ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 174

Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)

The Comparison of Distributive Semantics Models Applied to the Task of Short Job Requirements Clustering for the Russian Labor Market

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

Ivan Ryazanov
Institute of Information Technology
Chelyabinsk State University
Chelyabinsk, Russia
ryazanov.ivan@icloud.com

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

***Abstract*—In this article we compare different vector models (tf-idf, word2vec, fasttext, lda, lsi, artm) in the short text clustering task, using a dataset of job vacancy descriptions in Russian. A two-step experiment is proposed to determine the best model and its hyperparameters based on the quality of the resulting short text clusters. In the first stage, we investigate how various hyperparameters of each model can affect the clusters, produced by training a K-means model on each of the vector representations. In particular, we consider in detail, how the size of the output vector representation in each of our models can influence the quality of the final clusters. We also provide an extensive analysis of the effects of various regularization options for clusters, learned using the vectors produced by the ARTM algorithm. During the second stage, the models showing the best results in the previous step (word2vec, fasttext) are analyzed in greater detail. We compare the effectiveness of these models against datasets of different sizes, as well as using different structures of the source fragments (partial elements or full texts of vacancy descriptions). In our experiments, the highest quality of clusters (evaluated using the ARI metric) was achieved by word2vec, closely followed by the fasttext model. Finally, we perform a topic analysis for each of the resulting clusters and evaluate their homogeneity.**

Keywords—clustering, vector models, short texts, job vacancies, labour market

I. Introduction

The modern labor market is changing extremely fast due to the rapid development of information technologies. New professions and technologies appear constantly, and as a result, requirements for employees on the labor market are changing. Automatic information extraction of these requirements becomes quite urgent for the applicant. This article serves as the continuaton of our previous research [1] into the Russian labour market. The article provides a comparative analysis of various models of distribution semantics (tf-idf, word2vec, fasttext, lda, lsi, artm) as applied to the problem of short texts clustering on the example of vacancies in the Russian labor market. Our goal is to find the most effective method for separating each vacancy text into various categories of statements, such as skill and knowledge requirements, previous job experience, working conditions, etc.

II. Overview of existing approaches to labor marked analysis

This section discusses related works on short text clustering and natural language analysis of the labor market domain.

In [2], the authors have demonstrated how topic modeling can be applied in the field of labor market intelligence. Presented architecture of clustering, labelled Mixed Graph of Terms has shown interesting results in the task of classification: it can extract a demonstrative graph of relations between words from corpus. This model, however, can be simplified to the LDA module, which we have previously [1] evaluated only for it to be outperformed by ARTM.

Partially, the task of understanding the structure of requirements of labor market is being solved by classification. For example, [3] distributional semantic models in combination with different multi-label and multi-class classifiers such as LabelPowerset logistic regression show promising results. Although professions as classes, described in legal documents exist, this solution requires improvements in case of understanding the results of classifying. High score of metrics in classification does not show the structure of data, in contrast the analysis of mistakes can give another point of view.

The analysis of short job vacancy parts is quite similar to the analysis of twits [4], with the only difference being a much smaller and specialized vocabulary in the case of job vacancies, which is restrained and less expressive, teeming with clericalism. Such texts rarely contain emojis or emotional words, while having an abundance of jargon, specialized terms, and words with shifted semantics, such as “experience.”.

The authors of [5] evaluate their soft-clustering model on three datasets: search snippets, stack overflow and pubMed articles. These datasets also contain specialized vocabularies. Expectedly, the performance is quite poor (60% accuracy for stack overflow dataset).

New approaches to analysis and short texts in general [6,8] and labor market texts in particular [9,13] are currently being actively developed.

III. TRAINING DATASET EVALUATION

The primary dataset in our research has been assembled from job listings taken from hh.ru. It contains a total of 33 million individual vacancies. Approximately 10% of them are IT vacancies (~ 3 million). Out of these we have selected about 430K unique vacancies from most recent years [2,3]. The text of each vacancy was then split into sentences, resulting in a dataset of 13 million short text elements (13M dataset, labeled as 13M_parts below). For most of the comparative experiments on short texts clustering, we randomly selected 100 thousand elements (100K dataset, labeled as 100K_parts).

