Updating of Statistical Register by Web Scrapping

Jaroslav Sixta[1], Michal Čigáš[2], Jan Fojtík[3]
[1]The Czech Statistical Office, The Czech Republic
[2]The Czech Statistical Office, The Czech Republic
[3]The Prague University of Economic and Business, The Czech Republic

Abstract

Statistical offices are responsible for maintaining and updating of statistical registers that form the baseline for official statistics. In the Czech Republic, the Statistical Business Register counts about 2.7 mil legal units with different statistical attributes. However, the group of users contain not only statisticians and government authorities but also everyday users who are interested in the statistical attributes of companies, bodies etc. One of the most requested attributes is the statistical classification of economic activities abbreviated as NACE in European Union. As the statistical offices are under long-term pressure of modern governments to decrease the administrative burden of businesses it is still more and more difficult to carry out statistical surveys to detect economic activities of units in statistical business register in case it is not possible to use available administrative sources for this purpose. In addition, the assignment of NACE without automatic coding is very time consuming and requires significant workload on the staff of statistical offices. As a result, the Czech Statistical office launched the cooperation with the Prague University of Economics and Business to develop project for automatic updating of the economic activities and other variables in the statistical business register based on web scrapped data and machine-learning techniques. This paper presents how we intend to fill web addresses of enterprises into statistical business register, scrap them to obtain information about the economic activities and process the texts from webs to obtain NACE code. We also describe our experience with using ChatGPT4 for processing data from the internet database of companies. At the time of preparing this contribution, the project is still ongoing so it does not present final solution. However, the achievements that have already been reached might be fruitful and inspiring for all that intend to deal with this issue in the future.

Keywords: Statistical Business Register, Machine-learning, Web scraping, NACE classification, Automatic coding

1. Introduction

The Czech Statistical Office (CZSO) has long-term partnership with the University of Economic and Businesses[1], Prague especially in the field of national accounts and education of employees. Several members of CZSO top management studied in this university and now actively work on the various positions of its organisational structure. Several years ago, the CZSO started offering practical trainings for foreign students of this university in various areas of statistics. This activity proves to be fruitful for both sides and started cooperation on improving the CZSO statistical business register.

[1] see https://www.vse.cz/english/

2. The statistical business register

The CZSO's statistical business register (SBR)[2] covers complete economic population of the Czech Republic. It registers more than 2.7 million legal units and serves as a tool for coordination and grossing-up economic surveys and for production of business population statistics. Even though SBR is now in compliance with the EU Regulation no. 2019/2152 of the European Parliament and of the Council of 27 November 2019 on European business statistics, repealing 10 legal acts in the field of business statistics and its quality is highly evaluated by the users there are still problems with NACE activities misclassification. For updating of SBR the administrative and statistical sources are able to provide only textual descriptions of all economic activities without determining which is the main one. As a result, it is not always possible to match these descriptions with the appropriate NACE code and detect main economic activity. In many cases, this work has to be done manually and it is very burdensome and time-consuming activity for SBR administrators. In addition, the information on the change of economic activities is rarely recorded in administrative sources, thus the only source for updating it are statistical surveys, which do not cover whole population.

3. The experience with web scraping and machine learning

In order to decrease the number of misclassification the Statistical Register Department that is responsible for governing of SBR at CZSO, searched for the possibilities to update economic activities from the information presented by the economic subjects on their web pages by using web scraping and machine learning methods. This idea was based on the information acquired from the outcomes of the ESSnet Big Data published on CROS portal[3], presentations from the developing the OECD-UNSD global register on multinational enterprises[4] and experience of other EU member states with improving their business register by using web scraping and machine learning methods. Furthermore, it was also important that the CZSO top management continuously support using these techniques across the statistics to increase quality of the produced data. For several years, the CZSO has been using machine-learning methods in price statistics and the results obtained are very relevant. Currently there are several ongoing projects focused on using web scraping or machine learning in various statistical domains at CZSO.

4. The start of the project

As the Statistical Register Department does not have staff educated in the above-mentioned techniques it was necessary to search for the partner that will bring necessary knowledge to

[2] see https://www.czso.cz/csu/res/business_register
[3] see https://cros-legacy.ec.europa.eu/content/essnet-big-data-1_en
[4] see https://www.oecd.org/sdd/its/mne-platform.htm

this project. SINCE the CZSO's internal experts, which has knowledge with web scraping and machine learning methods, has already been engaged in other projects, it was decided to search for external support. As a result in 2022 Statistical Register Department engaged student of the University of Economics and Businesses to work as a trainee in this field. His main task was firstly to acquire available web information on using web scraping and machine learning methods for improving the quality of business register and secondly to propose relevant project for implementation at CZSO including the suggestion of concrete machine learning method that should be used in it. He was very successful in this task and provided us useful information for developing of the project. His traineeship finished at the end of 2022.

