dirichlet forest priors [2, 3], where words are assigned with "z-labels"; a z-label represents the topic this specific word should fall into, and the Interval Semi-Supervised LDA (ISLDA) model [10, 24] where specific words are assigned to specific topics, and sampling distributions are projected onto that subset.

2.3 ARTM

Topic modeling can be viewed as a special case of matrix factorization, where the problem is to find a low-rank approximation $\Phi\Theta$ of a given sparse matrix of term-document occurrences. Note, however, that the product $\Phi\Theta$ is defined only up to a linear transformation since

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta)$$

for any nondegenerate matrix S. Therefore, our problem is ill-posed and generally has an infinite set of solutions. Previous experiments on simulated data [34] and real social media data [10] show that neither PLSA nor LDA can ensure a stable solution. To make the solution more appropriate one must introduce additional optimization criteria, usually called *regularizers* [29].

The Dirichlet prior can be considered as a weak smoothing regularizer. Therefore, our starting point will be the PLSA model, completely free of regularizers, rather than the LDA model, although the latter is more popular in recent research works.

In *Additive Regularization of Topic Models* (ARTM) [31] a topic model (1) is trained by maximizing a linear combination of the log-likelihood $L(\Phi, \Theta)$ and r regularizers $R_i(\Phi, \Theta)$, $i = 1, \ldots, r$ with *regularization coefficients* τ_i:

$$R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta),$$

$$L(\Phi, \Theta) + R(\Phi, \Theta) \;\to\; \max_{\Phi, \Theta}.$$

Karush–Kuhn–Tucker conditions for this nonlinear problem yield (under some technical

restrictions) necessary conditions for the local maximum [34]:

$$p_{tdw} = \underset{t\in T}{\text{norm}}(\phi_{wt}\theta_{td}), \tag{6}$$

$$\phi_{wt} = \underset{w\in W}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right), \tag{7}$$

$$\theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \tag{8}$$

where $n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}$ and $n_{td} = \sum_{w\in d} n_{dw}p_{tdw}$. Again, this system of equations can be solved with the EM algorithm.

The strength of ARTM is that each additive regularization term R_i yields a simple additive modification of the M-step. Many models previously developed within the Bayesian framework can be easier reinterpreted, trained, and combined in the ARTM framework [33, 34]; e.g., PLSA does not use regularization at all, $R = 0$, and LDA with Dirichlet priors $\phi_t \sim \text{Dir}(\beta)$ and $\theta_d \sim \text{Dir}(\alpha)$ and maximum a posteriori estimation of Φ, Θ corresponds to the smoothing regularizer [5]. The regularizer can be interpreted as a minimizer of KL-divergences between the columns of Φ, Θ and fixed distributions β, α respectively.

3 Additive Regularization

3.1 General Approach

In this section, we consider an exploratory search problem of discovering all ethnic-related topics in a large corpus of blog posts. Given a set of ethnonyms as a query $Q \subset W$, we would like to get a list of ethnically relevant topics. We use a semi-supervised topic model with lexical prior to solve this problem; similar models have appeared for news clustering tasks [17], discovering health topics in social media [25] and ethnic-related topics in blog posts [10, 24]. In all these studies, researchers specify for each predefined topic a certain set of *seed words*, usually very small, e.g., a news category or ethnicity. This means that we must know in advance how many topics we would like to find and what each topic should be generally about. The *interval semi-supervised LDA* model (ISLDA) allows to specify more than one topic per ethnicity [10], but it is difficult to guess how many topics are associated with each ethnicity, and if an expert does not anticipate a certain subset of seed words, it will be impossible to learn in the model. Moreover, and in [10, 24], where the case study was similar to our present work, ISLDA was used to look for ethnic-related topics, but since seed words related to different ethnicities were separated into different topics, so no multi-ethnic topics could appear. In our present approach, the topic model has more freedom to decide the composition of subject topics in S. Moreover, all cases above include a large amount of preliminary work involved in associating seed words with predefined topics.

We address the above problems by providing a lexical prior determined by a set of ethnonyms Q common for all ethnically relevant topics. The model itself determines which ethnicity or combination of ethnicities make up each relevant topic.

We use an additive combination of regularizers for smoothing, sparsing, and decorrelation in order to make topics more interpretable, sparse, and diversified [34]. The ARTM framework lets us do all of these things seamlessly, without complicated inference and developing new algorithms. All these regularizers have been implemented as part of the BigARTM open-source topic modeling toolbox. We show that the combination of regularizers significantly increases the number of retrieved well-interpretable ethnical topics.

First of all, we split the entire set of topics T into two subsets: domain-specific *subject* topics S and *background* topics B. Regularizers will treat S and B differently. The relative size of S and B depends on the domain and has to be set in advance by the user. The idea of background topics that gather uninteresting words goes back to the *special words with background* (SWB) topic model [12], but unlike SWB, we define not one but many background topics in order to model irrelevant non-ethnic-related topics better, thereby improving the overall quality of the model.

3.2 Smoothing and Sparsing

A straightforward way to integrate lexical priors is to use smoothing and sparsing regularizers

with uniform β distribution restricted to a set of ethnonyms Q:

$$\beta_w = \frac{1}{|Q|}[w \in Q].$$

We introduce a smoothing regularizer that encourages ethnonyms $w \in Q$ to appear in ethnic-related topics S together with a sparsing regularizer that prevents ethnonyms from appearing in background topics B:

$$R(\Phi) = \tau_1 \sum_{t \in S} \sum_{w \in Q} \ln \phi_{wt} - \tau_2 \sum_{t \in B} \sum_{w \in Q} \ln \phi_{wt}.$$

In the exploratory search task, relevant content usually constitutes a very small part of the collection. In our case, the entire ethnicity discourse in a large dataset of blog posts is unlikely to add up to more than one percent of the total volume. Our goal is to mine fine-grained thematic structure of relevant content with many small but diverse and interpretable subject topics S, but also to describe a much larger volume of content with a smaller number of background topics B. Formally, we introduce a smoothing regularizer for background topics B in Θ and a sparsing regularizer that uniformly supresses ethnic-related topics S:

$$R(\Theta) = \tau_3 \sum_{d \in D} \sum_{t \in B} \ln \theta_{td} - \tau_4 \sum_{d \in D} \sum_{t \in S} \ln \theta_{td}.$$

The idea is to make background topics B smooth, so that they will contain irrelevant words, and subject topics S sparse, so that they will be as distinct as possible, with each topic concentrating on a different and meaningful subject.