As a result, 3 datasets for training vector models were formed (see table 1): one dataset from 430k full texts of vacancies and two datasets of short texts 100K and 13M, which include: requirements, duties, general working conditions, previous experience of the applicant, and various miscellaneous elements, which introduce semantic noise . The examples of these elements of the vacancies are presented in table 2.

TABLE I. DESCRIPTION OF DATASETS

Dataset	Number of documents	Number of tokens	Number of unique tokens (vocabulary)	Average number of words in a document
100K_parts	101K	607K	19K	6.02
13M_parts	13M	99M	194K	7.59
430K_fulltext	430K	62M	193K	146.15
valid_dataset	1200	14256	2655	6

Figure 1 shows that the size of the overwhelming majority of vacancy elements does not exceed 10-15 words.

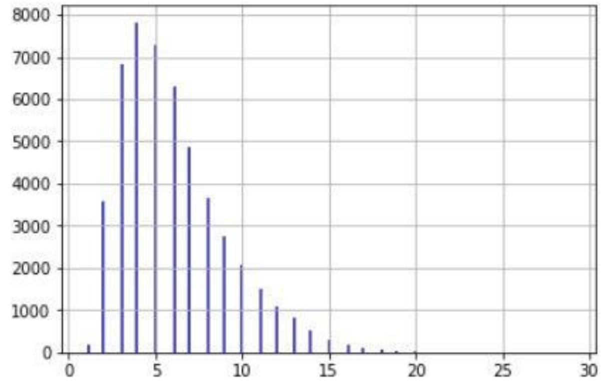

Fig. 1. The distribution of the word counts in the elements of the 100K dataset

IV. EVALUATION DATASET DESCRIPTION DUTIES

To assess our method's quality, we have manually labeled a set of 1000 elements from vacancies.Top level of labeling splits dataset into 6 labels (tab. 2).

Group 3 (duties) and group 4 (requirements for skills) share common vocabulary, for example, “website development using html, css, js technologies” (group 3) and “requirements: know web development technologies html, css, js” (group 4). Because of this, we have decided to separate these groups into second-level clusters. Thematic groups are presented in table 3. The examples from this dataset are presented in table 4.

TABLE II. FIRST LEVEL SUBJECT GROUPS

#	Name	Description
1	General experience	Experience in some position, development of something not related to technology
2	Requirements to soft-skills	General skills, negotiation, competent speech, English, teamwork, etc.
3	Duties	Actions to be taken as part of work duties
4	Requirements to hard-skills	Professional experience: skill, ability, understanding, ability to work with something, knowledge, acquaintance, experience with technology
5	Education (documents)	Availability of documents on education (certificate, diploma, certificate)
6	Other	Various common words, trash

TABLE III. SECOND LEVEL SUBJECT GROUPS

#	Names groups
1	Database
2	1C (SAP)
3	Network
4	Software Development, Programming Languages
5	Testing, quality assessment
6	WEB, CMS
7	Mobile development
8	Hardware, controllers, microcontrollers
9	MS Office, a confident user
10	Graphics 2D / 3D, VR
11	Data Recovery and Protection
12	Advertising, SEO, Digital Marketing
13	Security systems
14	ProjectMeneger, Scrum, Agile
15	Financier, accountant
16	Version Control Systems
17	OS Administration (Linux, Windows)
18	Sales
19	Integration
20	Documentation, reporting
21	Data Analysis
22	Business analyst, business processes

TABLE IV. EXAMPLES OF DATA FROM MANUALLY LABELED DATASET

id	Text	G1	G2
6403153	ООП, STL, boost, классические алгоритмы и структуры, шаблоны проектирования	4	4
7577237	Создание дизайна сайтов, баннеров, презентаций	3	6
2688504	Навыки эффективных коммуникаций, ответственность, стрессоустойчивость, инициативность	2	
5566825	Что будет плюсом:	6	
5289206	Знание основ HTML, CSS, Javascript, PHP	4	6
3841654	Высшее образование	5	

V. Description of models

We have used the following vector models in our experiments: tf-idf [14], word2vec [15], fasttext [16], lda [17], lsi[18], artm [19,20]. (table 5).

