Contribution ID: **334** Type: **Poster presentations**

Data gathering and wrangling for the monitoring of the Russian labour market

Monday, 10 September 2018 16:00 (1 hour)

This project is devoted to monitoring and analyzing the labour market based on the publicly available data on job offers, CVs and companies gathered from open data projects and recruitment agencies.
The relevance of project is that some work areas have already overcrowded, some are outdated or some may have a little need for new workers, and some new and growing industries will offer good jobs. The result obtained at the end will allow one to have a look on the labor market on different levels starting from the local one. 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.
One of the key tasks is the collection of job offers data from open sources and recruitment agencies. Before writing parsing-scripts, need to analyze existing open sources of vacancies and identify the final list from which the vacancy data will be downloaded.
No less important task is data pre-processing, where the main task is to remove duplicate job offers appear from different sources. Because sophisticated comparison of more than a million vacancies requires significant time, this step was realized using Apache Spark on a cluster. Also, this step involves using of machine learning algorithms. For the job offers, the vector representation is constructed using gensim word2vec, then the closest ones are selected. For the moment, more than a million of vacancies from Headhunter, Superjob, Trudvsem recruitment agencies have been already collected and processed.

Primary author: Mr JAVADZADE, Javad (JINR)

Presenter: Mr JAVADZADE, Javad (JINR)

Session Classification: Poster Session