3.3 Decorrelation

Diversifying the term distributions of topics is known to make the resulting topics more interpretable [27]. In order to make the topics as different as possible, we introduce a regularizer that minimizes the sum of covariances between ϕ_t vectors over all specific topics t:

$$R(\Phi) = -\tau_5 \sum_{t \in S} \sum_{s \in S \setminus t} \sum_{w \in W} \phi_{wt}\phi_{ws} + \tau_6 \sum_{t \in B} \sum_{w \in W} \ln \phi_{wt}.$$

The decorrelation regularizer also stimulates sparsity and tends to group stop-words and common words into separate topics [27]. To move these topics from S to B, we add a second regularizer that uniformly smoothes background topics.

3.4 Modality of Seed Words (Ethnonyms)

Another possible way to use lexical priors is to distinguish ethnonyms into a separate *modality*. Generally, modality is a kind of tokens in a document. Examples of modalities include a separate class of tokens (sample modalities include named entities, tags, foreign words, n-grams, authors, categories, time stamps, references, user names etc.). Each modality has its own vocabulary and its own Φ matrix normalized independently. A multimodal extension of ARTM has been proposed in [32] and implemented in BigARTM. We introduce two modalities: words and ethnonyms. The latter is defined by a seed vocabulary Q and matrix $\tilde{\Phi}$ of size $|Q| \times |T|$. In ARTM, the log-likelihood of a modality is treated as a regularizer:

$$R(\tilde{\Phi}, \Theta) = \tau_7 \sum_{d \in D} \sum_{w \in Q} n_{dw} \ln \sum_{t \in T} \tilde{\phi}_{wt}\theta_{td},$$

where regularization coefficient τ_7 is in fact a multiplier for word-document counters n_{dw} of the second modality.

In order to make ethnic-related topics more diverse in their ethnonyms, we introduce an additional decorrelation regularizer for the modality of ethnonyms:

$$R(\tilde{\Phi}) = -\tau_8 \sum_{t \in S} \sum_{s \in S \setminus t} \sum_{w \in Q} \tilde{\phi}_{wt}\tilde{\phi}_{ws}.$$

Note that we introduce decorrelation for subject topics S separately for words modality with Φ matrix and for ethnonyms modality with $\tilde{\Phi}$ matrix.

3.5 Putting It All Together

The BigARTM library[1] lets users build topic models for various applications simply by choosing a suitable combination of predefined regularizers. All of the regularizers listed above can be used in any combination; by using different mixtures one can achieve different properties for the resulting topic model. In one of the models (model 5 in Section 4.2), we combined all regularizers described above. Note that while the resulting models have relatively many hyperparameters, and optimal tuning of them may incur prohibitive computational costs, in practice it suffices to set the hyperparameters to some reasonable values found in previous experiments. In all results shown below, hyperparameters were tuned with a greedy procedure, one by one.

4 Evaluation

4.1 Datasets and Settings

From the sociological point of view, the goal of our project is to mine and monitor ethnic-related discourse in social networks, e.g., find how popular topics are related to various ethnic groups, perhaps in specific regions, and identify worrying rising trends that might lead to ethnic-related outbursts or violence. While multimodal analysis that would account for topic evolution in time and their geospatial distribution remains a subject for further work, we evaluate our models on a real life dataset mined from the most popular Russian blog platform *LiveJournal*.

The dataset contains ≈ 1.58 million lemmatized posts from the top 2000 *LiveJournal* bloggers embracing an entire year from mid-2013 to mid-2014. Data were mined weekly according to the *LiveJournal*'s rating that was quite volatile, which is why the number of bloggers in the collection comprized several dozens of thousands. The complete vocabulary amounted to 860K words, but after preprocessing (leaving only words that contain only Cyrillic symbols and perhaps a hyphen, are at least 3 letters long, and occur ≥ 20 times in the corpus) it was reduced to 90K words in ≈ 1.38 million nonempty documents.

To choose the number of topics, we have trained PLSA models with 100, 300, and 400 topics, evaluated (by a consensus of a team of human assessors) that the best result was at 400 topics, and hence chose to use 400 topics in all experiments. This corresponds to our earlier experiments with the number of topics in relation to mining ethnic discourse [10, 24].

The collection was divided into batches of 10000 documents each. All ARTM-based models were trained by an online algorithm with a single pass over the collection and 25 passes over each document; updates are made after processing every batch. For the semi-supervised regularizer, we have composed a set of several hundreds ethnonyms — nouns denoting various ethnic groups, based on literature review, Russian census and UN data, expert advice and other sources; 249 of those words occurred in the collection. Ethnonyms were considered the best candidates for improving mining topics that correspond to the sociological notion of ethnicity and inter-ethnic relations. The latter are understood as interpersonal or intergroup interactions and attitudes caused or justified by the ethnic status of participants; they should be differentiated from international relations where the main actors are countries, including nation-states, and their governments or individual official representatives, and the subject is not always related to the ethnic status of individuals or groups. International and inter-ethnic relations are closely connected and in some situations inseparable, however, intuitively it is clear that for preventing internal ethnic conflict monitoring attitudes to migrants expressed by bloggers is more relevant than mining news on world summits or international trade treaties. We, therefore, assumed that topics on ethnicity per se should be dominated by ethnonyms (Turks), while ethnic adjectives (Turkish) and country names (Turkey) would more probably refer to international relations. In the Russian language, these three categories are almost always different words, which in our mind could contribute to easier differentiation between topics on ethnicity and on international relations.

[1] http://bigartm.org/

4.2 Models

In our BigARTM experiments, we have trained a series of topic models. In all models with hyperparameters, we have tuned these hyperparameters to obtain the best models available for a specific model with a greedy procedure: start from reasonable default values, optimize the first parameter, fix it and optimize the second parameter and so on.

In total we have evaluated eight models with $|T| = 400$ topics each. For all models, we have chosen regularization coefficients manually based on the results of several test experiments. In all additively regularized models with lexical priors, we divided topics into $|S| = 250$ subject topics and $|B| = 150$ background topics. Next we list the different models compared in the experiments below and provide the motivation behind introducing and comparing these specific topics:

(1) plsa: reference PLSA model with no regularizers;

(2) lda: LDA model implemented in BigARTM with smoothness regularizers on Φ and Θ with uniform α and β and hyperparameters $\alpha_0 = \beta_0 = 10^{-4}$;

(3) smooth: ARTM-based model with smoothing and sparsing by the lexical prior, with regularization coefficients $\tau_1 = 10^{-5}$ and $\tau_2 = 100$ (tuned by hand); besides, in this and all subsequent regularized models we used the smoothing regularizer for the Θ matrix with coefficients $\tau_3 = 0.05$ and $\tau_4 = 1$;