TABLE V. VECTOR MODEL FAMILIES

Model \ dataset	Size vector	100K parts	13M parts	FullText 430K
Stage 1				
TFIDF	10, 50, 100, 200, 300, 500	+		
WORD2VEC	50, 100, 200, 300, 500	+		
FastText	50, 100, 200, 300, 500	+		
LDA	10, 50, 100, 200	+		
LSI	10, 50, 100, 200	+		
ARTM +regularization	10, 20, 50, 100, 200, 500	+		
Stage 2				
WORD2VEC	50, 100, 200, 300, 500	+	+ (sv 300)	+ (sv 300)
FastText	50, 100, 200, 300, 500	+	+ (sv 300)	+ (sv 300)

* information of datasets is presented in table 1

* the description of stage 1 and 2 is presented in the section “experiment description”

We have used the following naming convention for our models:

ALGORITHM_SIZE-VECTOR_[REGULARIZATION PARAM.]_DATASET.

Using various regularization parameters present in the ARTM library [21], we have constructed a total of 1944 variations of this topic model. This was done in order to assess the impact of regularization on the quality of a topic model, when it was used to facilitate short text clustering. The main regularizers are [22]:

- SmoothSparsePhiRegularizer provides the ability to smooth or discharge subsets of topics (values used: 0.0, -0.1, -0,25 -0.5, -1, - 1.25, -1.5, -1.75, -2.0)
- SmoothSparseThetaRegularizer provides the ability to smooth or parse subsets of topics in multiple documents (values used: 0.0, -0.1, -0,25 -0.5, -1, - 1.25, -1.5, -1.75, -2.0)
- Decorrelator provides the ability to decorrelate topics (for example, to make topics more distinct), which allows to increase the interpretability of topics (values used: 0, 1000, 3000)

VI. Structure of the experiment

Stage 1. At the first stage, all possible variations of vector models and their hyper parameters (see table 5), trained on a 100K dataset, were obtained. For each model, clustering quality estimates were calculated. In total, about ~ 2K models were constructed for all variations of the algorithms and their parameters. Two models with the best estimates passed to the next stage (the results are shown in Figures 2-3).

For each vector model we have built its own family of models with different sizes of output vectors. All models were further trained and clustered under equal conditions on a 100K dataset. It was done in order to further analyze and evaluate the impact of different datasets on the final clusterization assessment. The models that showed the best result on 100K were further trained and compared on the 13M dataset and 430K_fulltexts (stage 2).

Clustering was performed on a different number of clusters (5, 7, 10, 15, 20, 30, and 50) using the clustering method: K-MEANS for each vector model. The choice of this clustering method was determined by the analysis of the following articles [23,25]

The quality of clustering assessment. To evaluate resulting clusters, we utilized ARI [26]: a popular metric for quality assessment in clustering tasks. Since in step 3 clustering was performed with different number of clusters (namely 5, 7, 10, 15, 20, 30, and 50 clusters) for each vector model, it was decided to evaluate the total quality of the vector model as the average of all clustering estimates for this model. The assessment of the quality of clustering was carried out without taking into account group 6 (see table 1).

Stage 2. For two models of the leaders of stage 1, additional models were trained on 13M and 430K_fulltext datasets. Comparative assessments were made of the impact of the dataset size on the final grade (the results are shown in Figure 4). One model with the best ratings, passed to the next stage.

The results of the leader model were evaluated in detail, which at the previous stages received the highest estimates of the quality of clustering (the result is shown in Fig.5).

VII. Results

A. *ARTM and regularization*

Figure 2a shows that the best size of the topic vector (size_vector) for ARTM models is 20. The Figures 2c and 2d show that moderate regularization (SmoothSparsePhiRegularizer and SmoothSparseTheteRegularizer regularizers, respectively) can lead to a slight improvements, but in Figure 2b (regularizer Decorrelator) for any use of regularization leads to a deterioration of the final estimates.

Based on the totality of indicators for further comparative experiments, a topic model was selected with the following parameters: topic size 20, SmoothSparsePhiRegularizer (parameter tau) = -0.5, SmoothSparseThetaRegularizer (parameter tau) = -0.5, Decorrelator = 0.

