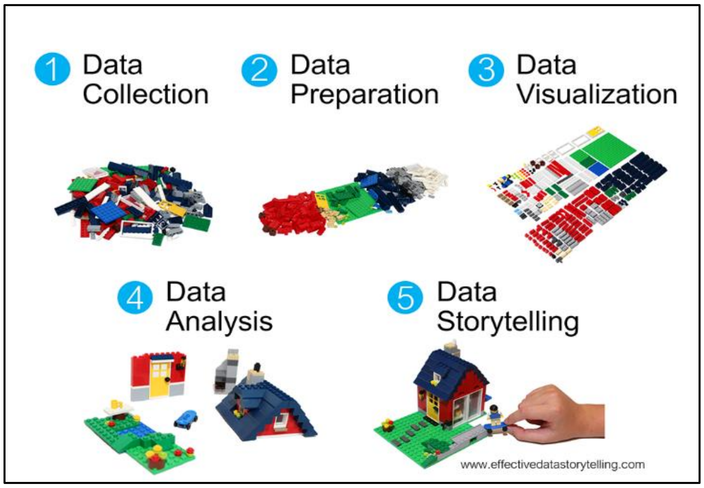
Blogpost: Aleksandar Kostadinov

Unlocking the Power of Web Data: Analyzing Occupations and Skills in demand through Big Data

New technologies provide opportunities for fast data collection. Using the web-scraping tools in Python, this post shows how the occupation and skills demand of more than 6,000 online job vacancies (OJVs) published in the most popular job portal in North Macedonia were collected, structured, visualized and analyzed.



The data sources and collection methods used for gathering labour market information are similar to those employed in other fields of studies and main characteristics of the labour market data sources can be divided broadly into three groups:

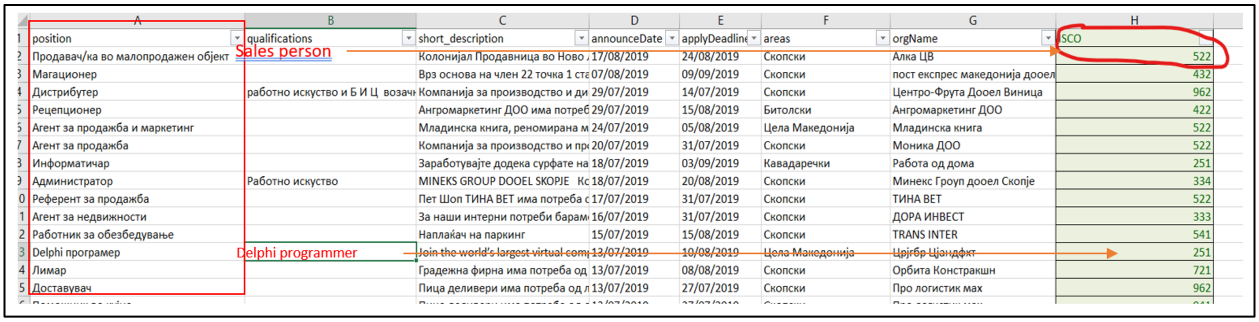
- Surveys
- Administrative data
- **Web portals (ex. Social media, Community portals , Market places, Forums, e-commerce)**

Methodology for collecting data from web portals

In this example, I used web scraping techniques and tools (Python) to download online job vacancies and the content from the most popular online job vacancy (OJV) portal in North Macedonia- www.najdirabota.com.mk. Overall, over 800 web pages were downloaded from the website, ranging between August 2018 and August 2019.

After downloading the web content, raw unstructured data was obtained, which needed to be cleaned and preprocessed for further analysis. In this step, special characters (^, [, \t,], +, |, [, \t,], +, $) and private data, such as employer addresses and phone numbers, were removed.

In the next step, it was identified the data column containing occupational names which is then compared and matched to the ISCO classification library on 3-digit level. This process allowed to obtain the ISCO 3 digit codes, as shown below.