(4) decorrelated: ARTM-based model that extends (3) with decorrelation with coefficients $\tau_5 = 5 \times 10^4$ and $\tau_6 = 10^{-8}$; the smoothing coefficient for ethnically relevant subject topics was $\tau_1 = 10^{-6}$;

(5) restricted dictionary: ARTM-based model that extends (4) by adding a modality of ethnonyms with coefficients $\tau_7 = 100$ and $\tau_8 = 2 \times 10^4$; the decorrelation coefficients was $\tau_5 = 1.5 \times 10^6$ and $\tau_6 = 10^{-7}$; subject words were smoothed with coefficient $\tau_1 = 1.1 \times 10^{-4}$; for this model we used a dictionary with $|Q| = 249$ ethnonyms;

(6) extended dictionary: same as (5) but with dictionary extended by adjectives and country names if respective ethnonyms did not occur; the positive outcome here would be that more relevant topics can be found with an extended dictionary, while the negative outcome is that ethnic topics could instead get lost within topics on international relations;

(7) recursive: the basic PLSA model trained on a special subset of documents, namely documents retrieved from topics that were considered ethnic-relevant by assessors in model 5 with a threshold of 10^{-6} in the Θ matrix for all subject topics; here, the hypothesis was that a collection with a higher concentration of relevant documents could yield better topics;

(8) keyword documents: PLSA model identical to (7) but trained on a subset of only those documents that contained at least one word from the dictionary.

Models 7 and 8 were introduced to test two different ways of enriching the initial collection. Model 8 was used as reference for model 7: it was to check if enriching the collection through a preliminary cycle of topic modeling would yield better results than retrieving texts via a simple keyword search.

Figure 1 shows several sample topics from some of the models (translated to English; superscript ^{adj} denotes an adjectival form of the word, usually a different word in Russian). It appears that later models, 6 and 7, yield topics that are better suited for the ethnic purpose of our study; in what follows, we will expand and quantify this observation.

4.3 Assessment

In the rest of this section, we discuss the qualitative and quantitative results of our study, starting from the assessment methodology and then discussing the results of our human coding experiments. However, results coming from the assessors were supplemented with values of the tf-idf coherence quality metric introduced earlier in [10, 24]. It has

Table 1. Sample ethnic-related topics from several models

Model	Sample topic
(1)	Muslim, religious, Islam, extrasensoric, sect, Christian, alley, radical, labyrinth, Uzbekistan, Christian^{adj}, Islam^{adj}
(2)	republic, Caucasus, sometimes, Chechen, Caucasian^{adj}, Dagestan, nationality, Checnya, region, power, Ingushetia
(3)	Armenia, Azerbaijan, Armenia^{adj}, Armenian, caravan, Yerevan, Tajik, Azeri, Azeri^{adj}, Uzbek, Alice, SSR, Tatar, survey
(5)	Uzbek, Russian, Russia, migrant, Uzbekistan, work^{adj}, Moscow, country, Tajik, janitor, place, work, citizen, home, Asia
(6)	Russian, Uzbek, Tajik, migrant, Russia, work, janitor, border, work, Uzbekistan, guest worker, place, town, Asia
(6)	Kazakhstan, Asia, region, central, Kyrgizia, Tajikistan, Afganistan, country, republic, Middle, Uzbekistan, territory, Russia
(7)	migrant, country, Russia, migration, Asia, illegal, migrant^{adj}, Tajikistan, guest worker, citizen, work^{adj}, work, Middle
(7)	Kazakhstan, region, country, Asia, republic, Kyrgizia, Russia, state, military, central, territory, defense, collaboration

Table 2. Average coherence and tf-idf coherence for all models in the study

Model	T	coh_{10}	tfidf_{10}	coh_{20}	tfidf_{20}
1 (plsa)	400	-325.3	-212.0	-1447.0	-1011.6
2 (lda)	400	-344.2	-230.9	-1539.8	-1121.2
3 (smooth)	400	-367.1	-261.2	-1583.9	-1210.2
4 (decorr)	400	-378.9	-274.0	-1651.2	-1296.1
5 (restr. dict.)	400	-310.0	-196.4	-1341.9	-908.4
6 (ext. dict.)	400	-321.7	-209.6	-1409.1	-995.3
7 (recursive)	400	-326.5	-212.1	-1415.6	-982.5
8 (keyword)	400	-328.8	-214.4	-1463.6	-1014.5

been shown that tf-idf coherence better matches the human judgment of topic quality than the traditional coherence metric [22].

Table 3. Intercoder agreement: share of differing answers

Question	Diff.
1 (general understanding)	0.28
2 (event/phenomenon)	0.30
3 (ethnonyms)	0.07
4 (ethnic issues)	0.06
5 (international relations)	0.08
6 (other)	0.25

Results on average coherence and tf-idf coherence for all topics in every model are shown in Table 2; we show two versions of coherence-based metrics, computed with top 10 words in a topic and computed with top 20 words. The distributions of all four metrics are also shown in more detail on Fig. 1, which shows the sorted metrics (coh_{10}, tfidf_{10}, coh_{20}, and tfidf_{20}) for each model, so a graph that goes above the other represents the better model. Table 2 and Fig. 1 show that while models 5 (restricted dictionary) and 6 (extended dictionary) win in all four cases, all models have comparable values with respect to the topic quality metrics except for models 3 (smooth) and 4 (decorrelated). This was supported by preliminary human evaluation, so we decided to drop these two sets of results from further consideration, choosing to use limited human assessment resources on the better models.

For all other models, assessors were asked to interpret the topics based on 20 most probable words in every topic of each model, except models 3 and 4 that demonstrated much lower quality as measured with coherence and tf-idf coherence [24] and thus were excluded from assessment. For each topic, two assessors answered the following questions, related both to the overall quality and to the ethnic nature of our study:

(1) Do you understand why these words are collected together in this topic? (1) absolutely not; (2) partially; (3) yes.

(2) If you answered "partially" or "yes" to question 1: do you understand which event or phenomenon can be discussed in texts related to this topic? (1) absolutely not; (2) partially; (3) yes.

(3) Is there an ethnonym among the top-words of this topic? Specify the total number of ethnonyms.

(4) If you answered "partially" or "yes" to question 2: is this event or phenomenon related to ethnic issues? (1) not at all; (2) partially or unclear; (3) yes.

(5) If you answered "partially" or "yes" to question 2: is this event or phenomenon related to international relations? (1) not at all; (2) partially or unclear; (3) yes.