B. *The comparison of vector models*

The next series of experiments was devoted to comparing the effects of the different types of and values of their main hyperparameter - the size of the output vector on the quality of the final value (average ARI).

Figure 3 shows that the best results are achieved by word2vec_50 and fasttext_50. An interesting effect can be observed in almost all models: an increase in the vector size (starting with 50) does not lead to a significant improvement in the quality of clusters. In case of word2vec (Fig. 3.b) it leads to a smooth decrease (deterioration) of the final estimate.

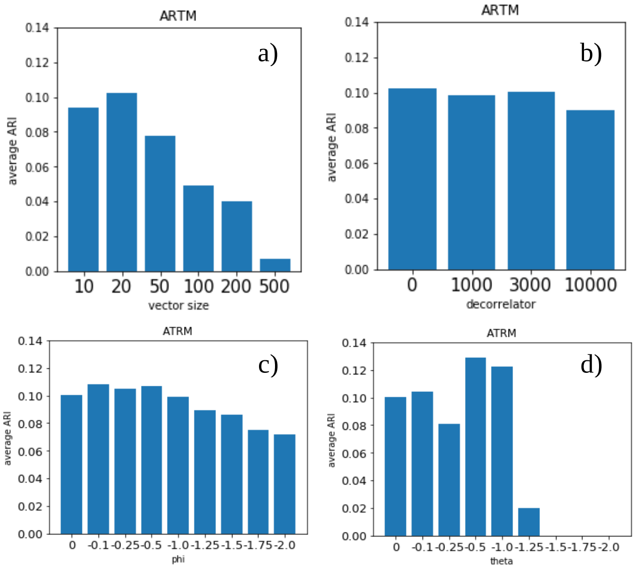

Fig. 2. Average ARI score for different sizes of the topic vector (a), and for different values of regularization parameters (b, c, d).

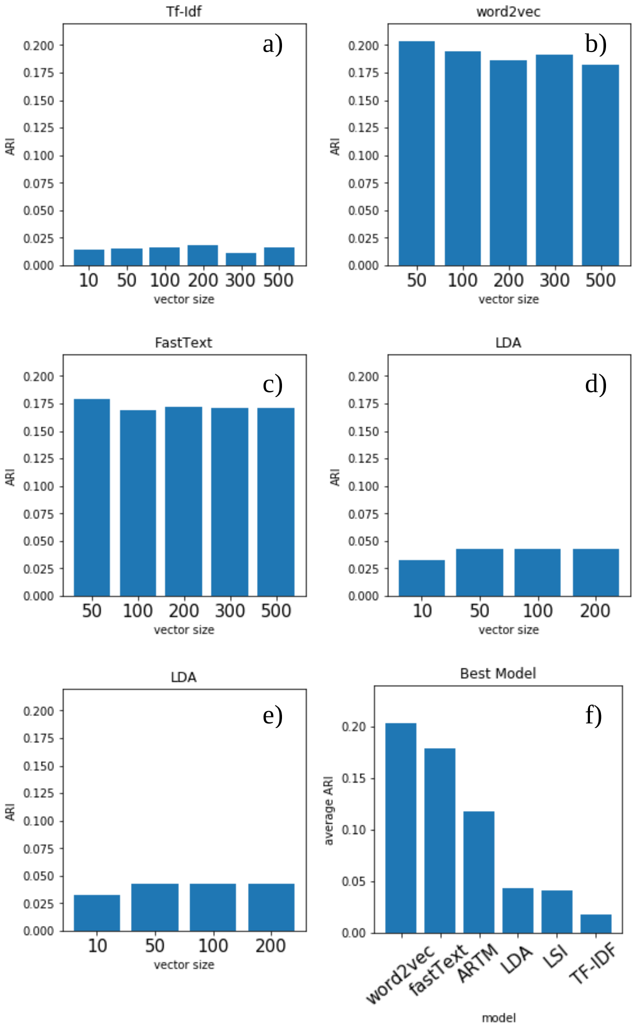

Fig. 3. Average ARI rating for different sizes of the output vector

Based on their performance, word2vec and fasttext models were chosen for further experiments, designed to assess the possible influence of the vector size on the quality of produced clusters. As a result, for the final assessment of clustering, word2vec and fasttext models were selected.