5. Establishing the cooperation with MODE research

Fortunately, the possibility to continue with the project occurred at the beginning of 2023. At that time, the CZSO top management secured the support of the MODE research[5]. This institution is research and consulting agency set up by the Faculty of Informatics and Statistics of the University of Economic and Businesses. It offers long-term experience with the statistical processing and analysis of data files, financial and cash-flow models, risk analysis and its diversification, computational process automation, demographic and regional analyses and forecast as well as data mining and web scraping. In addition, it provide training in the field of using statistical programmes and methods. The agency assigned two experts for this project.

6. The main objectives of the project

The project is divided into three interconnected phases. The primary goal of the first stage is to add the relevant web address to the enterprises in SBR and scrap them in order to obtain information on economic activities. In the second phase, the objective is to develop model based on the machine-learning techniques in order to predict appropriate NACE code from the scrapped information. The final stage is focused on providing information and training the CZSO staff to be able to secure the sustainability and further development of the outcomes in the future.

7. Identifying valid web pages

The first activities of the project began with searching for the available sources of information on web addresses. We contacted the CZ.NIC association that represents the interests of more than 115 internet providers in the Czech Republic. They provided us file with nearly 2 million URLs but not all of them led to existing web page. Since the recognition of a valid web page was crucial was crucial step to reduce the computational requirements and processing time, it

[5] see https://mode.vse.cz/english/#

was necessary to exclude non-existing pages from the file. The invalid web pages were defined as those that lacked any text or presented only HTTP 404 error status. This filtering step reduced the file to 477 thousand URLs with valid web pages.

8. Linking web pages to appropriate legal units

This step was focused on linking web page with appropriate unit in SBR. The Czech legislation obliges economic entities to publish on their web sites ICO or VAT number. They are unique identifiers of legal units and are kept in SBR. As a result, we were searching for these numbers in the text of each web page. For ICO, we tried to detect the string with the 8 consecutive digits, potentially separated by space in the middle. The VAT number consists of the string “CZ”, followed by 8 to 10 digits, where the digits usually represent ICO. Both identifiers respect the specific rules of MODULO 11, where the weighted sum of the ID numbers should be divisible by 11 without any remainder. Finally, we try to match the IDs identified on the web pages with the existing ones in SBR.

During the scraping process, it was observed that IDs are mostly missing in the web page linked to the URL because they usually referred to the main page. The information about IDs is usually part of the separate page with contact information. Therefore, we had to search for the link to the contact pages. To do that we tried to detect the links with the text string “onta”, which refers to the Czech words for contact. If we were successful, we were searching for ICO and VAT number on both the main and contact pages.

With this approach, we were able to identify around 80 thousand web pages with ICO and VAT numbers. After exclusion of duplicities, we finally linked one or more web pages to 57 thousand legal units in SBR.

9. Using the data from firmy.cz

The second source for getting information on web pages was firmy.cz. It is the most visited and updated internet database of companies in the Czech Republic. Moreover, for identification of enterprise it uses the same IDs as statistical business register. Because the representatives of firmy.cz did not react on our requests for data, we used Apify tool, which serves as a platform, where developers build, deploy, and monitor web scraping and browser automation tools. It helped us to identify 100 239 companies and after excluding the duplicities we link web pages to other 48 thousands legal units in SBR. Furthermore, we were able to acquire some additional, mostly contact information, which can be used for updating the SBR.

10. Analyzing the data

In total, we were able to secure at least one web page for 105.532 legal units in SBR. We analyzed them to find out whether they are relevant and suitable to start building NACE

predicting model. The data were breakdowned by number of employees, turnover, NACE or legal form. From the results we learned that there are more than 22 thousand units which have NACE codes regularly updated from statistical surveys and therefore should be accurate. As a result we decided to use them for verification of results from NACE predicting model.

