BIG DATA TECHNOLOGIES FOR LABOUR MARKET ANALYSIS

S.D. Belov[1,2 a], J.N. Javadzade[1,2 b], I.S. Kadochnikov[1,2 c], V.V. Korenkov[1,2 d], P.V. Zrelov[1,2 e]

[1] *Joint Institute for Nuclear Research, 6 Joliot-Curie St, Dubna, Moscow Region, 141980, Russia*

[2] *Plekhanov Russian University of Economics, 36 Stremyanny per, Moscow, 117997, Russia*

E-mail: [a] belov@jinr.ru, [b] jjavadzade@yandex.ru [c] kadivas@jinr.ru, [d] korenkov@jinr.ru,
[e] zrelov@jinr.ru

This paper discusses some approaches to the intellectual text analysis in application to automated monitoring of the labour market. The scheme of construction of an analytical system based on Big Data technologies for the labour market is proposed. Were compared the combinations of methods of extracting semantic information about objects and connections between them (for example, from job advertisements) from specialized texts. A system for monitoring of the Russian labour market has been created, and the work is underway to include other countries in the analysis. The considered approaches and methods can be widely used to extract knowledge from large amounts of texts.

Keywords: text analysis, Big Data, labour market monitoring

1. Previous work

Recently, the prospects of "digitalization" of economic processes have been actively discussed. This is an extremely difficult task that has no solution in the framework of traditional methods. The prospects for their qualitative development in the article are illustrated by the example of using Big Data analytics and text mining to assess the labor force needs of regional labor markets. There is also an important question of studying the interaction between labour market and professional education system [1]. The problem was solved using the automated information system developed by the authors for monitoring the compliance of employers' personnel needs with the level of specialist training. The information base for collecting information was open sources. The presented system creates additional opportunities for identifying qualitative and quantitative relations between the education sector and the labor market. It is aimed at a wide range of users: authorities and administrations of regions and municipalities; management of universities, companies, recruitment agencies; graduates and university graduates.

In previous work [2] we described the approaches and the prototype of the labour market monitoring system. Now, having enough real data from the market, it is possible to make more elaborated analysis, allowing a more detailed understanding of market requirements.

2. Data processing infrastructure

Every day, about 1.5 million active vacancies are updated and subject to analysis and preservation. In order to track the dynamics of indicators and lay the basis for forecasting the state and needs of the labor market, we need effective storage, intellectual analysis, and visualization of data on vacancies for the maximum available time (we are considering data from 2015, which at the moment is already 5 years). Therefore, the system was based on Big Data technologies. First of all, the following freely distributed software products were used: Spark, Hadoop, Kafka, Flume, Marathon, Chronos, Docker. Data processing schema is presented on Fig. 1.

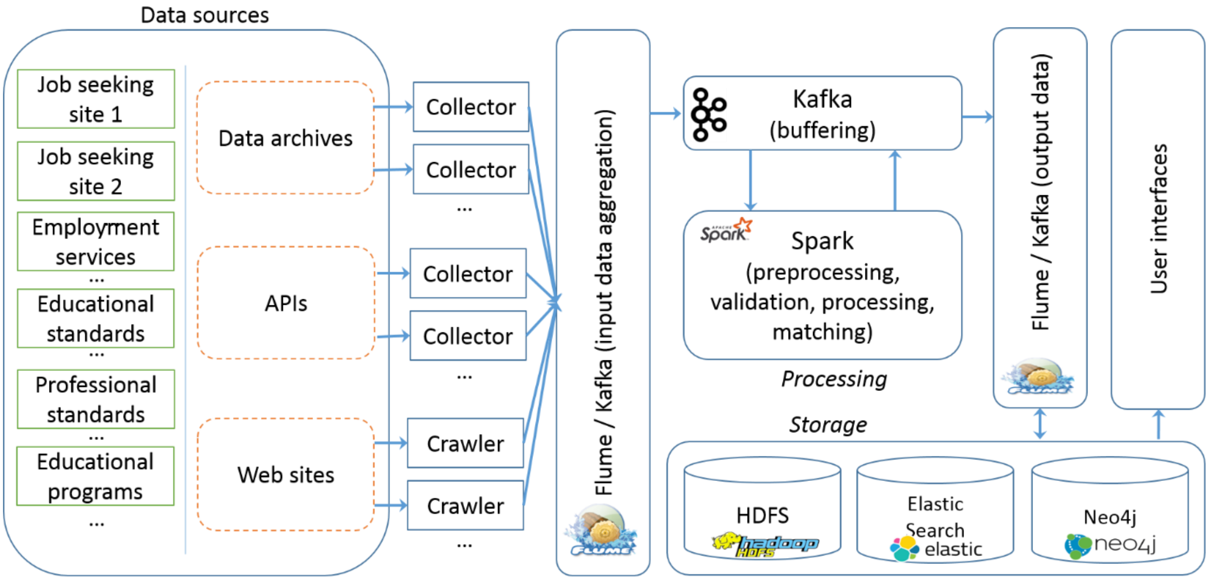

Figure 1. Infrastructure and data flow scheme

There are three big main data sources (job-seeking portals): trudvsem.ru, hh.ru, sj.ru. Information on job offers is publicly available here, and custom collectors were developed to fetch data from there.