The graph (Fig. 4) illustrates that overall word2vec_50 is outperforms fasttext_50 for all sizes of the output vector. Also for fasttext increasing the output vector size has almost no effect on the quality of clustering. However, increasing the vector size for word2vec leads to a smooth decrease in the quality of clustering (within 1-2%). It can be concluded that for the task of clustering short texts it is not necessary to choose a large output vector size (this conclusion is also relevant for other types of vector models, as follows from Figure 3).

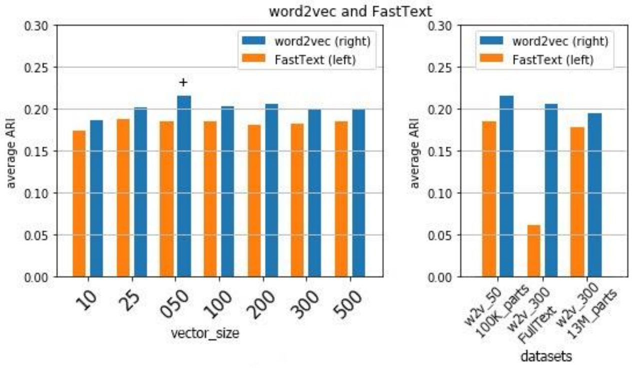

Fig. 4. Comparison of average ARI for different vector sizes for word2vec and fasttext models

C. Determining the number of clusters

For the best model (word2vec_50_100K), the clustering was evaluated for 5, 7, 10, 15, 20, 30, 50, 75, 100 clusters. The results are presented in Fig. 5. The highest quality is achieved with 30 classes. This number of clusters was taken as the basis for further actions.

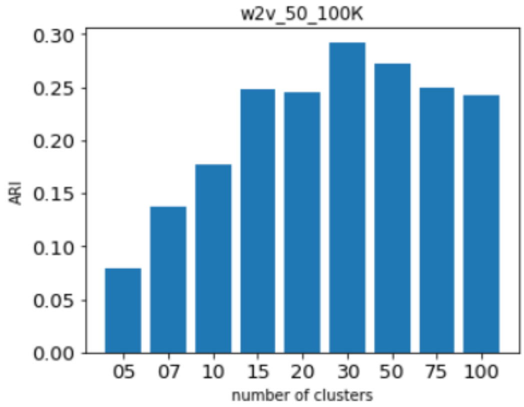

Fig. 5. Assessment (ARI) of word2vec_50_100K clustering quality for different numbers of clusters

D. Defining Clustering Features

For the final experiment, the word2vec_50_100K model was used to create 30 clusters. Fig. 6 provides the distribution of clusters in groups that are presented in table 2. We can distinguish groups 1, 2, 3, 5, 8, 9, 12, 14, 16, 17, 18, 19, 20, which are mostly covered by 1-2 clusters, which may indicate that clustering based on the selected model learns features of short texts in these groups rather well.

Special attention should be paid to groups 4 and 6, for which the original, manually labeled dataset was significantly unbalanced, making it much more difficult for the clustering algorithm to discern unique features in these groups.

Another metric of quality for clustering is the homogeneity of the resulting clusters. Clusters, predominantely distributed between 1 or 2 classes, are considered to be highly homogenous. Figure 6 shows the distribution of clusters #01-10 for 22 groups from table 2.. Total number of clusters is 30, the first 9 clusters were selected for demonstration. Clusters 00 and 07 did not contain a single record and therefore were not included in the graph. Figure 6 shows that the majority, specifically 7 clusters: 01, 03, 04, 05, 06, 08, 09, are distributed over one dominant group, which may indicate that the clustering algorithm manages to create fairly homogeneous groups of short texts that coincide with separate manually labeled classes (see table 3).

Table 6 presents short texts taken from clusters #01-10.

Analyzing the results in table 3 in detail, it can be noted that the short texts selected by the clustering algorithm in clusters 1, 3, 4, 5, 8 are very close both in terms of vocabulary and in meaning. They form fairly homogeneous clusters, but in cluster # 2 examples of short texts are quite different, both in terms of vocabulary and content, which makes it non-homogenous (this conclusion is fully supported by the graphs in Figure 6).