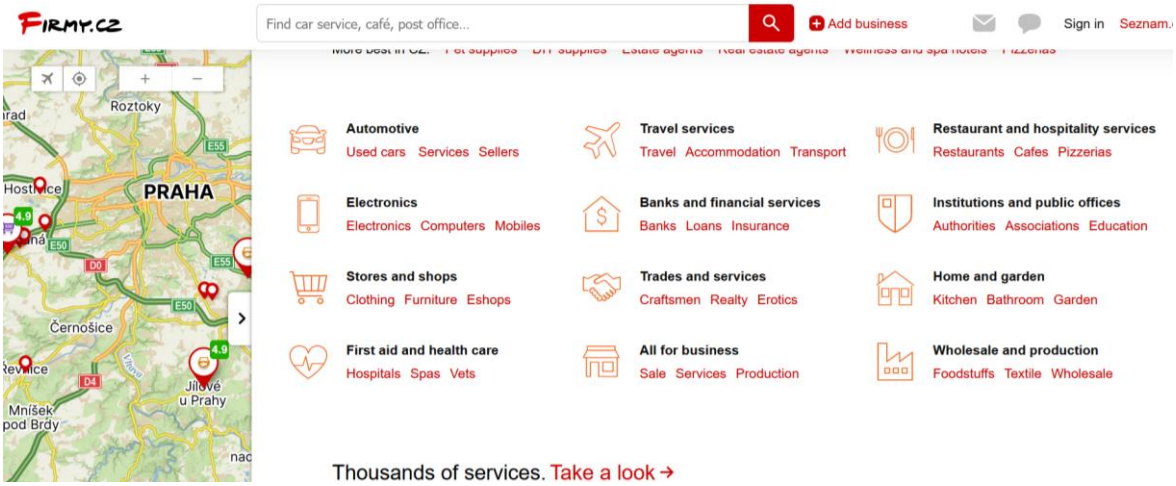
11. Development of NACE prediction model

This activity began shortly before writing this document so only the preliminary results are available. The idea is to build classification model predicting NACE codes from the text of the web page. Currently the first prototype of the model was developed. It uses the random forest classifier with no hyperparameter tuning. By this time, it was tested on training sample of 15 thousands units and the main objective was to predict correct NACE section. During the first testing, we were able to reach accuracy of 56%. Randomly selected cases with incorrect predictions from test sample were reviewed manually and, if necessary, corrected. The corrected cases are integrated into the training sample to enhance the model's classification performance. In the future steps we intend to improve the classification by tuning hyperparameters, trying different text processing techniques or classification method. Later we will concentrate on predicting the NACE codes on more detailed level. We also intend to develop dictionary with the key words for each NACE. In this field we would like to use NACE index, which is currently developed by EUROSTAT.

12. Processing the information on economic activities from firmy.cz

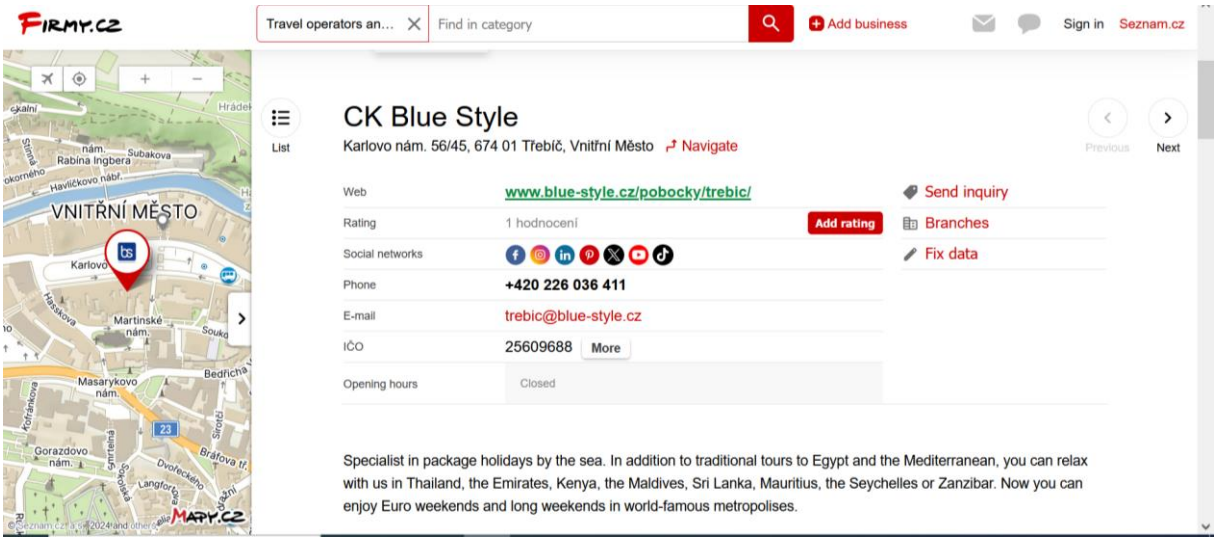
The internet database firmy.cz, which is described above, not only contains information about web pages and contacts but it also informs about their economic activities. For this purpose firmy.cz created catalogue of businesses and institutions, which encompass several economic categories (see figure 1).

