ATLANTIS PRESS

Atlantis Highlights in Computer Sciences, volume 3

21st International Scientific Workshop on Computer Science and Information Technologies (CSIT 2019)

Use of Topic Modelling for Improvement of Quality in the Task of Semantic Search of Educational Courses

Ivan Nikolaev
Institute of Information Technology
Chelyabinsk State University
Chelyabinsk, Russia
ivan_nikolaev@csu.ru

Dmitry Botov
Institute of Information Technology
Chelyabinsk State University
Chelyabinsk, Russia
dmbotov@gmail.com

Yuri Dmitrin
Institute of Information Technology
Chelyabinsk State University
Chelyabinsk, Russia
dmitrinyuri@gmail.com

Julius Klenin
Institute of Information Technology
Chelyabinsk State University
Chelyabinsk, Russia
jklen@yandex.ru

Andrei Melnikov
Ugra Research Institute of Information Technologies
Khanty Mansiysk, Russia
MelnikovAV@uriit.ru

***Abstract*— This paper proposes an approach, improving the quality of the original educational course programmes semantic search algorithm, based on vector representations, produced by distributional semantic. The proposed approach works by providing an expert with interpretable topic filtering of courses in search results. Application of probabilistic topic modeling based on additive regularization ensures the interpretability of vector components in representations of texts, allowing the expert, in the process of exploratory search, to narrow down the set of relevant documents found previously by using the vector model. In our experiments, we study the applied task of educational course search, using current requirements of the labor market (requirements described in professional standards serve as search queries). The implementation of topic filtering is based on the open-source library BigARTM. We investigate the influence of hyperparameters and the choice of regularizers in the construction of a topic model on the improvement of quality of educational course semantic search using various vector models: word2vec, fasttext, TF-IDF are investigated.**

Keywords — topic modeling, topic filtering

I. INTRODUCTION

With the active development of massive open online course (MOOC) platforms and the increasing rate of change of the labor market needs as we transition into the digital economy, the task of information retrieval of educational content becomes increasingly important for the formation of educational programs and individual educational trajectories. Today teachers, educational programme developers, and students have to spend considerable effort and time to navigate the full diversity of open educational content, content by leading universities and the context of the ever-changing demands of the labor market, as well as rapid obsolescence of professional standards.

Previously, the authors of this study proposed a method of semantic search for training courses using neural network language models [1]. Various neural network models, such as word2vec and fasttext, were analyzed and compared with the statistical vector model TF-IDF. Averaged word2vec and fasttext showed the best results according to the MAP (Mean Average Precision) metric, while the fasttext also provided better quality ranking as measured by the nDCG (normalized discounted cumulative gain) metric.

In this paper, we propose a method of filtering the search results of online courses and academic disciplines, by applying topic modeling based on the open-source BigARTM library. Search for training courses can be performed under the specified requirements of professional standards, which will improve the relevancy of retrieved educational content despite significant differences of vocabularies between the educational field (educational programs, online courses) and the professional field (professional standards, job requirements).

II. APPLICATION OF TOPIC MODELLING TO INFORMATION SEARCH

Topic modeling is one of the applications of machine learning to statistical text analysis. Topic models explore the hidden thematic structure of the corpus, by proposing that topic t is a probability distribution $p(w|t)$ over words, while document d is a probability distribution $p(t|d)$ over topics t. Simply put, a topic is a collection of words, by looking at which, one could say which subject they describe. Using this approach, the number of topics becomes much smaller than the number of words in the document; therefore, a sufficiently strong compression of the document occurs, while the most essential information about its subject matter is preserved.

Currently, methods based on probabilistic topic modeling (PTM) are being actively developed. PTM defines each topic as a discrete probability distribution over a set of words, and each document, as a discrete probability distribution over a set of topics [2]. The PTM performs joint clustering or "soft-clustering" of documents and words on topics, which means that a document or a word can belong to several topics at the same time with different probabilities.

One of the applications of topic modeling is informational (exploratory) search [3, 4]. According to this concept, a document or a collection of documents acts as a

Copyright © 2019, the Authors. Published by Atlantis Press SARL.
This is an open access article under the CC BY-NC license 4.0 (http://creativecommons.org/licenses/by-nc/4.0/).

search query, and information about the subject area is generated as a response. This can be especially useful in a situation where it is not clear how to formulate a search query: in the case of the absence of words or the problem of homonyms. This approach in some cases allows forming a big picture of the subject area.

New approaches to information retrieval, based on topic modeling are currently being actively developed [5-9].

III. Overview of Topic Modelling Approaches

Currently, two methods and their variations are primarily used for probabilistic topic modeling: pLSA (probabilistic latent semantic analysis) and LDA (Latent Dirichlet allocation).

At the heart of pLSA [10] is an algorithm that associates hidden variables representing topics with each word and document, allowing each to be related to several topics at once, with a certain probability.

This approach is based on the formula of total probability and the hypothesis of conditional independence, according to which the distribution of terms in a document $p(w|d)$ is described by a probabilistic mixture of distributions of terms over topics $\varphi_{wt} = p(w|t)$ and topics over documents $\theta_{td} = p(t|d)$.