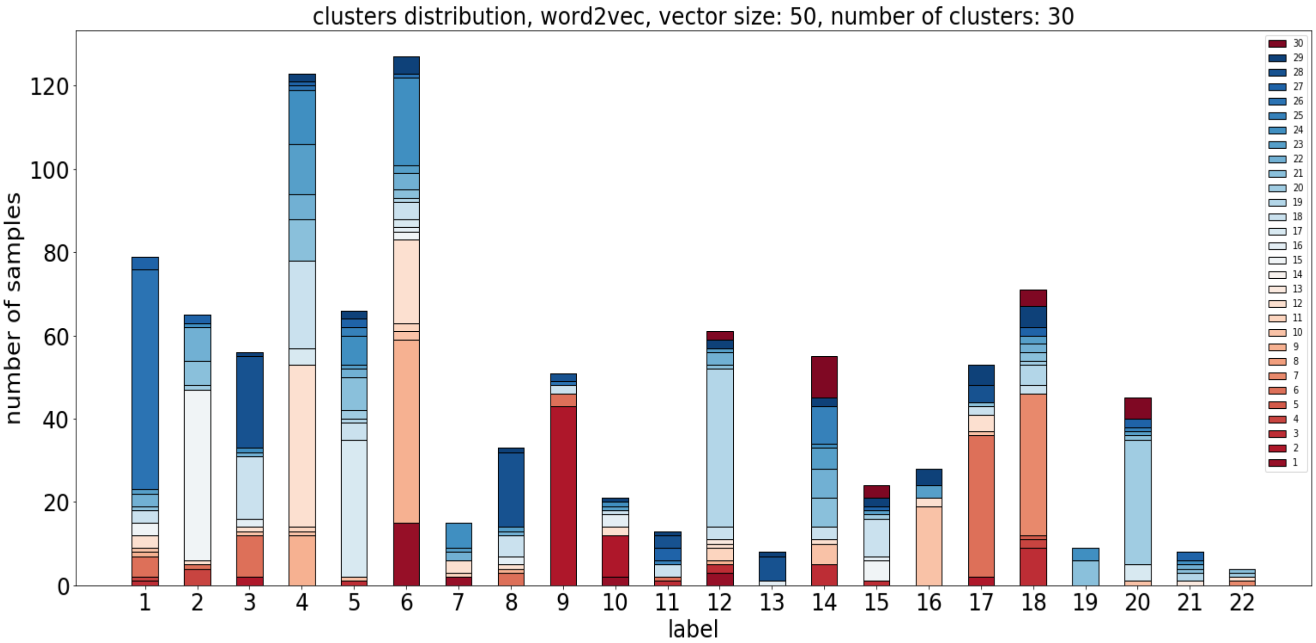

Fig. 6. The distribution of clusters in groups from table 2

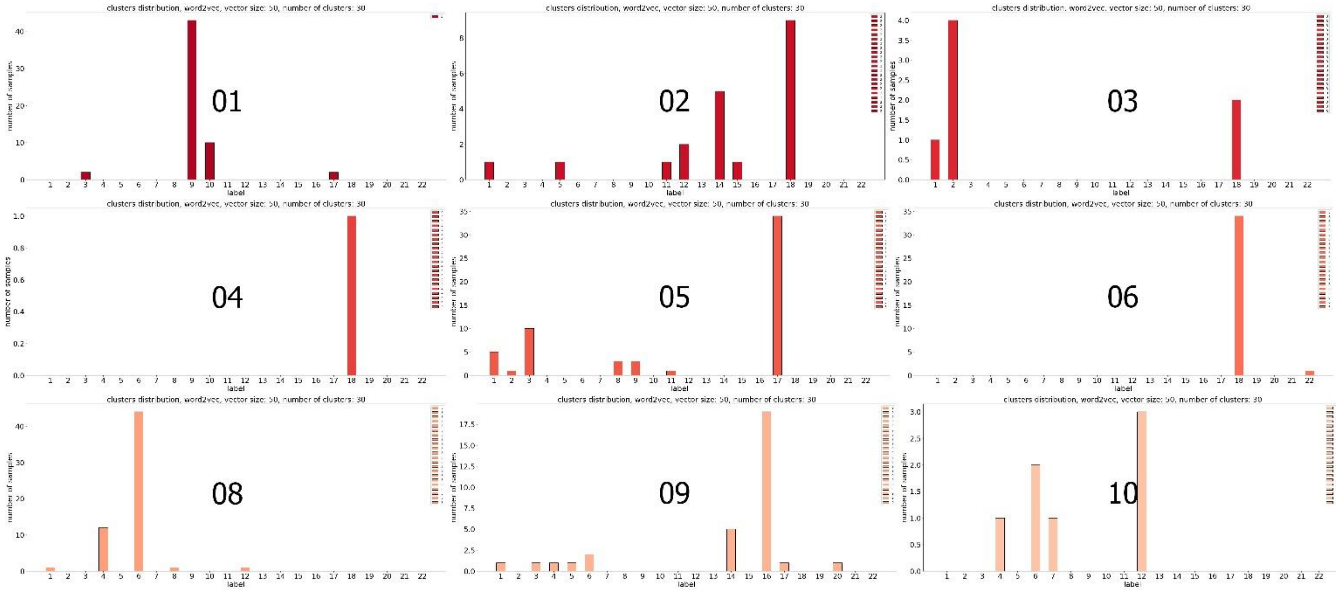

Fig. 7. The distribution of select clusters (No. 01-10) by classes from table 2.
(* clusters 00 and 07 did not contain a single record)

TABLE VI. EXAMPLES OF SHORT TEXTS FROM DIFFERENT CLUSTERS

Examples of short texts	**Cluster #**
MS Windows на экспертном уровне	1
Уверенное владение компьютером и пакетом MS Office (Word, Excel, Visio)	1
Уверенный пользователь ПК,	1
Профессиональное знание ПК и офисных приложений MS Office;	1
Продвинутый пользователь MS Excel.	1
Знание Microsoft Word, Excel	1
Участие в работе по постановке системы процессного управления, создании процессного офиса.	2
Выполнение инжиниринга для нужд программных комплексов объектов энергетики	2
Поддержка работоспособности предоставляемых компанией услуг	2
Перевозка персонала ЭТЦ для выполнения работ на сетях связи.	2
Разработка и внедрение трейд-маркетинговых программ в координации с отделом маркетинга;	2
Работы по сопровождению и сервисному обслуживанию банкоматов в банках;	2
Предоставление консультации по внесению платы Владельцам Транспортных Средств (далее ВТС)	3
Сопровождение пользователей в период опытно-промышленной эксплуатации системы;	3
Опыт общения с пользователями;	3
консультировать по техническим вопросам высокого уровня сложности;	3
5) Консультации пользователей	3
Грамотная устная и письменная речь;	4
Сильные навыки командной работы и эффективного взаимодействия.	4
Отличные коммуникативные навыки	4
