Research on the Demand for IT Jobs in Hungary

Data Monitoring and Textual Analysis

Capstone Project Summary for CEU Business Analytics MSc

Viktória Kónya

2022 June

Contents

1 Client Information

The Client Company is a Hungarian small enterprise that is focused on Internet research and consulting services. During my Capstone Project I worked in close cooperation with the Data Mining team of the company. The team provides data mining service to all corporate and institutional customers whose activities accumulate a large amount of data, but which, at the same time, want to support their decisions and develop their operational and business activities along their inherent interconnections.

2 Project Overview

The aim of this research project was to build a data pipeline that allows for the monitoring and the trend analysis of the demand for professionals in the IT job market in Hungary. In order to do this, a scraper was built that collects publicly available job data from 7 Hungarian job search websites on a daily basis. The open job listings from the employment websites then were used as a proxy of the demand for IT professionals.

The analytical challenge that I worked on was rooted in the fact that the dataset contained job postings from multiple IT fields starting from Software Engineers to Business Analysts. In order to quantify the demand for certain roles, the heterogenous job listings needed to be grouped into less granular job categories with similar roles. As the scraped job titles were messy, no reliable labels were available in the dataset. Because of this I used unsupervised learning technique (Topic modelling) for trying to organize the individual job postings into less granular job categories.

3 Project Summary

The project consisted of three main parts: data collection and database building, data exploration and textual analysis. Codes for each stage were written in Python.

3.1 Data collection and database creation

The goal of the first part of the project was to create a data pipeline with a clean database that the Client can use for the continuous monitoring of IT related job postings in the future. In order to ensure that anyone from the Data Mining team can continue to work on the project after the Capstone submission, all codes with the dependencies were pushed to Client's *GitHub* repository.

The source of the database is publicly available data from seven Hungarian job search websites (profession.hu, cvonline.hu, randstad.hu, jobline.hu, kellyservices.hu, kozigallas.hu, dreamjobs.hu) that were scraped on a daily basis. The automatic execution of scraper codes was scheduled with *Jenkins* jobs and are currently running from the Client's Virtual Machine. The scraped data is then uploaded to the Client's Azure File Storage.

The scraped content was difficult to work with as it contained unstructured raw data and lacked indicators that are needed for the analysis. The issues with the data quality that I spotted during the explanatory analysis were iteratively corrected in the database loading script. The main corrections impacted the title of the jobs and the hiring company on which I applied multiple string manipulations with *Regular Expressions*.

The location of the place of work was brought to a standardized administrative unit level using the *GoogleMaps* library of Python. The database was then enriched with different levels of administrative units (county, district, country) using an external data source provided by the Client.

For the text analysis part of the project, it was crucial to separate job postings with job descriptions in Hungarian and English languages. I used the *Googletrans* library of Python for the language detection and created a language flag field in the database.

Lastly, keyword search-based feature extraction was used to detect technical skills (hard skills) from the textual content of the job descriptions.

3.2 Explanatory data analysis

For the analytical part of the project the daily scraped data accumulated between 19th April 2022 and 27th May 2022 was used. Within this interval I could examine 6895 unique job postings.

The seven scraped websites altogether listed about 3000 active IT job postings each day. Profession.hu, cvonline.hu and jobline.hu are the main advertisers of open positions in IT with an average of 1025, 979 and 634 active IT job postings each day.

About 53% of the advertised IT job listings had job descriptions in Hungarian language, 46% in English language and the remaining job postings (<1%) were in German. Two of the monitored job search websites (kellyservices.hu, randstad.hu) had more than 80% of their IT job listings in English.

The most demanded IT positions were Java Developers, DevOps Engineers and Data Engineers. However, there is also a strong demand for positions that require knowledge in both business and IT such as Scrum Masters, Business Analysts and Project Managers.

Not surprisingly, 8 of the top 10 hiring companies with the most job listings are recruiting agencies. The recruiting agency with the most job postings (Randstad Hungary Kft.) had more than 900 active job postings during the roughly 1-month period.

The place of work of the job postings is concentrated in the capital with 72% of the positions located in Budapest. 15% of the job listings have mentions of opportunities for home office in the title of the job or on the main page of the website.

Regarding the technical skills SQL, Java and Python are the most demanded hard skills (excluding MS Office from the list). There is also a strong demand for professionals with knowledge in cloud technologies such as AWS or Microsoft Azure and experience with agile framework such as JIRA or Scrum/Kanban.

Regarding the soft skills communication skills, language skills, problem solving skills and analytical skills are the most frequently mentioned requirements in the job postings. There were about 1100 mentions of communication skills and about 400 mentions of language skills in the descriptions.

3.3 Text analysis

In the text analysis part of the project the goal was to organize the individual job postings into job categories based on the similarity of the job descriptions. As no trustable labels were available in the dataset, I used LDA (Latent Dirichlet Allocation) topic model for the classification problem.

The first LDA model that I built relied on the assumption that hard skills are the key determinants of the job classes. The model was built on a restricted corpus of terms which only contained the extracted technical skills. The final model classified the job postings into 13 job categories.

In the second LDA model I used the full text of the job descriptions as input. In order to improve the interpretability of the model results I applied multiple transformations and restrictions during the text preprocessing such as lemmatizing the text, forming n-grams from the frequently co-occurring terms, applying frequency thresholds and tokens with noun-type structures were kept only. The final model classified the job postings into 14 job categories.

As a sanity check I calculated the similarity between the topic vector representation of job postings with identical job titles.

4 Summary of the results

LDA is a popular topic model that provides a simple way to organize large volumes of unlabelled text. However, LDA has several drawbacks that I encountered with during the text analysis. First of all, the interpretation of the resulting topics is not always straightforward and is not free from subjective factors. Second, the number of resulting topics needs to be set in advance. Despite that there are measures that we can use to define the optimal number of topics, this will not always result in the model with the best interpretation. In this case we might need to deviate from the topic number suggested by an objective measure. Third, the model works well with a large corpus which puts limitations to my findings of the model that I built on the extracted hard skill terms.

5 Learning outcomes

By working on this research project, I could gain hands-on experience with several tools and topics that we were acquainted with during the Business Analytics programme including working with Virtual Machines, scheduling tasks with Jenkins, Python coding and text analysis. Next steps of the project could be the creation of a SQL database from the currently used datasets.

∞