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

Formula (1) describes the process of generation for a collection of topics, given $p(w|t)$ and $p(t|d)$, and searches for φ_{wt} and θ_{td}, such that the model sufficiently accurately approximates the frequency estimates of conditional probabilities. The model parameters are found by solving the likelihood maximization problem (likelihood logarithm) (2-4):

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

$$\sum_{w\in W} \varphi_{wt} = 1;\ \varphi_{wt} \geq 0 \quad (3)$$

$$\sum_{t\in T} \theta_{td} = 1;\ \theta_{td} \geq 0 \quad (4)$$

This method has the following disadvantages:

- The model is prone to overfit and is not applicable to large datasets.
- It is impossible to calculate the probability for a document not from the original dataset.
- The learning process cannot be interpreted.

LDA [11] is a Bayesian version of the pLSA. This approach uses the Dirichlet a priori distribution as the topic distribution over documents and over words, which allows the model to generalize better.

LDA has the following disadvantages:

- It is not grounded in linguistics and does not model any linguistic phenomenon, and its application is dictated solely by the convenience of analytical integration in the Bayesian output;
- The bayesian approach is too complicated to combine more than 2–3 parameters.

The additive regularization approach to topic models (ARTM) [12] is based on the pLSA method. Problem (2) has an infinite set of solutions, and in the general case is solved by using the regularization mechanism, which works by adding weighted sums of regularizers to the optimization criterion. ARTM maximizes the linear combination of the likelihood logarithm (6) and regularizers $R_i(\Phi,\Theta)$ with non-negative coefficients $\tau_i, i = 1,\dots,k$ (5), with consideration for non-negativity and normalization (3, 4):

$$R_i(\Phi,\Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta) \quad (5)$$

$$\sum_{d\in D}\sum_{w\in d} n_{dw} ln \sum_{t\in T} \varphi_{wt}\,\theta_{td} \to max_{\Phi,\Theta} + R_i(\Phi,\Theta) \quad (6)$$

The authors of the ARTM, have proposed various sets of regularizers that increase the interpretability, sparsity, and variation in the topics, produced by the model. In a series of experiments, the level of quality was achieved, comparable if not superior to the common approaches of the word2vec family [13].

In [14], a search of online educational courses for a given text of the discipline's programme shows that both pLSA and LDA performed worse in the task of semantic search, than the word2vec word vectors, averaged across the document and weighted by IDF (inverse document frequency). However, the neural network language models have a drawback - the lack of interpretability of vector components. In this paper, we propose to use the ARTM algorithm for additional filtering, allowing for thematic interpretation.

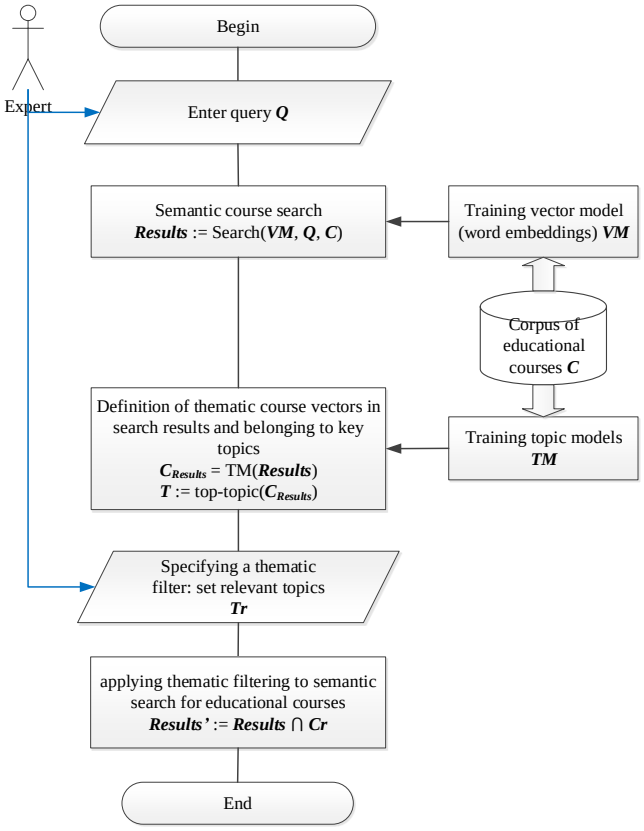

Fig. 1. Algorithm for applying thematic filtering to improve the quality of semantic search for educational courses

IV. TOPIC MODELING BASED METHOD OF SEMANTIC SEARCH FOR EDUCATIONAL COURSES

Figure 1 describes the proposed method, based on topic modeling as a thematic filter, used to search for training courses according to the queries set by an expert.

Model training stage. We train vector and topic models, using the corpus of educational courses. The vector model allows searching for courses based on cosine similarity with the query [1]. The topic model generates two matrices: word distributions over topics, and topic distributions over documents. Each course is assigned its topic vector—the distribution of the probabilities of assigning topics to the course (matrix td). An example of a thematic vector is presented in Figure 2. Each topic is characterized by its list of keywords that determine the meaning of this topic and allow the expert to assess the relevance of the topic to their search query. This stage is described in more detail in paragraph 4.