To support the processing, lambda-architecture was implemented. The analysis should be as fast as possible, so in-memory processing is the main technology used. The key part of the infrastructure is an Apache Spark [3] cluster.

3. Job offers analysis and classification

Basic information about the state of the labor market is obtained by analyzing the database of collected vacancies. To obtain correct statistics, it is necessary to solve, first of all, the following tasks:

1) Determination of duplicate vacancies. Even if you use one source, job ads can be duplicated, but if you use multiple sources, such checks are necessary.

2) Classification of vacancies by branches of professional activity.

3) Analysis of the job offer content, analysis of individual requirements for skills and competencies.

The need to delete identical vacancies is connected with the fact that the sample we use consists of uploading data from several sources, and on each of the sources the same vacancy can be republished repeatedly with some time interval. In order that the data of the same vacancy were not processed several times, it was decided to implement search of identical and similar vacancies with further removal of duplicates. Despite a direct comparison under the name of employer, job title and address, it is necessary to take into account the fact that the name of the position and the content of jobs may change if re-published or the information could be just written on a slightly different way.

Previously, to compare the meaning of text fields, the method of comparing the vector representation of texts in semantic space was used (using the *word2vec* approach). Further, to make the analysis more specific, it was necessary to distinguish words and expressions characteristic of certain professions and fields of activity. For this purpose, the statistical indicator TF-IDF [4] (term frequency - inverted document frequency) was used, which is mainly used to assess the importance (weight) of a particular word (term) in the context of the entire document included in the general collection (base).

Due to the data from *hh.ru* and *superjob.ru* are already structured, it can be used as training data for a kind of multi-label classification [5]. That is, initially there is a sample with about one million marked data and it is possible to operate on it. The next step is to extract only the data necessary for classification. These are the duties, requirements, as they contain basic information about the job and a list of professional areas and specializations to which the job belongs. After the preprocessing: removal of stop words, tokenization and lemmatization of the text, there is everything necessary for further classification of the vacancy. Job offers than were classified against professional areas and required competencies. For the classification, it was trained and used a neural network implementation from the *scikitlearn* library. When jobs are classified, it is, moreover, easier to find identical records in the database.

3. Conclusions

A system for monitoring of the Russian labour market has been created, and the work is underway to include other countries in the analysis. The considered approaches and methods can be widely used to extract knowledge from large amounts of texts (it works fine on text data of terabyte-scale volume). Using together Big Data technologies, statistical methods and machine learning techniques allowed us to significantly accelerate the analysis and conduct it in a reasonable time for researchers.

4. Acknowledgment

The work was supported by the Russian Foundation for Basic Research (RFBR), grant 18-07-01359 "Development of an information-analytical system of monitoring and analysis of labour market's needs for graduates of Universities on the basis of Big Data analytics".

References

[1] Azmuk, N. (2015). The interaction of labour markets and higher education in the context of digital technology. Economic Annals-XXI, 7–8(1), 98–101.

[2] S. Belov, I. Filozova, I. Kadochnikov, V. Korenkov, R. Semenov, P. Smelov, P. Zrelov Labour market monitoring system, *CEUR Workshop Proceedings*, ISSN:1613-0073, Vol. 2267, pp. 528-532

[3] Armbrust, M., Xin, R.S., Lian, C., Huai, Y., Liu, D., Bradley, J.K., Meng, X., Kaftan, T., Franklin, M.J., Ghodsi. A., & Zaharia, M. 2015. Spark SQL: Relational Data Processing in Spark. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (SIGMOD '15). ACM, New York, NY, USA, 1383-1394. DOI: https://doi.org/10.1145/2723372.2742797

[4] Mark Needham. scikit-learn: TF/IDF and cosine similarity for computer science papers // Available at: https://markhneedham.com/blog/2016/07/27/scitkit-learn-tfidf-and-cosine-similarity-for-computer-science-papers/ / (accessed 01.12.2019)

[5] Rocco Schulz. Performing Multi-label Text Classification with Keras // Available at: https://blog.mimacom.com/text-classification/ (accessed 01.12.2019)