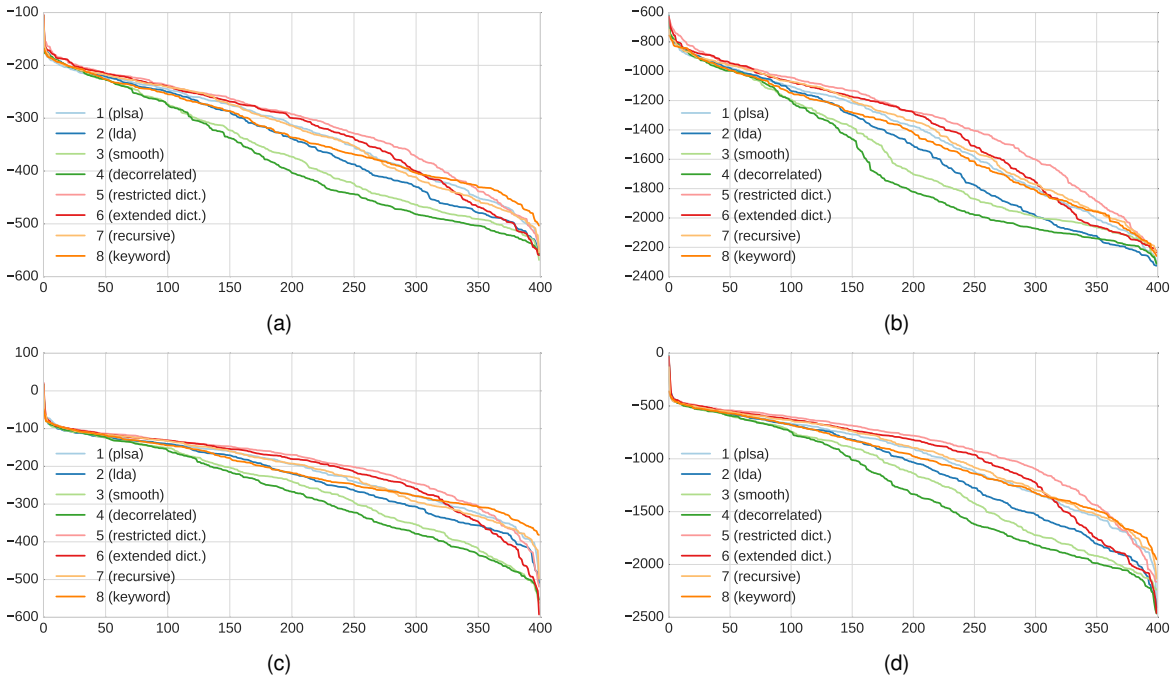

Fig. 1. Sorted topic quality metrics: (a) coh_{10}; (b) tfidf_{10}; (c) coh_{20}; (d) tfidf_{20}

(6) If you answered "partially" or "yes" to question 2: is this event or phenomenon related to some other category of topics, not related to ethnicity? (1) not at all; (2) partially or unclear; (3) yes.

Assessors were clearly instructed on all matters, including the differences between ethnicity and international relations. We have asked assessors about both of these issues because from our previous experience with semi-supervised approaches [10, 24] we know that the international relations topics are often retrieved instead of ethnic-related topics or tend to blend with them. This, ultimately, produces high probabilities for documents devoted to global political conflicts/relations or just travel abroad and fails to bring up texts related to internal ethnic conflict, everyday interethnic communication, including hate speech, or national policies on ethnicity issues — everything that was considered important in this case study. We, therefore, wanted to discriminate between the algorithms good at retrieving international relations topics and those able to retrieve exactly what we want — ethnic discourse.

We have collected the answers of seven assessors; Table 3 summarizes total intercoder agreement values, showing the share of differing answers for every question. In general, these results show good convergence between the assessors, on the level of our previous experiments with similar evaluation [26]. When the assessors disagreed in assigning a topic to a category, rather than averaging their results we produced two sets of scores: in the first set, we assigned each topic a maximum from the assessors' scores; in the second set, we did the opposite — that is, assigned a topic the minimal score. We thus obtained the upper and the lower bounds of the human judgment and compared the models.

For every model, Table 4 also shows the average tf-idf coherence metric. Note that although our results match previous experiments regarding the comparison between coherence and tf-idf coherence well (correlation with tf-idf coherence is in our experiments approximately 10-12% better

Table 4. Experimental results: general interpretability and coherence for partially, highly, and generally interpretable models

	#	coh_{10}	tfidf_{10}	coh_{20}	tfidf_{20}
Partially interpretable topics					
1 (plsa)	139	-258.7	-145.3	-1145.9	-696.9
2 (lda)	192	-274.9	-163.3	-1224.1	-777.5
5 (restricted dict.)	237	-284.6	-163.0	-1247.9	-768.8
6 (extended dict.)	146	-258.6	-141.2	-1156.0	-686.1
7 (recursive)	239	-281.9	-166.3	-1235.7	-788.1
8 (keyword)	114	-256.3	-140.2	-1141.4	-682.8
Highly interpretable topics					
1 (plsa)	119	-318.0	-206.6	-1414.7	-982.5
2 (lda)	120	-389.5	-273.1	-1743.7	-1324.6
5 (restricted dict.)	87	-330.7	-227.0	-1410.7	-1028.2
6 (extended dict.)	103	-313.8	-199.9	-1372.6	-936.4
7 (recursive)	58	-349.2	-241.1	-1498.1	-1086.1
8 (keyword)	106	-310.0	-198.9	-1354.3	-914.8
Both partially and highly					
1 (plsa)	258	-286.0	-173.6	-1269.9	-828.7
2 (lda)	312	-319.0	-205.5	-1424.0	-988.0
5 (restricted dict.)	324	-297.0	-180.2	-1291.6	-838.5
6 (extended dict.)	249	-281.5	-165.5	-1245.6	-789.6
7 (recursive)	297	-295.1	-180.9	-1287.0	-846.3
8 (keyword)	220	-282.2	-168.5	-1244.0	-794.6

than correlation with standard coherence), still in this study human judgments correlate with tf-idf coherence only at the level of approximately 0.5, so there is still a long way ahead to develop better quality measures.

Since the models we test here all attempt to extract a certain number of high-quality topics while filtering out "trash" topics into a specially created "ghetto", it makes little sense to compare the models by the overall quality of all topics. It is much more important to look at the coherence of those topics that were found either good or relevant by the assessors.

4.4 Relevance and Coherence Scores

Table 4 summarizes the most important results on quality understood as interpretability (question 2) and its relation to tf-idf coherence. In this table, "partially interpretable" topics are those that were scored "1" by at least one of the assessors answering question 2; "highly interpretable" are those that were scored "2" respectively (but it is enough for only one assessor to give the high mark, i.e. this is the optimistic evaluation). The two leaders are models 5 and 6 (restricted dictionary and extended dictionary, respectively). We can see in Table 4 that model 6 (extended dictionary) outperforms all the rest by the overall quality, that is, by coherence and tf-idf coherence calculated over all topics. Model 5 (restricted dictionary) does produce higher values of coherences and tf-idf coherences in the groups of interpretable topics, but note that the number of interpretable topics is lower. This means that model 5 finds fewer topics, but the topics it finds are on average better.

Table 5 summarizes our most important findings regarding how relevant the topics are to our goal. "Partially relevant" topics are those that were scored "1" by at least one of the assessors answering questions 5 and 6; "highly relevant" are, respectively, those that were scored "2" by at least one assessor. "All relevant" topics in Table 5 include topics that are either partially or highly relevant to either ethnicity or international relations. Average interpretability was calculated as the mean evaluation scores given to the respective topics by assessors answering question 2. Here we again see the same two leaders, models 5 and 6, and the former outperforms the latter in terms of tf-idf coherence of relevant topics, while the latter outperforms the former in terms of the number of topics considered relevant by the assessors. This is true both for ethnic and international relations topics, and for both levels of relevance. This means that our extension of the seed dictionary brings more topics found by assessors both generally interpretable and relevant to both international relations and ethnicity, although average coherence of these topics becomes somewhat lower. Ethnic topics, thus, do not get substituted by or lost among topics on international relations.

Table 6 shows human-evaluated interpretability of the topics: it shows the average score given by the assessors to topics from each subset and for the two general questions, e.g., the top left corner shows that on average, assessors scored 1.80 on question 1 (general interpretability) for topics that are highly relevant to ethnic issues. Note that, interestingly, now model 6 outperforms model 5 in terms of interpretability: according to this measure, in model 6 relevant topics are

Table 5. Topics' relevance and coherence

Topics	Partially relevant					Highly relevant					Both partially and highly				
	#	coh_{10}	tfidf_{10}	coh_{20}	tfidf_{20}	#	coh_{10}	tfidf_{10}	coh_{20}	tfidf_{20}	#	coh_{10}	tfidf_{10}	coh_{20}	tfidf_{20}
1 (plsa)															
ethnic	5	-313.2	-190.2	-1399.2	-904.8	12	-334.0	-207.1	-1480.9	-996.3	17	-327.9	-202.1	-1456.9	-969.4
IR	20	-279.1	-150.7	-1227.0	-733.8	19	-315.3	-194.0	-1410.7	-946.8	39	-296.8	-171.8	-1316.5	-837.6
all relev.	20	-289.6	-163.0	-1271.2	-784.9	25	-315.9	-194.3	-1408.0	-938.7	45	-304.2	-180.4	-1347.2	-870.3
2 (lda)															
ethnic	2	-239.7	-124.4	-1158.5	-646.0	13	-306.8	-190.0	-1369.1	-927.9	15	-297.9	-181.3	-1341.0	-890.3
IR	21	-285.1	-158.9	-1266.2	-763.1	29	-353.3	-225.7	-1580.6	-1097.5	50	-324.7	-197.7	-1448.6	-957.1
all relev.	18	-289.4	-162.3	-1287.3	-777.7	37	-336.3	-212.2	-1496.3	-1023.0	55	-320.9	-195.9	-1427.9	-942.7
5 (restricted dictionary)															
ethnic	18	-288.7	-164.7	-1264.2	-798.5	30	-331.6	-222.3	-1419.0	-1015.8	48	-315.5	-200.7	-1360.9	-934.3
IR	33	-269.1	-142.5	-1190.8	-707.7	26	-323.1	-207.4	-1358.1	-917.3	59	-292.9	-171.1	-1264.5	-800.1
all relev.	36	-267.2	-142.0	-1177.6	-695.1	47	-322.7	-211.1	-1374.5	-958.4	83	-298.7	-181.1	-1289.1	-844.2
6 (extended dictionary)															
ethnic	8	-288.4	-160.5	-1315.2	-805.1	22	-280.7	-150.0	-1226.8	-713.8	30	-282.8	-152.8	-1250.4	-738.2
IR	18	-250.0	-126.3	-1130.6	-641.1	29	-287.4	-156.3	-1240.9	-740.8	47	-273.1	-144.8	-1198.7	-702.6
all relev.	22	-261.2	-136.5	-1199.9	-707.7	37	-285.5	-158.3	-1234.6	-741.8	59	-276.4	-150.2	-1221.7	-729.1
7 (recursive)															
ethnic	18	-308.2	-181.3	-1418.7	-952.6	22	-320.1	-201.8	-1431.0	-971.4	40	-314.7	-192.6	-1425.5	-962.9
IR	30	-283.3	-161.6	-1236.8	-780.4	30	-291.4	-171.4	-1292.9	-827.3	60	-287.4	-166.5	-1264.9	-803.9
all relev.	34	-285.4	-161.3	-1269.0	-810.6	47	-299.0	-180.1	-1331.3	-869.8	81	-293.3	-172.2	-1305.1	-844.9
8 (keyword)															
ethnic	5	-289.7	-161.1	-1315.9	-805.0	37	-297.9	-175.6	-1318.9	-834.7	42	-297.0	-173.9	-1318.6	-831.1
IR	18	-264.7	-138.4	-1168.7	-670.7	32	-278.5	-164.3	-1240.7	-782.9	50	-273.5	-155.0	-1214.8	-742.5
all relev.	17	-279.5	-154.3	-1230.7	-741.3	52	-282.5	-165.5	-1260.1	-793.1	69	-281.8	-162.8	-1252.8	-780.4

not only more numerous, but also slightly more interpretable than in model 5; however, fewer of them are clearly related to specific events (question 2). For sociologists, a larger number of relevant topics is an advantage since they are not very numerous anyway and can be double-checked for relevance and interpretability manually, while, had they been filtered out automatically, they may never be brought to the expert's attention, so model 6 looks preferable.

At the same time, the dictionary of model 6 has been situational: it substituted the missing ethnonyms with adjectives and country names, while the ethnic groups whose ethnonyms were present in the collection were not supplemented by adjectives or country names. This principle of dictionary construction means that different adjectives and country names should be excluded each time even if some of them are present in the collection. It also may have lead to some overfitting in our best model. To make this model more practical and the quality assessment more reliable, in the future we suggest to rerun it with the full dictionary of ethnonyms, adjectives and country names that will be made universal.

Interesting results are produced by models 7 (recursive) and 8 (keyword texts). By evaluating both the number of relevant topics and coherence, the recursive model looks similar to model 5 with restricted dictionary (fewer, but more coherent topics of interest); keyword-based model is similar to model 6 (more numerous and a little less coherent topics of interest) (see Table 3).