Search stage. At this stage, the expert sets the query used to search for semantically similar courses, using one of the vector models [1]. The result of this stage is the ranked list of N courses, most similar to the original query.

Thematic filter preparation stage. For each course in the search results, its topic vector is determined, which, in turn, is used to select a list of key topics—topics, which the topic model considered most likely for this course.

To determine the list of key topics, it is necessary to sort the topic vector in order of decreasing probabilities of topics and select the top N topics.

However, if the variance of the topic vector is not large enough, and the number of topics with a pronounced influence in the thematic vector is less than N, then the topics with relatively small influence could find their way into the list of key topics, which in turn can adversely affect the quality of thematic filtering. To avoid this, we propose the introduction of an additional lower constraint—a k-average of a topic vector. This parameter is determined as the average of the topic vector, multiplied by the coefficient k.

An experiment for the calculation of N and k_avg is described in paragraph 5.2.

As a result, for each query, a list of the most frequently encountered key course topics is compiled, which is displayed to the expert as a list of keywords. The expert can then select a number of the most relevant topics for a given query by keyword, and apply a thematic filter.

Filtering stage. For each course, a list of key topics is compared with a list of the most relevant topics, as selected by the expert. For the final list of search results, only courses, which have at least one relevant topic, are selected.

The described method of thematic filtering allows the expert to influence the final search results, and to order the courses based on the topics he selects. The main advantage of this approach is the interpretability of topics chosen by an expert, which allows for more intelligent and detailed control over the search.

V. TOPIC MODEL CONSTRUCTION

A. Description of the text corpus of educational courses used by the topic model

A text corpus of 3027 educational courses was used to train the topic model. The main characteristics of the dataset are presented in table 1.

Table 2 presents the detailed data on the MOOC platforms and the corpus of educational course programmes, collected from the websites of 10 large Russian universities for disciplines, related to information technology.

TABLE I. EDUCATIONAL CORPUS CHARACTERISTICS

Corpus	Number of sources	Number of documents (training courses)	Number of tokens	Number of unique tokens (vocabulary)
MOOC Corpus	4 MOOC-platforms	2051	249k	14k
Educational Courses Corpus	10 universities	976	351k	16k
Total	**14**	**3027**	**600k**	**23k**

TABLE II. DETAILING THE CORPORA OF ONLINE COURSES AND EDUCATIONAL COURSES

#	Stepik	OpenEDU	Coursera	Intuit	Educational courses
Number of documents	554	315	275	907	976
Number of tokens	39394	93866	39669	75942	351459
Number of unique tokens	6098	6574	6113	6693	15688
Average token count	71	297	144	83	360

B. Description of topic model training approach

All texts go through a preprocessing stage: deletion of non-textual characters, lowering character's case, lemmatization using pymorphy2 library, and removal of stopwords. After that, the texts are converted to UCI-Bag-of-words format acceptable by the BigARTM.

In the process of learning, a list of regularizers and a list of results (perplexion and sparsity of the word (Φ) and document matrices (Θ), purity and size of the core, coherence) are selected, which control the quality of the model's training.

Determining the number of topics

One of the most important parameters of a topic model is the number of topics into which words and documents are to be distributed. The number of topics should be sufficient to provide enough of a variety, yet not too large so that the topics retain enough specificity of the subject area and could be easily interpreted by the expert.

Interpretability is very difficult to formalize. In papers [14-16], the mechanism for assessing interpretability—coherence—is proposed. This estimate is based on a pointwise mutual information formula (PMI) of the k most likely words from a topic, that are found together in documents. This assessment is best correlated with expert estimates of interpretability. Model coherence is thus defined as the average for all topics.

1) Selection of regularizers

Combining and selecting regularizer parameters is a complex empirical task. By selecting a combination of a regularizer and regularization coefficients, it is possible to significantly improve sparseness, contrast, purity, and coherence at the same time with a slight loss of credibility of the model, as well as an improvement of the topic interpretability.

Smoothing and increasing the sparsity of matrices Φ and Θ (SmoothSparsePhi, SmoothSparseTheta), as well as decorrelation of terms (DecorrelatorPhi) are recommended as primary and most common regularizers. The increase in the sparsity of matrices Φ and Θ, leads to the reduction of weak components and an increase in the proportion of stronger ones. DecorrelatorPhi also acts to increase model sparsity, however, it does so in such a way that each topic aggregates words that are not found in other topics, making the topics differ more from each other.

A combination of these regularizers was used for topic exploratory search [17]. It turned out that it significantly improves the overall quality of search results.

To assess the impact of the regularization, 4 thematic models with different regularization parameters were used (Table 3).

Figure 2 describes an example of a topic vector for the course "Computer Science and Programming". Among the topics, that make the greatest contribution to this vector, are the following topics: 9, 16, 32, 48, 49. The list of keywords for these topics is provided in table 4.

Table 2 presents the detailed data on the MOOC platforms and the corpus of educational course programmes, collected from the websites of 10 large Russian universities for disciplines, related to information technology.