Будешь применять на практике навык эффективных телефонных переговоров.	4
Навыки ведения переговоров;	4
- Сильные аналитические и лидерские способности, отличные коммуникативные навыки;	4
Знание серверных ОС Windows Server 2008/2012.	5
Опыт работы с продуктами MS Office от 2010.	5
опыт построения локальной сети с использованием доменной иерархии (Windows и Samba);	5
знание принципов работы и опыт настройки ПК и периферийных устройств;	5
Знание принципов работы СУБД, резервирования, резервного копирования и восстановления.	5
3+ лет Python/Django с "боевым" опытом.	8
Знание основ HTML, CSS, Javascript, PHP	8
Знание HTML5, CSS3, JavaScript, Ajax, jQuery, SASS;	8

VIII. Conclusion

The article provides the comparative analysis of vector models applied to the task of clustering short texts in the field of Russian labor market. A comparative assessment of the influence of the size of vector representations on the quality of clustering is shown. We conclude that the choice of vector size value in the region of 50 is optimal and preferable. It is also shown that increasing the size of the output vector for most models does not improve the quality of clustering. For the ARTM model moderate regularization results positively on the quality of clustering. Additionally, we have proved that the size of the dataset does not significantly improve the clustering quality.

In the final series of experiments with the best model (word2vec_50_100K, clustered by k_means 30), we have shown that a large number of clusters turn out to be fairly homogeneous, and this model copes well with its.