It, thus, means that re-iteration of topic modeling on a subset of texts extracted during the first iteration does not bring improvement, or even brings deterioration, and therefore is excessive and useless. In terms of numerical results, single-iteration modeling on a collection selected by keyword produces the results similar to or not dramatically worse than the best model (model 6), but the sets of ethnicity-related topics found by these two approaches are significantly different, so to get the best possible coverage one should probably use a combination of these techniques, one possible direction for further work.

Table 6. Interpretability results for the topics relevant for ethnic and international relations subjects

	Question 1			Question 2		
Topics	part.	highly	all	part.	highly	all
1 (plsa)						
ethnic	1.80	1.75	1.76	1.20	1.50	1.41
IR	1.90	1.68	1.79	1.75	1.26	1.51
all relevant	1.85	1.72	1.78	1.65	1.36	1.49
2 (lda)						
ethnic	2.00	1.92	1.93	2.00	1.62	1.67
IR	2.00	1.69	1.82	1.86	1.21	1.48
all relevant	2.00	1.76	1.84	1.83	1.32	1.49
5 (restricted dictionary)						
ethnic	2.00	1.40	1.62	1.89	1.27	1.50
IR	1.85	1.42	1.66	1.85	1.35	1.63
all relevant	1.89	1.45	1.64	1.86	1.32	1.55
6 (extended dictionary)						
ethnic	2.00	1.73	1.80	1.75	1.27	1.40
IR	1.94	1.72	1.81	1.72	1.17	1.38
all relevant	1.95	1.62	1.75	1.68	1.16	1.36
7 (recursive)						
ethnic	1.78	1.59	1.68	1.00	0.95	0.97
IR	1.87	1.87	1.87	1.43	1.20	1.32
all relevant	1.94	1.72	1.81	1.35	1.09	1.20
8 (keyword)						
ethnic	2.00	1.76	1.79	1.20	0.89	0.93
IR	1.94	1.91	1.92	1.33	1.16	1.22
all relevant	1.94	1.83	1.86	1.41	1.08	1.16

4.5 Prefiltering and Two-Stage Topic Modeling

In the final series of computational experiments, we tested a natural extension of the ideas expressed in previous models: to filter the original collection with respect to the resulting subject topics and try topic modeling again. To test this idea, we have chosen documents from the original collection that contained top words from subject topics discovered on the previous step. Then, the much reduced collection was again subject to topic modeling; in this experiment, we have compared several variations of ARTM models. The reduced collection contained approximately 320K documents with the same set of ethnonyms as the large models.

The reduced collection has allowed us to perform a large-scale comparison of ARTM models with different parameters. In the paper, we show a sample of nine models with characteristic parameters that may result in different behaviour. Table 7 shows their parameters; note that model 9 has the same parameters as model 8 but has been trained for three epochs over the entire dataset compared to a single pass in model 8.

Table 7. Second stage topic models. Model 9 is the same as model 8 but with three passes instead of one

#	Parameters							
	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8
1	10^{-4}	-10^2	0.05	-1.0	0	10^{-7}	1.0	0
2	0	0	0	0	0	0	100.0	0
3	$2.5 \cdot 10^{-4}$	-10^2	0.05	-1.0	0	10^{-7}	1.0	0
4	10^{-4}	-10^2	0.05	-1.0	0	10^{-7}	100.0	0
5	$2.5 \cdot 10^{-4}$	-10^2	0.05	-1.0	0	10^{-7}	100.0	0
6	$5 \cdot 10^{-3}$	-10^2	0.05	-1.0	0	10^{-7}	100.0	0
7	$5 \cdot 10^{-5}$	-10^2	0.05	-1.0	$2 \cdot 10^5$	10^{-7}	100.0	$2 \cdot 10^4$
8	$5 \cdot 10^{-5}$	-10^2	0.05	-1.0	10^5	10^{-7}	100.0	10^4
9	$5 \cdot 10^{-5}$	-10^2	0.05	-1.0	10^5	10^{-7}	100.0	10^4

To make the results comparable with full models, we have trained all models with the same number of topics, 250 subject (ethnic) topics and 150 general (background) topics and computed coherence scores on the entire dataset rather than the reduced one (those scores would, naturally, be much better). Table 8 shows coherence results for new models. The top nine rows show average coherence scores for all topics and can be directly compared with Table 2; we see that the best second-stage models. models 4 and 5, have better coherence scores than the best first-stage models from Table 2. Comparing models 8 and 9, we also see that additional passes over the corpus do indeed improve the topics but only very slightly, so in case of a large corpus, when it is costly to double or triple the training time, one pass should be sufficient.

Table 8 also provides separate average estimates for coherences and tf-idf coherences of subject (ethnic) and background (general) topics. Note an interesting effect: background topics have consistently better scores than subject topics across all models. This is due mainly to the fact that we have chosen a far larger number of ethnic topics (250) than necessary since we need to make sure all ethnic topics are captured by the model, and a false positive (a junk ethnic topic) is not a problem. We show some sample topics from one of the best second stage models in Table 9. While ethnic topics do indeed have plenty of good ethnic- or nationality-related topics, they also have a lot of uninterpretable junk topics (e.g., topics 92 and

Table 8. Second stage models: coherence and tf-idf coherence

#	T	coh_{10}	tfidf_{10}	coh_{20}	tfidf_{20}
All topics					
1	400	-367.6	-259.3	-1587.0	-1203.5
2	400	-328.1	-215.6	-1451.4	-1015.7
3	400	-367.4	-258.8	-1589.6	-1210.6
4	400	-299.3	-191.3	-1289.0	-869.8
5	400	-299.3	-191.2	-1289.9	-870.1
6	400	-329.0	-220.0	-1417.6	-1011.8
7	400	-365.2	-286.9	-1548.7	-1296.4
8	400	-353.6	-264.5	-1519.9	-1223.8
9	400	-351.3	-256.6	-1518.4	-1199.5
Subject (ethnic) topics					
1	250	-432.7	-323.1	-1865.3	-1505.1
2	250	-319.3	-208.1	-1411.7	-980.3
3	250	-432.8	-322.6	-1871.2	-1517.6
4	250	-312.6	-198.4	-1343.3	-909.6
5	250	-313.0	-198.6	-1347.1	-912.7
6	250	-366.0	-251.9	-1581.9	-1171.1
7	250	-424.6	-358.1	-1797.0	-1626.7
8	250	-404.6	-320.3	-1748.1	-1506.1
9	250	-406.7	-313.3	-1770.3	-1491.6
Background (general) topics					
1	150	-258.9	-152.9	-1123.2	-701.0
2	150	-342.6	-228.1	-1517.7	-1074.9
3	150	-258.5	-152.5	-1120.2	-699.0
4	150	-277.3	-179.6	-1198.6	-803.6
5	150	-276.3	-178.9	-1194.6	-799.2
6	150	-267.3	-166.9	-1143.8	-746.2
7	150	-266.3	-168.3	-1134.8	-746.0
8	150	-268.7	-171.5	-1139.5	-753.4
9	150	-259.1	-162.0	-1098.4	-712.8

232 in Table 9); at the same time, background topics are not ethnic-related but are indeed more coherent on average.