TABLE III. TOPIC MODELS

#	Title	Smooth Sparse Phi	Smooth Sparse Theta	Decorrelator Phi
1	60_0_0_0	0	0	0
2	60_-0,02_-0,03_2500	-0,02	-0,03	2500
3	60_-0,02_-0,5_5000	-0,02	-0,5	5000
4	60_-0,02_-1_10000	-0,02	-1	10000

C. Effect of regularization

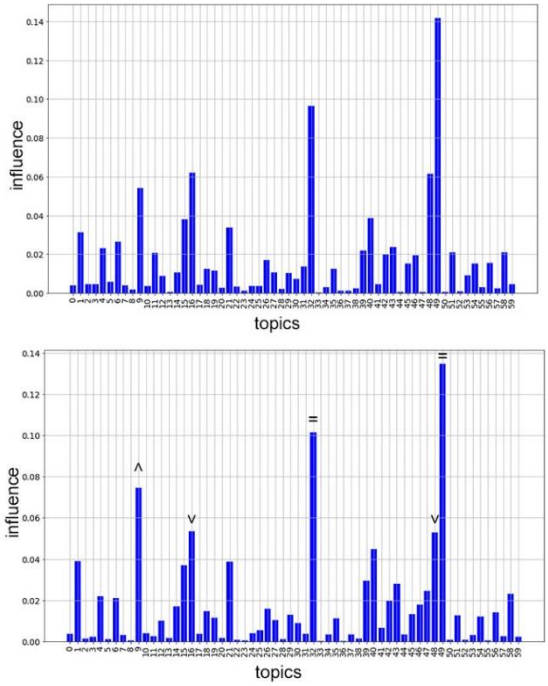

Fig. 2. Topic vector for the course "Computer Science and Programming" for model 1 (60_0_0_0) (top) and model 2 (60_-0,02_-0,3_2500) (bottom)

TABLE IV. KEY WORDS FOR KEY TOPICS FOR THE COURSE "COMPUTER SCIENCE AND PROGRAMMING"

№ theme	Russin	English
9	класс, ориентировать, язык, программа, объектный, среда, java, основа, разработка, объект	class, orient, language, program, object, environment, java, basis, development, object
16	базовый, основа, изучение, основный, практический, сложный, использовать, программа, студент, практика	basic, basis, learning, basic, practical, complex, use, program, student, practice
32	алгоритм, структура, сложность, реализация, поиск, дерево, задача, алгоритмический, сортировка, операция	algorithm, structure, complexity, implementation, search, tree, task, algorithmic, sorting, operation
48	дисциплина, область, метод, навык, студент, цель, умение, вид, основный, формирование	discipline, area, method, skill, student, goal, skill, type, basic, formation
49	программирование, язык, функциональный, программа, тип, введение, стандартный, пример, выражение, структура	programming, language, functional, program, type, introduction, standard, example, expression, structure

D. Assessing quality of the topic model

To assess the quality of topic models, a combined approach of internal and external metrics is used.

The main internal quality estimates of the model are, the sparsity of matrices Φ and Θ, and parameters of the topic cores (topic_kerrnel_score).

Perplexity shows the rate of convergence for the model. It is based on the likelihood logarithm and characterizes how well the corpus is described by the model.

Figure 3 shows that while learning, perplexity converges fairly quickly. This is because the size of the dataset is relatively small, and the topic model manages to quickly learn a given number of topics. Figure 4 shows a coherence plot for different numbers of topics. Models with the size of topic vector 40 and 60 achieved the best result. According to the aggregate estimation of experts and the average coherence plots, the total size of the topic vector was set to 60. This size of the topic vector allows obtaining sufficiently informative and human-interpretable topics. Examples of keywords are provided in table 4.

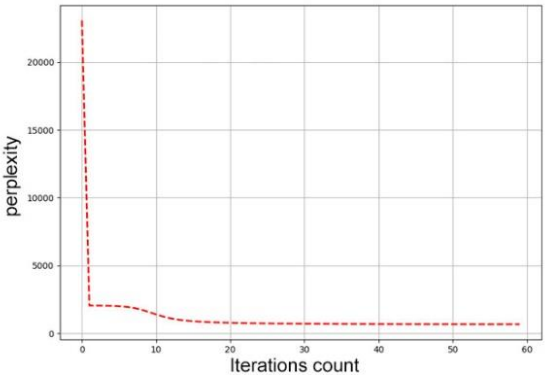

Fig. 3. Perplexity to vector size for basic topic model 0

Figure 5 presents the evaluation of the sparsity of matrices Φ and Θ (SparsityPhi on the left and SparsityTheta on the right). From the results of our experiments, we can note that the value of the SmoothSparsePhi regularizer below -0.02 does not lead to a greater increase in the sparsity of the matrix Φ. At the same time, if SmoothSparsePhi is within the range [0, -1], it allows for the sparsity of the matrix Θ to reach 49%.

