DATA GATHERING AND ANALYSIS FOR THE MONITORING OF THE RUSSIAN LABOUR MARKET

J. Javadzade [1,2,a], S. Belov [1,3]

[1] *Laboratory of Information Technologies, Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region, 141980, Russia*

[2] *University Centre, Joint Institute for Nuclear Research, 6 Joliot-Curie, Dubna, Moscow region, 141980, Russia*

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

E-mail: [a] jjavadzade@yandex.ru

This project is devoted to monitor and analyze the labour market using the publicly available data on job offers, CVs and companies gathered from open data sources and recruitment agencies. The relevance of the project is that some current work areas have already overcrowded, some are outdated, or may have a little need for the new employees, and some new and growing industries are proposing good job offers. The result obtained at the project allows one to have a look on the labor market at different levels for each region of Russia. This information is useful not only for school graduates, students and people who is just looking for a better job for themselves, but also for the employers and bodies of authority. It is also can be useful for universities to estimate the relevance of the educational programs they offer.

Keywords: labour market, education, Big Data, machine learning, Apache Spark

1. Motivation

Like for many countries, one of the most important issues in Russia is the problem of employment.

According to the official state statistic, in January 2018 the unemployment rate in Russia was about 5.2% [1], and about 22 million people live below the subsistence level line [2]. This could say about some problems in job market in Russia, and to overcome them the labour market should be carefully studied.

Labour market is a big and complex system, interconnected with many other systems of the society and state. The demand for professions changes every year, vacancies and industries which yesterday have been relevant, already today may not be in irrelevant. And this is also with professional skills.

This project's aim is to provide country-level monitoring and analysis of the labour market basing on the publicly available data on job offers, CVs and companies gathered from open data projects and recruitment agencies. Certainly, this topic involves educational institutions, state enterprises, employers, households, citizens, etc. Universities teach students in obsolete areas that are not demanded by the labor market. If employers offer unattractive conditions and people leave the region, in this way in some regions there are no specialists in some professional areas etc. Results of this project will be useful for all the parties mentioned.

2. Labour market data gathering and analysis

The process of data gathering and processing consists of three stages. Technologies used on each level are shown on the Figure 1.