5 Conclusion

In this work, we have shown that *additive regularization of topic models* (ARTM) can provide social scientists with an effective tool for mining specific topics in large collections of user-generated content. Our best model has outperformed basic LDA both in terms of the number of relevant topics found and in terms of their quality, as it was found in experiments with topics related to ethnicity.

What is especially important for digital humanities, additive regularization allows one to easily construct nontrivial extensions of topic models without mathematical research or software development. By combining built-in regularizers from the BigARTM library, one can get topic models with desired properties. In this work, we have combined eight regularizers and constructed a topic model for exploratory search that can take a long list of keywords (in our case, ethnonyms) as a query and output a set of topics that encompass the entire relevant content. This model can be used to explore narrow subject domains in large text collections. In general, this study shows that ARTM provides unprecedented flexibility in constructing topic models with given properties, outperforms existing LDA implementations in terms of training speed, and provides more control over the resulting topics. Both specific regularizers introduced here and the general ARTM approach can be used in further topical studies of text corpora concentrating on different subjects and/or desired properties of the topics.

However, further experiments are needed to make our comparisons more precise. First, it would be interesting to compare our best model with semi-supervised non-interval LDA, where, instead of ascribing small bunches of words to multiple small ranges of topics, the entire dictionary would be ascribed to a large range of topics (akin to ARTM-produced models). Second, as has been mentioned above, it would be interesting to experiment with the universal dictionary of ethnonyms, adjectives, and country names. Finally, the results should be tested for stability via multiple runs of each model; stability of topic models is an interesting problem in its own right [18]. In general, semi-supervised learning approaches exhibit a good potential for mining not only ethnicity-related topics but also other types of specific topics of which the end-users may have incomplete prior knowledge.

Acknowledgments

This work was supported by the Russian Science Foundation grant no. 15-18-00091.

References

1. **Agrawal, A., Fu, W., & Menzies, T.** (**2016**). What is Wrong with Topic Modeling? (and How to Fix it Using Search-based SE). *ArXiv e-prints*.

Table 9. Sample topics from second stage model 4

Topic no.	Top words
	Sample ethnic topics
(1)	Irish, Ireland, time, day, beer, saint, country, friend, place, good, life
(12)	migrant, Uzbek, Russian, Russia, work, Moscow, place, Uzbekistan, job, country, janitor
(14)	Scottish, Scotland, whiskey, drink, beer, time, bottle, place, good, century, English, day, measure
(173)	Syrian, Syria, weapon, militant, Ali, army, Damascus, terrorist, region, mountain, military
(92)	mother, unroll, client, fabric, deer, Kupriyanovich, Putinga, orthopedic, Jehova, Marfino, rounding
(232)	sort, NSA, travel, Jean, Krasnovka, cezve, Soviet, oyster, Krasnodar, torture, Tashkent
	Sample general topics
(35)	woman, man, family, girl, female, life, beautiful, wife, red
(48)	price, cost, real estate, rent, buying, buyer, good, average, product, square
(36)	hospital, medical, operation, patient, clinic, medicine, healthy, cure, public health
(99)	game, team, play, player, season, soccer, stadium, win, soccer^{adj}, championship, sport, fan
(101)	color, red, black, green, white, blue, flower, color^{adj}, shade, place, pink

2. **Andrzejewski, D. & Zhu, X.** (**2009**). Latent Dirichlet allocation with topic-in-set knowledge. *Proc. NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, SemiSupLearn'09, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 43–48.
3. **Andrzejewski, D., Zhu, X., & Craven, M.** (**2009**). Incorporating domain knowledge into topic modeling via Dirichlet forest priors. *Proc. 26th Annual International Conference on Machine Learning*, ICML'09, ACM, New York, NY, USA, pp. 25–32.
4. **Apishev, M., Koltsov, S., Koltsova, O., Nikolenko, S. I., & Vorontsov, K.** (**2016**). Additive regularization for topic modeling in sociological studies of user-generated texts. *Proc. 15th Mexican International Conference on Artificial Intelligence*.
5. **Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W.** (**2009**). On smoothing and inference for topic models. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI'09, AUAI Press, Arlington, Virginia, United States, 27–34.
6. **Blei, D. M. & Lafferty, J. D.** (**2006**). Correlated topic models. *Advances in Neural Information Processing Systems*, 18.
7. **Blei, D. M. & Lafferty, J. D.** (**2006**). Dynamic topic models. *Proceedings of the* 23^{rd} *International Conference on Machine Learning*, ACM, New York, NY, USA, pp. 113–120. doi:10.1145/1143844.1143859.
8. **Blei, D. M. & McAuliffe, J. D.** (**2007**). Supervised topic models. *Advances in Neural Information Processing Systems*, 22.
9. **Blei, D. M., Ng, A. Y., & Jordan, M. I.** (**2003**). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5), 993–1022.
10. **Bodrunova, S., Koltsov, S., Koltsova, O., Nikolenko, S. I., & Shimorina, A.** (**2013**). Interval semi-supervised LDA: Classifying needles in a haystack. *Proc. 12th Mexican International Conference on Artificial Intelligence*, volume 8625 of *Lecture Notes in Computer Science*, Springer, 265–274.
11. **Chang, J. & Blei, D. M.** (**2010**). Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4(1), 124–150.
12. **Chemudugunta, C., Smyth, P., & Steyvers, M.** (**2007**). Modeling general and specific aspects of documents with a probabilistic topic model. *Advances in Neural Information Processing Systems*, volume 19, MIT Press, 241–248.
13. **Chen, X., Zhou, M., & Carin, L.** (**2012**). The contextual focused topic model. *Proceedings of the* 18^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, pp. 96–104. doi:10.1145/2339530.2339549.
14. **Griffiths, T. & Steyvers, M.** (**2004**). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (Suppl. 1), pp. 5228–5335.
15. **Hoffmann, T.** (**2001**). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1), 177–196.
16. **Hofmann, T.** (**1999**). Probabilistic latent semantic analysis. *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, pp. 289–296.

17. **Jagarlamudi, J., Daumé, H., III, & Udupa, R.** (**2012**). Incorporating lexical priors into topic models. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'12, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 204–213.
18. **Koltcov, S., Koltsova, O., & Nikolenko, S. I.** (**2014**). Latent dirichlet allocation: Stability and applications to studies of user-generated content. *Proceedings of the 2014 ACM conference on Web science (WebSci 2014)*, pp. 161–165.
19. **Koltsov, S., Nikolenko, S. I., Koltsova, O., Filipov, V., & Bodrunova, S.** (**2016**). Stable topic modeling with local density regularization. *Proc. 3rd international conference on Internet Science*, volume 9934 of *Lecture Notes in Computer Science*, Springer, pp. 176–188.
20. **Li, S. Z.** (**2009**). *Markov Random Field Modeling in Image Analysis*. Advances in Pattern Recognition, Springer, Berlin Heidelberg.
21. **Lin, C., He, Y., Everson, R., & Ruger, S.** (**2012**). Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering*, 24(6), 1134–1145. doi:10.1109/TKDE.2011.48.
22. **Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A.** (**2011**). Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 262–272.
23. **Nikolenko, S. I.** (**2015**). SVD-LDA: Topic modeling for full-text recommender systems. *Proc. 14th Mexican International Conference on Artificial Intelligence*, volume 9414 of *Lecture Notes in Computer Science*. Springer, pp. 67–79.
24. **Nikolenko, S. I., Koltsova, O., & Koltsov, S.** (**2015**). Topic modelling for qualitative studies. *Journal of Information Science*. doi:10.1177/0165551515617393.
25. **Paul, M. J. & Dredze, M.** (**2014**). Discovering health topics in social media using topic models. *PLoS ONE*, 9(8).
26. (**2013**). *Sociopolitical processes in the internet*. Laboratory for Internet Studies. Internal report, National Research University Higher School of Economics, reg. no. 01201362573, Moscow.
27. **Tan, Y. & Ou, Z.** (**2010**). Topic-weak-correlated latent dirichlet allocation. *7th International Symposium Chinese Spoken Language Processing (ISCSLP)*, pp. 224–228.
28. **Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M.** (**2005**). Sharing clusters among related groups: Hierarchical Dirichlet processes. *Advances in Neural Information Processing Systems*, 17, 1385–1392.
29. **Tikhonov, A. N. & Arsenin, V. Y.** (**1977**). *Solution of ill-posed problems*, W. H. Winston, Washington, DC.
30. **Tutubalina, E. & Nikolenko, S. I.** (**2015**). Inferring sentiment-based priors in topic models. In *Proc. 14th Mexican International Conference on Artificial Intelligence*, volume 9414 of *Lecture Notes in Computer Science*, Springer, pp. 92–104.
31. **Vorontsov, K.** (**2014**). Additive regularization for topic models of text collections. *Doklady Mathematics*, 89(3), 301–304. ISSN 1064-5624.
32. **Vorontsov, K., Frei, O., Apishev, M., Romov, P., Suvorova, M., & Yanina, A.** (**2015**). Non-Bayesian additive regularization for multimodal topic modeling of large collections. *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, TM'15, ACM, New York, NY, USA, pp. 29–37.
33. **Vorontsov, K. V. & Potapenko, A. A.** (**2014**). Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. *AIST'2014, Analysis of Images, Social networks and Texts*, volume 436, Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), pp. 29–46.
34. **Vorontsov, K. V. & Potapenko, A. A.** (**2015**). Additive regularization of topic models. *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications*, 101(1), 303–323.
35. **Wang, C., Blei, D. M., & Heckerman, D.** (**2008**). Continuous time dynamic topic models. *Proceedings of the 24^{th} Conference on Uncertainty in Artificial Intelligence*.
36. **Wang, X. & McCallum, A.** (**2006**). Topics over time: a non-Markov continuous-time model of topical trends. *Proceedings of the 12^{th} ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA, pp. 424–433. doi:10.1145/1150402.1150450.
37. **Williamson, S., Wang, C., Heller, K. A., & Blei, D. M.** (**2010**). The IBP compound Dirichlet process

and its application to focused topic modeling. *Proceedings of the* 27^{th} *International Conference on Machine Learning*, pp. 1151–1158.

38. **Yohan, J. & H., O. A.** (**2011**). Aspect and sentiment unification model for online review analysis. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM'11, ACM, New York, NY, USA, pp. 815–824. doi:10.1145/1935826.1935932.

Murat Apishev is an M.Sc. student at the Moscow State University and Junior Developer at the Search Department, Yandex, Moscow, Russia. He received his B.Sc. degree from the Moscow State University at 2015. His research interests include machine learning, parallel algorithms, and topic modeling.

Sergei Koltcov is the Deputy Director of the Laboratory for Internet Studies and the Associate Professor at the Department of Applied Mathematics and Computer Science at the National Research University Higher School of Economics, St.Petersburg. He received his Ph.D. in physics from the Institute for Analytical Instrumentation of the Russian Academy of Science at St.Petersburg in 2000. His research interests include mathematical modeling in various fields: topic modeling, sentiment analysis, electronic/ionic optics, mass spectrometry, gas dynamics, and statistical physics.

Olessia Koltsova is the Director of the Laboratory for Internet Studies and Associate Professor at the Department of Sociology at the National University Higher School of Economics, St. Petersbugr. As an academic commited to interdicplinary data driven research, she leads various collective projects in the sphere of Internet and society, as well as in methods of large-scale automatic internet data analysis for social science. In recent years, she has published on online community structure, user content topical composition and sentiment, relation of internet to protests, electoral preferences, entrepreneurial success, and other topics. She is also the author of News Media and Power in Russia, Routledge, 2006.

Sergey Nikolenko is a Senior Researcher at the Laboratory for Internet Studies, National Research University Higher School of Economics, and Laboratory of Mathematical Logic at the Steklov Institute of Mathematics at St. Petersburg. He received his M.Sc. summa cum laude from St. Petersburg State University at 2005 and Ph.D. from the Steklov Institute of Mathematics at St. Petersburg at 2009. His research interests include networking algorithms and systems, machine learning and probabilistic inference, bioinformatics, and theoretical computer science.

Konstantin Vorontsov is the Head of Intelligent Systems Department at the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Professor at the Moscow Institute of Physics and Technology (State University), and Professor of the Russian Academy of Sciences. He received his Sc.D. from the Computing Center of RAS at 2010. His research interests include machine learning, information retrieval, generalization bounds, topic modeling, and exploratory search.

Article received on 25/06/2016; accepted on 20/08/2016.
Corresponding author is Murat Apishev.