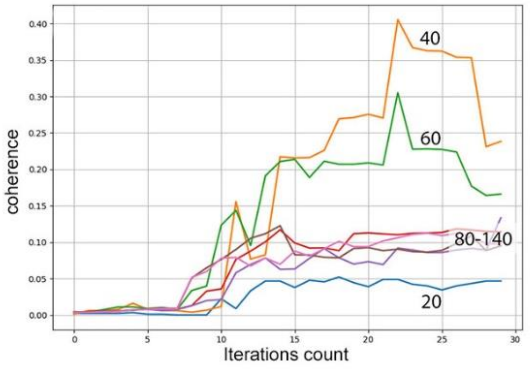

Fig. 4. Average coherence to vector size for basic topic model 0

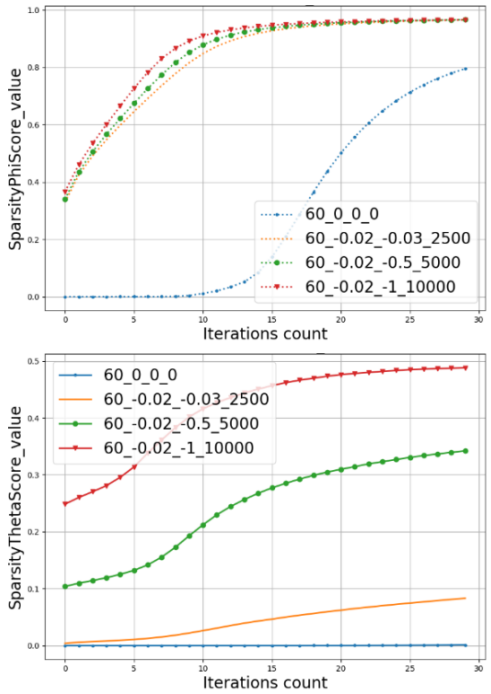

Fig. 5. Sparsity for matrices Φ (phi) (top) and Θ (theta) (bottom)

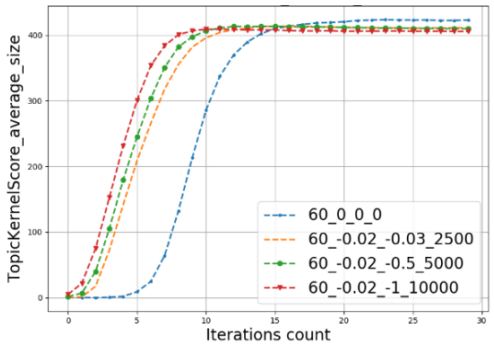

Fig. 6. Average topic core size depending on the degree of regularization

Figures 6 and 7 show that regularization has almost no effect on the average core size of topics, and holds very little improvements to the average contrast and purity of topics. This suggests that even without the regularization of the topic model, we managed to learn quite well from this corpus of text documents.

External evaluation of the quality of the topic model include the degree of suitability of the topic model to some "external" task; in our case, this task was the task of the semantic search for educational courses, fitting the requirements of professional standards. The following describes an experiment on the external quality assessment of the topic model applied to this task.

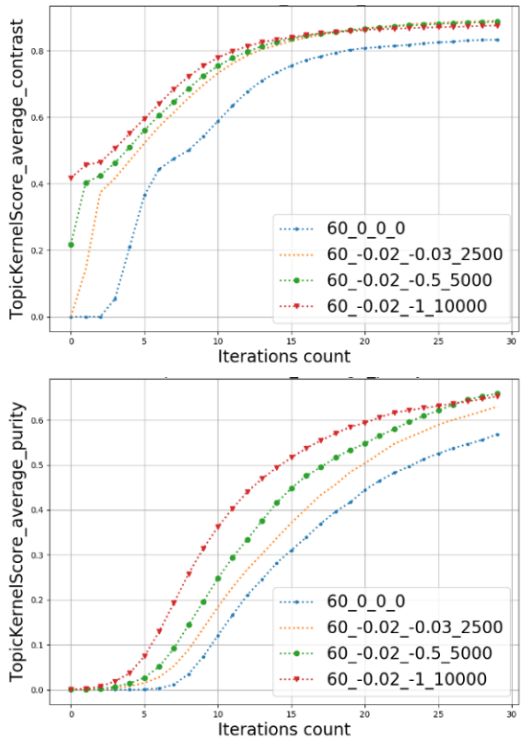

Fig. 7. Average contrast (top) and average purity (bottom) depending on the degree of regularization

VI. Evaluation of Topic Filtering in the Task of Semantic Search for Courses

A. Dataset description

To evaluate the quality of the topic model, we used a dataset, which was assembled in 2 stages.

During the first stage, several vector models were prepared (see Table 1), which were used to determine the list of the most similar courses for each of the queries. The queries were the general functions of the three professional standards (professions): "Programmer", "Information Systems Specialist", "Information and Communication Systems Systems Administrator" (hereinafter referred to as queries).

TABLE V. GRADED DATASET

Model	Unique queries <90%	Grades course-query
w2v_d300	17 / 28	536
tfidf	22 / 28	472
fasttext	12 / 28	407

For the second stage, the experts evaluated the similarity in course-query pairs for each vector model (using a five-point scale: 5 - texts are similar, 1 - texts are very dissimilar). For each query, lists of relevant (with estimates 5, 4, 3) and irrelevant courses (with estimates 2, 1) were formed. To evaluate the thematic filter, we have only used those queries, which had the proportion of relevant courses less than 90%. (see table 5), i.e. queries in which vector models made the greatest number of errors.

B. Description of the evaluation method

To evaluate the quality of the topic models as a thematic filter, we propose to consider a selection of relevant courses and the filtering of irrelevant courses as an assessment of the quality in the task of classifying courses by relevance for topic models with varying degrees of regularization.

Topic relevance class formation stage. For each query, based on expert's assessments, the course was assigned a label "relevant" (for grades 3,4,5) or "irrelevant" (1,2). Then all the key topics of all courses within the frame of the query were combined into a new aggregated topic vector of the query (Fig. 8) by formula 7.

$$c \in C_R \implies (c, t_c) \in P_R \; ; \; c \in C_I \implies (c, t_c) \in P_I \quad (7)$$

where c - course query, C_R - set of relevant courses, C_I - set of irrelevant courses, t_c - key topic of the course, P_R - set of relevant pairs (c, t_c) - for topic t of aggregated query topic vector. P_I - set of irrelevant pairs (c, t_c) - for topic t of aggregated query topic vector.

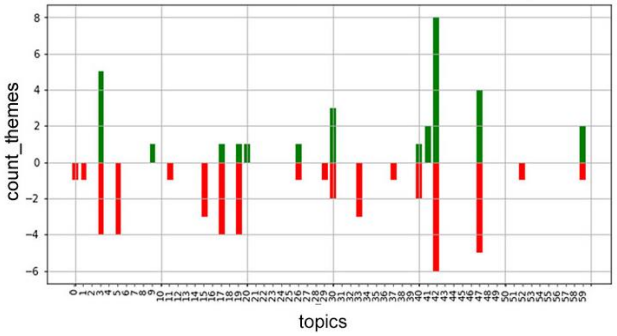

Fig. 8. Aggregated query topic vector

Each element of this aggregated topic vector can be assigned one of the 4 classes of topic relevance:

If $(|P_R| \geq |P_I| \;\&\; |P_I| = 0) \implies t \in T_R$, where t - a topic of a query topic vector, T_R - set of topics, consisting only of key topics of relevant courses.

If $(|P_R| \geq |P_I| \;\&\; |P_I| \neq 0) \implies t \in T_{R+}$, where t - a topic of a query topic vector, T_{R+} - set of topics, consisting mostly of key topics of relevant courses.

If $(|P_R| < |P_I| \;\&\; |P_I| = 0) \implies t \in T_I$, where t - a topic of a query topic vector, T_I - set of topics, consisting only of key topics of irrelevant courses.

If $(|P_R| < |P_I| \;\&\; |P_I| \neq 0) \implies t \in T_{I-}$, where t - a topic of a query topic vector, T_{I-} - set of topics, consisting mostly of key topics of relevant courses.

After this, the courses of the query are clustered based on the clustering of the key topics of the course by the relevance classes of topics (8-11).

$$c \in C_R \;\&\; \exists t_c \in (T_R \cup T_{R+}) \implies c \in C'_R \quad (8)$$

$$c \in C_R \;\&\; \forall t_c \in T_{I-}) \implies c \in C'_{R-} \quad (9)$$

$$c \in C_I \;\&\; \exists t_c \in T_{R+} \implies c \in C'_{I+} \quad (10)$$

$$c \in C_I \;\&\; \forall t_c \in (T_I \cup T_{I-}) \implies c \in C'_I \quad (11)$$

TABLE VI. THE DISTRIBUTION OF COURSES BY TOPIC RELEVANCE CLASSES

		actual	
		Relevant	***Irrelevant***
predicted	***Relevant***	C'_R	C'_{I+}
	Irrelevant	C'_{R-}	C'_I

Then, the main metrics for evaluating the quality of clustering were calculated (12-15):

$$accuracy = (C'_R + C'_I)/(C_R + C_I) \quad (12)$$

$$recall = C'_R/C_R \quad (13)$$

$$precision = C'_R / (C'_R + C'_{I+}) \quad (14)$$

$$\text{F1\_score} = 2 * (precision * recall)/(precision * recall) \quad (15)$$

As a base evaluation, we used the ratio of relevant courses to the entire set of courses was used:

$$\text{init\_}precision = C_R / (C_R + C_I) \quad (16)$$

C. *Determining hyperparameters for the calculation of the number of key course topics*

The results of the hyperparameter selection for the topic model, according to the method described in paragraph 5.2.1, for model 3 with average regularization.

From the plots in figure 9 we note optimal criteria for key topics of courses selection: max_themes = 3, k_avg = 3. The described method allows for the selection of the most effective number of key topics for each course.

D. *Evaluation of semantic search for educational courses with thematic filtering*

Based on the method described in paragraph 5.2. we calculate the quality of topic models with different degrees of regularization in comparison with vector models. The results are presented in Table 7.