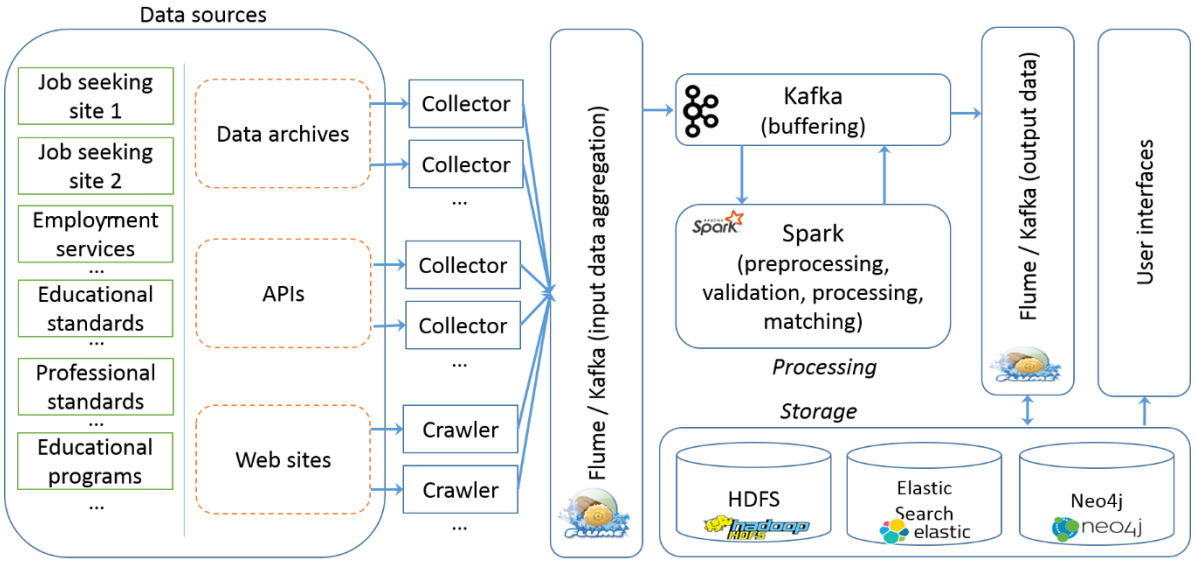

Figure 1. Data processing workflow

The first step is data gathering and joining from several sources. One of the key tasks is the collection of job offers data from open sources and recruitment agencies, because we will analyze this data to get the result of the study. While preparing the data gathering, a dozen of open sources was considered. To fully evaluate the labour market in each region of the country, it was decided to use data from largest open sources: HeadHunter.ru, SuperJob.ru, TrudVsem.ru (in the total amount it is about 1 500 000 vacancies active vacancies daily). Each of sources have API, that proved to be very useful for parsing all vacancies with necessary data like a requirement, responsibility, professional area salary, city and others. To collect data, we use Python with «request» module and with connecting to the API of sites.

The second step is data preprocessing. After first step follows no less important task — phase of data pre-processing: structuring, search and removal duplicate job offers from different sources etc. To find duplicate records, the text categorization approach is used [3].

At the third step the data analysis is been made. The core part here is matching the job offers with the official professions reference. This analysis is performed in a vector space: basing on the vectors for single words and phrases, the vectors for job offers and the descriptions of the professions are constructed using the specially tuned algorithm. Then vacancies and professions are being matched by finding the closest vector if any. For example, we use TF-IDF for identify most important requirement for each job market area [4] and sqrtcos similarity between texts [5]. Also, this step involves using of machine learning algorithms the vector representation is constructed using gensim word2vec [6], then the closest ones are selected basing on cosine distance [7]. For Russian language, RusVectōrēs models [8] are used.

Data pre-processing and analysis are very resource-intensive and involves huge amount of data, so it was decided to use the cluster system Apache Spark. All this process based on a cluster system Apache Spark [9, 10]. In-memory computations on a cluster for the considered task gives almost linear scalability against the number of CPU cores involved.

3. Acknowledgement

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

4. Conclusion

The result obtained at the end will show a detailed situation of job market by each region, it will allow one to have a look on the labor market on different levels. There will be a great opportunity to check which vacancies are relevant in one or another area, identify average wage, what should to know for employment for each professional area, knowledge of which specialists should be trained in higher education institutions etc.

This information is useful not only for school graduates, students and people who is just looking for a better job for themselves, but also for the employers. It is also can be useful for universities to estimate the relevance of the educational programs they offer. And surely it will have an excellent effect on the level of unemployment in the country for better development.

References

[1] Federal Service of State Statistics, http://www.gks.ru/bgd/free/B04_03/IssWWW.exe/Stg/d03/36.htm (in Russian)

[2] Rossiyskaya Gazeta, https://rg.ru/2017/12/14/rosstat-nazval-chislo-bednyh-rossiian.html (in Russian)

[3] Fabrizio Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys (CSUR), v.34 n.1, p.1-47, March 2002

[4] Yilin Niu, Chao Qiao, Hang Li, Minlie Huang, Word Embedding based Edit Distance, arXiv:1810.10752

[5] S. Sohangir, D. Wang, "Improved Sqrt-cosine Similarity Measurement", Journal of Big Data, pp. 4-25, 2017.

[6] Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. CoRR, abs/1301.3781.

[7] Eva Martínez Garcia, Cristina España-Bonet, Lluís Màrquez (May 2015). "Document-Level Machine Translation with Word Vector Models". Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT), pages 59-66.

[8] Kutuzov A., Kuzmenko E. (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham

[9] M. Zaharia et al., Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. NSDI 2012. April 2012.

[10] M. Armbrust et al., Spark SQL: Relational Data Processing in Spark. SIGMOD 2015. June 2015.