Acknowledgment

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).

References

[1] Vinel, Mikhail, et al. "Experimental Comparison of Unsupervised Approaches in the Task of Separating Specializations Within Professions in Job Vacancies." Conference on Artificial Intelligence and Natural Language. Springer, Cham, 2019.

[2] Colace, F., De Santo, M., Lombardi, M., Mercorio, F., Mezzanzanica, M., & Pascale, F. (2019, January). Towards labour market intelligence through topic modelling. In Proceedings of the 52nd Hawaii International Conference on System Sciences.

[3] Botov, D., Klenin, J., Melnikov, A., Dmitrin, Y., Nikolaev, I., & Vinel, M. (2019, June). Mining Labor Market Requirements Using Distributional Semantic Models and Deep Learning. In International Conference on Business Information Systems (pp. 177-190). Springer, Cham.

[4] Chaturvedi, V., Pramanik, A., Ghosh, S., Bhadury, P., & Mondal, A. (2020). A Supervised Approach to Analyse and Simplify Micro-texts. In Emerging Technology in Modelling and Graphics (pp. 61-67). Springer, Singapore.

[5] Hadifar, Amir, et al. "A self-training approach for short text clustering." Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). 2019.

[6] Banerjee, Somnath, Krishnan Ramanathan, and Ajay Gupta. "Clustering short texts using wikipedia." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007.

[7] Hu, Xia, et al. "Exploiting internal and external semantics for the clustering of short texts using world knowledge." Proceedings of the 18th ACM conference on Information and knowledge management. 2009.

[8] Sriram, Bharath, et al. "Short text classification in twitter to improve information filtering." Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010.

[9] Boselli, Roberto, et al. ”Using machine learning for labour market intelligence.” Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2017.

[10] Colombo, Emilio, Fabio Mercorio, and Mario Mezzanzanica. ”Applying machine learning tools on web vacancies for labour market and skill analysis.” (2018).

[11] Wowczko, Izabela. ”Skills and vacancy analysis with data mining techniques.” In-formatics. Vol. 2. No. 4. Multidisciplinary Digital Publishing Institute, 2015.

[12] Spirin, Nikita, and Karrie Karahalios. ”Unsupervised approach to generate informative structured snippets for job search engines.” Proceedings of the 22nd International Conference on World Wide Web. ACM, 2013.

[13] Muthyala, Rohit, et al. ”Data-driven Job Search Engine Using Skills and Company Attribute Filters.” 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2017.

[14] Ramos, Juan. "Using tf-idf to determine word relevance in document queries." Proceedings of the first instructional conference on machine learning. Vol. 242. 2003.

[15] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

[16] Joulin, Armand, et al. "Fasttext. zip: Compressing text classification models." arXiv preprint arXiv:1612.03651 (2016).

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

[18] 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.

[19] 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.

[20] Vorontsov, Konstantin, et al. "Bigartm: Open source library for regularized multimodal topic modeling of large collections." International Conference on Analysis of Images, Social Networks and Texts. Springer, Cham, 2015.

[21] Vorontsov, Konstantin, and Anna Potapenko. "Additive regularization of topic models." Machine Learning 101.1-3 (2015): 303-323.

[22] Vorontsov, Konstantin, Anna Potapenko, and Alexander Plavin. "Additive regularization of topic models for topic selection and sparse factorization." International Symposium on Statistical Learning and Data Sciences. Springer, Cham, 2015.

[23] Deokar, Sanjivani Tushar. ”Text documents clustering using k means algorithm.” International Journal of Technology and Engineering Science [IJTES] 1.4 (2013): 282-286.

[24] Zhu, Yan, Jian Yu, and Caiyan Jia. ”Initializing k-means clustering using affinity propagation.” 2009 Ninth International Conference on Hybrid Intelligent Systems. Vol. 1. IEEE, 2009.

[25] Guan, Renchu, et al. ”Text clustering with seeds affinity propagation.” IEEE Trans-actions on Knowledge and Data Engineering 23.4 (2011): 627-637.

[26] Steinley, Douglas. "Properties of the Hubert-Arable Adjusted Rand Index." Psychological methods 9.3 (2004): 386