From Table 7 we can see the effect regularization has on improving the quality, however, if thematic models 1, 2, and 3 with moderate regularization show a stable quality improvement, then thematic model 4 with high regularization parameters for the TF-IDF vector model shows a slight degradation. It can also be noted that the lower the base estimate of the vector model (init_precision) is, the higher the gain from the use of topic modeling in these models becomes, i.e. it can be said that topic modeling in several cases has corrected the deficiencies of vector models.

Figure 10 shows, using the example of topic models with different degrees of regularization, the effect of regularization on the final quality of the thematic classification of courses according to classes of topics. It can be noted that moderate regularization leads to an improvement in the quality of the topic model, while excessive regularization can lead to a decrease in quality.

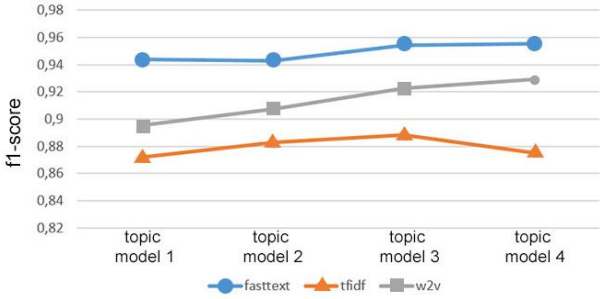

Fig. 9. Average F1-score for different topic models

Table 8 presents the results of issuing the fasttext vector model and thematic filter for the request “Software Design”.

VII. CONCLUSION

In this paper, a thematic filtering method based on the ARTM topic modeling algorithm is proposed. We show that the use of topic modeling as a thematic filter improves the quality of the semantic search for educational courses for specific requirements of professional standards. Results show that the improvement in the search task occurred for all vector models, and the more the vector model was mistaken before applying the filter, the more tangible was the positive effect of it. The use of moderate regularization during the training of the topic model also has a positive effect on the quality of thematic filtering. In the future, the authors plan to develop a thematic filtering algorithm and add possible ranking of courses in search results by specifying not only relevant but also irrelevant topics.

ACKNOWLEDGMENTS

The research is carried out with the support of the Russian Foundation for Basic Research in the course of the project №18-47-860013 p_a “Intelligent system of formation of educational programs based on neural network models of natural language, taking into account the needs of the digital economy of the Khanty-Mansiysk Autonomous Okrug—Ugra” (contract №18-47-860013\18).

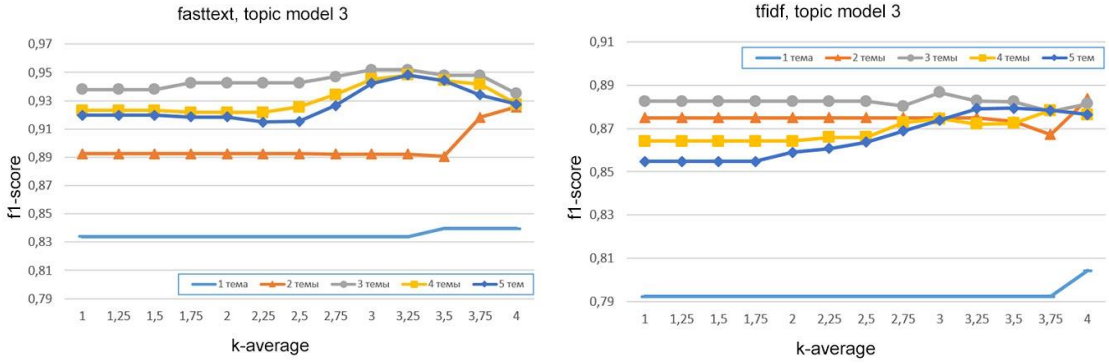

Fig. 10. f1-score to the number of topics (left) and to the k-average parameter (right)

TABLE VII. AVERAGE ESTIMATES OF THEMATIC CLASSIFICATION FOR QUERIES

Topic model	Vector model	F1-score	Precision	Recall	Average topic count C_R+C'_R
Model 1	fasttext	**0,9437 (+0,1526)**	**0,9289** (+0,2691)	**0,9673 (-0,0327)**	**9,00**
	tfidf	0,8717 (+0,1095)	0,8432 (**+0,2818**)	0,9172 (-0,0828)	10,14
	w2v	0,8956 (+0,052)	0,8828 (+0,1395)	0,9484 (-0,0516)	13,59
Model 2	fasttext	**0,9426 (+0,1516)**	**0,9164** (+0,2626)	**0,9792 (-0,0208)**	**9,08**
	tfidf	0,8826 (+0,1203)	0,8304 (**+0,2873**)	0,9473 (-0,0527)	10,32
	w2v	0,9074 (+0,0637)	0,8775 (+0,1411)	0,9614 (-0,0386)	13,71
Model 3	fasttext	**0,9544 (+0,1634)**	**0,9262** (+0,2735)	**0,9896 (-0,0104)**	**9,17**
	tfidf	0,8884 (+0,1261)	0,8619 (**+0,3007**)	0,9223 (-0,0777)	10,18
	w2v	0,9225 (+0,0789)	0,8835 (+0,1551)	0,972 (-0,028)	13,76
Model 4	fasttext	**0,9555 (+0,1645)**	0,9163 (+0,2777)	1,0000 (0,0000)	**9,25**
	tfidf	0,8753 (+0,1131)	0,8525 (**+0,2872**)	0,9162 (-0,0838)	10,18
	w2v	0,929 (+0,0854)	0,8876 (+0,1606)	0,9785 (-0,0215)	13,82

TABLE VIII. RESULTS OF RUNNING THE FASTTEXT VECTOR MODEL AND THEMATIC FILTER FOR THE REQUEST "ПРОЕКТИРОВАНИЕ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ" ("SOFTWARE DESIGN")

Name_course	Filtered
Анализ и проектирование систем управления (Analysis and design of control systems)	+
Технологии разработки программного продукта (Software development technologies)	+
Анализ требований к программному обеспечению (Analysis of software requirements)	+
Проектирование программных систем (Design of software systems)	+
Технологии разработки программного обеспечения (Software development technologies)	+
Программная инженерия (Software engineering)	+
Верификация и аттестация программного обеспечения (Software verification and certification)	-
Тестирование программного обеспечения (Software testing)	-
Тестирование и отладка программного обеспечения (Software testing and debugging)	-
Методы тестирование программного обеспечения (Software testing methods)	-
Тестирование программного обеспечения (Software testing)	-
Документирование и сертификация (Documentation and certification)	-

REFERENCES

[1] Botov D.S. Semantic search of educational courses according to job market requirements using neural language models. [Semanticheskiy poisk uchebnykh distsiplin pod trebovaniya rynka truda na osnove neyrosetevykh modeley yazyka] / D.S. Botov, Yu.V. Dmitrin, Yu.D. Klenin // SUSU Herald, "Computer technologies, management, radio electronics" series. [Vestnik YuUrGU. Seriya «Komp'yuternye tekhnologii, upravlenie, radioelektronika»]. – 2019. – T. 19, № 2. – S. 5–14.

[2] Daud A., Li J., Zhou L., Muhammad F. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China. 2010. Vol. 4. № 2. Pp. 280-301.

[3] Marchionini G. Exploratory search: From finding to understanding // Commun. ACM. - 2006. - Vol. 49, no. 4. - Pp. 41-46.

[4] White R. W., Roth R. A. Exploratory Search: Beyond the Query-Response Paradigm. Synthesis Lectures on Information Concepts, Retrieval, and Services. – Morgan and Claypool Publishers, 2009.

[5] Steyvers M., Griffiths T. Finding scientific topics // Proceedings of the National Academy of Sciences. — 2004. — Vol. 101, no. Suppl. 1. — Pp. 5228–5235.

[6] Paul M. J., Girju R. Topic modeling of research fields: An interdisciplinary perspective // RANLP. — RANLP 2009 Organising Committee / ACL, 2009. — Pp. 337–342.

[7] Blei D., Lafferty J. A correlated topic model of Science // Annals of Applied Statistics. — 2007. — Vol. 1. — Pp. 17–35.

[8] Bolelli L., Ertekin S., Giles C.L. Topic and trend detection in text collections using latent dirichlet allocation // ECIR. — Vol. 5478 of Lecture Notes in Computer Science. — Springer, 2009. — Pp. 776–780.

[9] Wang C., Blei D. M. Collaborative topic modeling for recommending scientific articles // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — New York, NY, USA: ACM, 2011. — Pp. 448–456.

[10] Thomas Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50-57, New York, NY, USA, 1999. ACM.

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993-1022, 2003.

[12] Vorontsov K. V. Additive regularization of topic models of text document corpora [Additivnaya regulyarizatsiya tematicheskikh modeley kollektsiy tekstovykh dokumentov] // RAN Reports [Doklady RAN]. - 2014. - T. 456, № 3. - S. 268-271.

[13] Klenin, J. Comparison of vector space representations of documents for the task of information retrieval of massive open online courses / J. Klenin, D. Botov, Y. Dmitrin // Communications in Computer and Information Science: Proceedings of the 6th Conference on Artificial Intelligence and Natural Language (St. Petersburg, Russia, September 2017). – 2018. – vol. 789. – pp. 156-164.

[14] Newman D., Karimi S., Cavedon L. External evaluation of topic models // Australasian Document Computing Symposium. — December 2009.— Pp. 11-18.

[15] Newman D., Lau J. II., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — HLT TO. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2010.— Pp. 100-108.

[16] Newman D., Noh Y., Talley E., Karimi S., Baldwin T. Evaluating topic models for digital libraries // Proceedings of the 10th annual Joint Conference on Digital libraries. — JCDL TO.— New York, NY, USA: ACM, 2010,- Pp. 215-224.

[17] Vorontsov K. V., Potapenko A. A. Regularization of probabilistic topic models for improvement of interpretability and topic count detection. [Regulyarizatsiya veroyatnostnykh tematicheskikh modeley dlya povysheniya interpretiruemosti i opredeleniya chisla tem] // International conference on computer linguistics Dialog-2014. [Mezhdunarodnaya konferentsiya po komp'yuternoy lingvistike Dialog-2014.