Figure 1: Firmy.cz – catalogue of businesses and institutions



The company can create its profile and assign it to the appropriate economic category. In the profile it provides ID number (IČO), name, contact information, opening hours and also shortly describes its activity (see figure 2 – the activity is described below),

Figure 2: Firmy.cz – example of company profile



During the project we scraped 276 726 profiles and after excluding the duplicities (the company can present its profile in several economic categories) we were able to identify 52 thousands legal units with textual description of economic activity.

In the next step we selected approximately 6 500 enterprises from business register which are statistically significant (had 5 or more employees) and we were not able to define NACE code from administrative or statistical sources. We searched their occurrence in the file of legal units scraped from firmy.cz and were successful in 1500 cases. We assigned them 4 digit NACE by using large language model (LLM) in ChatGPT4. Finally, the business register administrators were asked to verify results manually and they confirmed the correctness of NACE codes for 1200 enterprises. As a result, we were able to decrease the population of statistically significant units with undefined NACE codes by 18,5%.

We also tried to apply this method for enterprises with undefined NACE, which have less 5 employees. In this case, we were able to assign NACE code for 3.500 out of 60.000 units we have in statistical business register. Currently, the statistical business register administrators are verifying these units.

This exercise helped us to get experience with using the ChatGPT4 language model for automatic coding of economic activities. It is very suitable model that was able to assign correct NACE code to more 80% units. According to our analysis, there is no other language model, which would be such successful without additional learning. The total cost of using ChatGPT4 for this exercise was 18 $. On the other hand, the model is not suited for dealing with large volume of data. It allows processing only 150 requests per day and in case the

number of units exceeds 100.000, the cost of using ChatGPT4 may increase to thousand dollars. Therefore, we decided to focus on using language model BERT (Bidirectional Encoder Representations from Transformers) in the future, which is not so developed, but can be gradually improved by learning.

In the next steps of the project, we intend to analyse usability of other internet databases of companies (e.g. yellow pages) for scraping. In addition, we would like to use this method as a basement for automatic coding of textual descriptions of activities obtained from administrative sources and statistical surveys in order to streamline current manual coding.

13. Conclusion

Even though the project is “in the nappies“, it is possible to come up with some conclusions. The application of web scraping or machine learning techniques is spreading every year in official statistics. They provide us time-effective and relatively cheap way to obtain data that might be impossible or difficult to gain from administrative or statistical sources. They help us to produce new statistical outcomes or improve the quality of existing ones. Therefore, the CZSO promotes applying these techniques in working procedures of the office. The cooperation with the University of Economic and Business, which is able to offer experts with many years of experience with these techniques, helps us to overcome problem with the lack of knowledge of the CZSO staff.

In this project, securing the sustainability is as important objective as improving the quality of business register. The internal staff must be able to keep and further develop the achieved outcomes without support from external supplier. Thus, we will mitigate the risk that these activities will be stopped due to budget restrictions. The involvement of experts with university background, which have long-term pedagogical experience, will make sharing the knowledge more smooth.

We believe that successful finishing of this project will decrease the number of NACE misclassification in SBR. In addition, we also intend to use acquired prediction model for determining NACE from the texture descriptions of the economic activities, which we obtain from administrative sources. This step can significantly decrease the burden of SBR administrators that have to assign the selected NACE codes manually at present. Moreover, we believe that the results of this project will ease the implementing of new NACE revision in our SBR. Last, but not least we plan to publish the results on intranet to share them with the colleagues that carry out web scrapping and machine learning techniques in other CZSO projects.

References

1. Statistical business register of CZSO, https://www.czso.cz/csu/res/business_register
2. ESSnet Big Data, https://cros-legacy.ec.europa.eu/content/essnet-big-data-1_en
3. The University of Economic and Businesses, Prague, https://www.vse.cz/english/
4. OECD-UNSD global register on multinational enterprises, https://www.oecd.org/sdd/its/mne-platform.htm
5. MODE research, https://mode.vse.cz/english/#
6. Firmy.cz, https://www.firmy.cz/
7. APIFY, https://apify.com/
8. Language model BERT, https://en.wikipedia.org/wiki/BERT_(language_model)
9. ChatGPT4, https://openai.com/gpt-4
10. NACE classification Rev. 2, https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF