Connecting Professional Skill Demand with Supply

Erik Novak
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova cesta 39
1000 Ljubljana
erik.novak@ijs.si

Inna Novalija
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana
inna.novalija@ijs.si

ABSTRACT

Todays job market demand from the job seekers to continuously learn new skills. When applying for a job position one must have the required skill set. If the applicant is missing a skill it can be learned by attending a course. Finding the appropriate courses can be tedious but necessary work to be up-to-date with the job market demand. In this paper, we present a dashboard which connects the job market skill demand with the courses that give the required skill knowledge. We developed a pipeline for continuous crawling of job postings and courses which feeds the dashboard with the appropriate data. The dashboard allows searching by keywords and returns relevant job postings, courses and basic statistics relevant to the given search query.

General Terms

Job Market, Skill Set, Courses, Lectures, Design

Keywords

Information Retrieval, Data Mining, Analysis, Wikifier, VideoLectures.NET

1. INTRODUCTION

In todays job market the required skills are constantly evolving. This can be seen in more technical fields such as web development and data science where new tools and libraries are developed and available to the public with an increasing rate. This is visible in both research and industry sectors where a job position might require a previously unseen skill and the applicant needs to learn it to be qualified. Finding the courses that would give the skill knowledge can be tedious and does not guarantee its sufficiency.

To this end, we developed a dashboard which would connect the job market skill demand with the courses that give the required skill knowledge. We focused on job positions that require data science skills and courses that are provided by acknowledged course providers.

Our contributions are a) creating a sizable data set of data science related job postings containing the job postings title, description, locations and other information, and b) developing a dashboard which for a given query shows relevant job postings as well as courses and lectures which give the appropriate skills. The dashboard is daily updated with new job postings showing the most recent changes. Basic statistics such as the most popular job locations and skills are also shown.

The remainder of the paper is structured as follows. In section 2 we present related work. Next, data acquisition is explained in section 3 followed by the presentation of the dashboard in section 4. Finally, we discuss and conclude our work in section 5.

2. RELATED WORK

There are multiple blogs that write about top skills needed for getting a job in data science. One such blog is [14] which lists both non-technical and technical skills a data scientist should have in the coming years. Another blog [15] lists the top data science skills and courses where they can be learned. A lot of these blogs are not up-to-date and not reflecting the current state.

A research report [11] writes about connecting supply and demand in Canada's youth labor market. They were interested in finding what skills young adults acquired during their education, how employers demand is conveyed to students and those who support them and how well are the acquired skills utilized on the job. They presented their results but did not develop an application that would help to narrow the gap between the skill demand and supply.

Another report [13] talks about the mismatch of the skills young adults get during their education and the skills the companies demand. They found that skills are a critical asset for individuals, businesses and societies and that many employers report difficulties in finding suitably skilled workers. Additionally, they find that a sizable qualification mismatch is one of the biggest problems.

The company *Year Up* [9] helps young adults get the appropriate skills and the needed work experience. They identify motivated individuals and companies that are prepared to help, send the individuals to learn new skills and afterwards apply the newfound skills at the companies, getting the critical work experience for their career. The work is done manually which can be expensive and time consuming.

3. DATA ACQUISITION

Open job positions can be found using job search services. These services aggregate job postings by location, sector, applicant qualifications and skill set or type. One such service is Adzuna [6], a search engine for job ads which mostly covers English speaking countries. Another service is Trovit [7], a leading search engine for classified ads in Europe and Latin America. The service is available in 13 different languages and provides listings of jobs as well as cars, real estate and other products.

When applying for a job position the applicant requires to have a certain skill set. If the requirements are not fulfilled, he can enroll in courses to get the missing skills. Additionally, watching certain lectures can give a deeper understanding of a particular problem which can increase the probability of getting accepted for a job position. VideoLectures.NET [8] is an award-winning free and open access educational video lectures repository. It contains videos of individual lectures as well as lectures given at renown conferences.

Crawling. Since we needed a continuous flow of data, we developed a pipeline for acquiring job postings, courses and lectures. This will allow us to provide the dashboard, presented in section 4, with the most recent data. For job postings we targeted the portals like Adzuna with an emphasis on positions in Data Science and for courses we targeted different course providers, including Coursera [2], providing courses from top universities, and Hackr.io [4], a service which finds the best online programming courses & tutorials. We also targeted VideoLectures.NET to acquire video lectures containing the Data Science tag. The tags are given manually by the VideoLectures team.

For data acquisition and enrichment, we collected data either using dedicated APIs, including Adzuna API [1] as well as custom web crawlers. The data was formatted to JSON to aid further processing and enrichment.

Enriching. The next step of data preprocessing is *wikification* - identifying and linking textual components to the corresponding Wikipedia pages [16]. This is done using Wikifier [10] which also supports cross and multi-linguality enabling extraction and annotation of relevant information from job postings, courses and video lectures in different languages. Wikification will allow us to search for job postings, courses and lectures in multiple languages.

Next, we use the Skill and Recruitment Ontology (SARO) [17] to extract Data Science skills from job postings. For each job posting we match the Wikipedia concepts with the skills found in SARO ontology and declare the matched concepts as Data Science skills. These skills are then added to the job posting profile.

Finally, to allow searching by locations and countries the job postings were further enriched by using GeoNames ontology [5] to include the latitude and longitude and the corresponding GeoNames ID and the location name.

Data Set Statistics. The job postings data set contains almost 3.3M job postings acquired in the period of 18 months. Job postings were located for 144 different countries, the majority of them from Europe. Figure 1 shows the top fifteen countries with most found job postings. The UK dominates other countries with 906k job postings, followed by France with almost 539k.

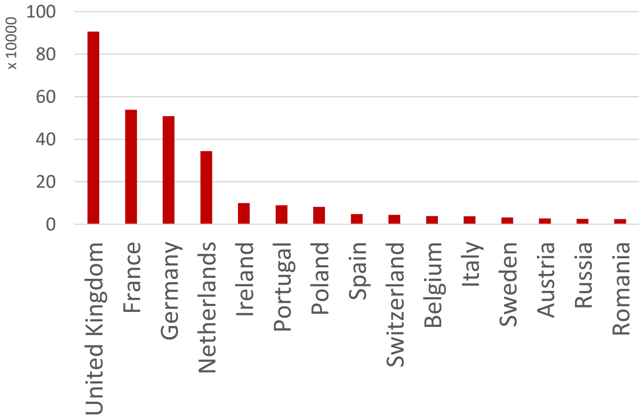


Figure 1: Top fifteen countries with most found job postings. The greatest number of job postings were found for UK, followed by France and Germany.

There were 650 unique Data Science skills extracted from the data set. These include soft skills, such as leadership and management, knowledge of a particular domain, such as machine learning and artificial intelligence, and programming languages. Figure 2 show the most demanded skills in the data set.

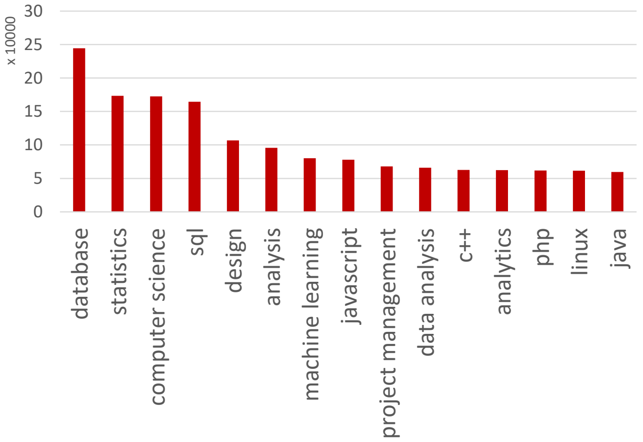


Figure 2: Top fifteen most demanded skills. They are mostly comprised of high-level skills, such as “database” and “computer science”, and programming languages.

The course data set contains over 63k course information including their title, description and course providers. The data set is comprised of over 8k courses available online and 55k offline courses. Figure 3 shows the distribution of online courses by course providers. The most courses were acquired from Coursera with above 4k, followed by Hackr.io at 2k.

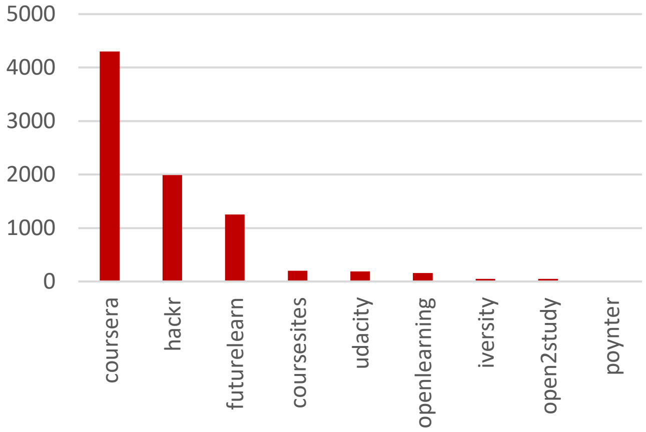


Figure 3: The distribution of online courses by course providers. The most courses were acquired from Coursera, followed by Hackr.io.

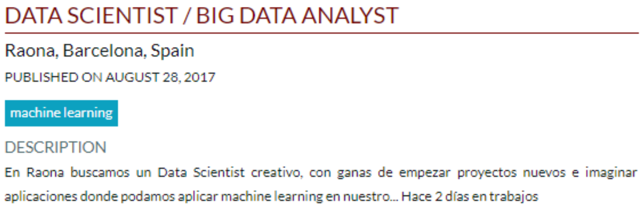
Finally, we acquired a data set of over 20k lectures published on VideoLectures.NET. It contains information about the lectures available on the video repository including title and description and link to the lecture.

4. DASHBOARD

Our objective is to automatically connect Data Science skill demand with the provided courses. To this end, we developed a dashboard [3] which enables its users to search for their desired job position, find out what is the required skill set and which are the appropriate learning materials and courses to acquire the missing skills. Additionally, the dashboard shows the most demanded skills and hiring location for the given results. In this section, we present the content retrieval methodology and describe the different components of the dashboard.

Methodology. Here we present the methodology used for retrieving the demand and supply content. The content is retrieved by inserting a query text in the search input. The user may add additional query conditions by selecting the Data Science skills, locations, countries and a time interval in which the job postings were published. Upon submitting, the query is used to fetch the content that matches the conditions. While all query values are used for retrieving job postings, only the input text and skills are used for retrieving the courses and video lectures content. Since courses and video lectures are available online the location and time interval are irrelevant for retrieving the supply content. To retrieve the content we first need to set an appropriate index. The job posting data set is indexed by Wikipedia concepts, Data Science skills, locations, countries and published date while the course and lecture data sets are indexed only by Wikipedia concepts. The query text is sent through wikification to acquire Wikipedia concepts which are used for retrieving the relevant content. Next, additional query conditions are used to filter out the content. The remaining content is used to calculate the most demanded skills and hiring locations. Finally, the query results are returned and used to update the dashboard components. This process is developed using QMiner [12], a data analytics platform for processing large-scale real-time streams containing structured and unstructured data.

Components. The dashboard is composed of different components. The largest component is a list of job postings. Each job posting is presented by its extracted information, including the Data Science skills extracted from the title and description. Figure 4 shows an example of a job posting in the list. Since Wikifier supports cross and multi-linguality the list consist of job postings written in different languages.



DATA SCIENTIST / BIG DATA ANALYST

Raona, Barcelona, Spain

PUBLISHED ON AUGUST 28, 2017

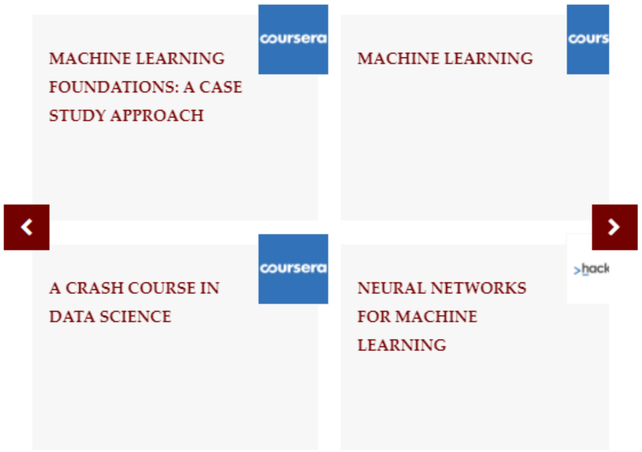
machine learning

DESCRIPTION

En Raona buscamos un Data Scientist creativo, con ganas de empezar proyectos nuevos e imaginar aplicaciones donde podamos aplicar machine learning en nuestro... Hace 2 días en trabajos

Figure 4: Example of a job posting returned by the query "machine learning". Even though the job posting is written in Spanish the methodology finds it relevant.

If the user does not have the required skill set it can be acquired by enrolling into courses shown in the course list. The list shows courses offered by different online course providers that are relevant to the users input query. Figure 5 shows the component containing the course list. Left and right arrows are used to navigate through the list where each course is presented by its name and a course provider.



MACHINE LEARNING FOUNDATIONS: A CASE STUDY APPROACH

MACHINE LEARNING

A CRASH COURSE IN DATA SCIENCE

NEURAL NETWORKS FOR MACHINE LEARNING

Figure 5: A sample of recommended courses for the query "machine learning". Clicking on a course redirects the user to the course provider where he can enroll.

Additionally, the user can watch lectures to get a deeper understanding of a problem. Similar to courses the video lectures list show relevant content found on VideoLectures.NET. Clicking the lecture redirects the user to the video lecture homepage.

The dashboard also shows them most demanded skills and job posting timeline. The timeline shows how did the ratio between queried and all job postings change since the start of the year 2016. Additionally, this shows a trend of the skill demand in the queried job posting subset. Figure 6 shows the visualizations used to show the skill demand and timeline.

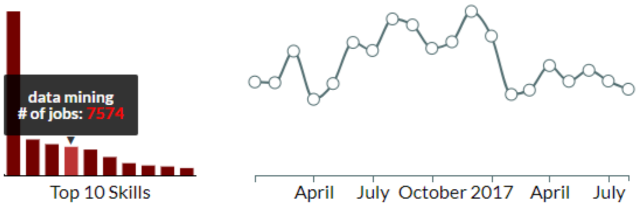


Figure 6: On the left the ten most demanded skills histogram, and on the right the number of job positions timeline, for the query "machine learning". Hovering over the histogram column shows the number of queried jobs demanding the skill.

Finally, a world map shows the most popular hiring locations extracted from the queried job postings. The locations are at first clustered where upon zooming the clusters divide and the individual locations are shown. Figure 7 show an example of clustered locations.

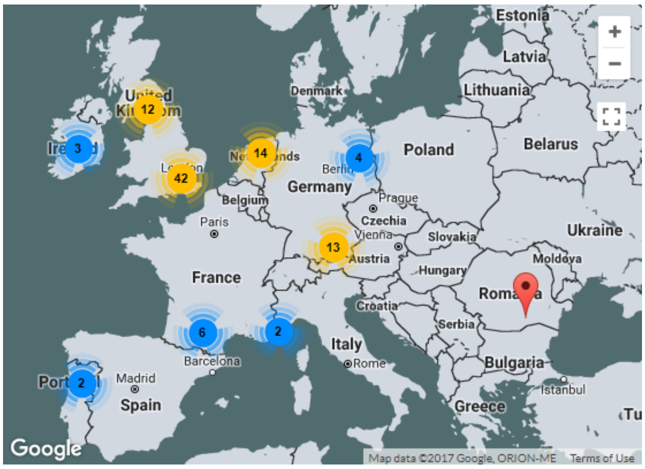


Figure 7: Top hundred hiring locations for the query "machine learning". The clusters show the number of locations it contains.

5. CONCLUSION AND FUTURE WORK

In this paper we present the methodology for automatically connecting skill demand and supply. We acquired a sizable job posting and course data set, developed a methodology for retrieving job postings, courses and lectures relevant to the user query and created a dashboard for showing the retrieved content.

In the future we wish to improve the data enriching process by handling skills that are not in the SARO ontology and add new features and improvements to the dashboard.

6. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency, EDSA EU H2020 project (Contract no: H2020-ICT-643937) and RENOIR EU H2020 project under Marie Skłodowska-Curie Grant Agreement No. 691152.

7. REFERENCES

[1] Adzuna api. `https://developer.adzuna.com/`. Accessed: 2016-09-07.

[2] Coursera | online courses from top universities. join for free. `https://www.coursera.org/`. Accessed: 2017-08-22.

[3] European data science academy dashboard. `http://jobs.videolectures.net/`. Accessed: 2017-08-23.

[4] Find the best online programming courses & tutorials - hackr.io. `https://hackr.io/`. Accessed: 2017-08-29.

[5] Geonames. `http://www.geonames.org/`. Accessed: 2017-08-23.

[6] Job search - find every job, everywhere with adzuna. `https://www.adzuna.com/`. Accessed: 2017-08-23.

[7] Trovit - a search engine for classified ads of real estate, jobs and cars. `https://www.trovit.com/`. Accessed: 2017-08-23.

[8] Videolectures.net - videolectures.net. `http://videolectures.net/`. Accessed: 2017-08-22.

[9] Year up - closing the opportunity divide. `http://www.yearup.org/`. Accessed: 2017-08-22.

[10] J. Brank. Wikifier. `http://wikifier.org/`. Accessed: 2017-08-23.

[11] R. Brisbois, L. Orton, and R. Saunders. *Connecting Supply and Demand in Canada's Youth Labour Market*. 2008.

[12] B. Fortuna, J. Rupnik, J. Brank, C. Fortuna, V. Jovanoski, M. Karlovcec, B. Kazic, K. Kenda, G. Leban, A. Muhic, et al. ■ qminer: Data analytics platform for processing streams of structured and unstructured data ■, software engineering for machine learning workshop. In *Neural Information Processing Systems*, 2014.

[13] W. E. Forum. Matching skills and labour market needs: Building social partnerships for better skills and better jobs. `http://www3.weforum.org/docs/GAC/2014/WEF_GAC_Employment_MatchingSkillsLabourMarket_Report_2014.pdf`, 2014.

[14] M. Mayo. KDnuggets analytics big data data mining and data science. `www.kdnuggets.com/2016/05/10-must-have-skills-data-scientist.html`. Accessed: 2017-08-22.

[15] E. Mcnulty. Top 10 data science skills, and how to learn them. `http://dataconomy.com/2014/12/top-10-data-science-skills-and-how-to-learn-them/`. Accessed: 2017-08-22.

[16] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384. Association for Computational Linguistics, 2011.

[17] E. Sibarani, S. Scerri, N. Mousavi, and S. Auer. Ontology-based skills demand and trend analysis, July 2016.