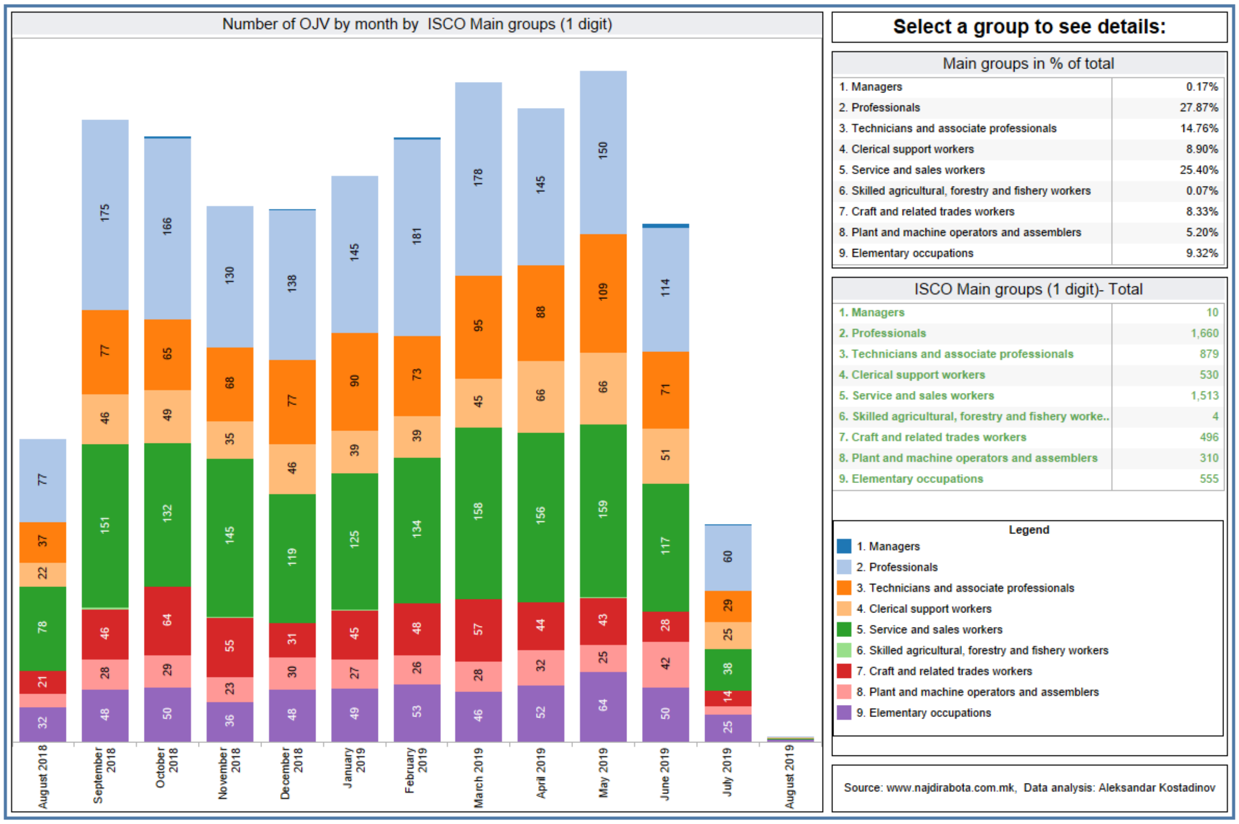


	A	B	C	D	E	F	G	H
1	position	qualifications	short_description	announceDate	applyDeadline	areas	orgName	ISCO
2	Продавач/ка во малопродажен објект		Колонијал Продавница во Ново	17/08/2019	24/08/2019	Скопски	Алка ЦВ	522
3	Магационер		Врз основа на член 22 точка 1 ста	07/08/2019	09/09/2019	Скопски	пост експрес македонија дооел	432
4	Дистрибутер	работно искуство и Б И Ц возач	Компанија за производство и ди	29/07/2019	14/07/2019	Скопски	Центро-Фрута Дооел Виница	962
5	Рецепционер		Ангромаркетинг ДОО има потреб	29/07/2019	15/08/2019	Битолски	Ангромаркетинг ДОО	422
6	Агент за продажба и маркетинг		Младинска книга, реномирана м	24/07/2019	05/08/2019	Цела Македонија	Младинска книга	522
7	Агент за продажба		Компанија за производство и пр	20/07/2019	31/07/2019	Скопски	Моника ДОО	522
8	Информатичар		Заработувајте додека сурфате на	18/07/2019	03/09/2019	Кавадаречки	Работа од дома	251
9	Администратор	Работно искуство	MINEKS GROUP DOOEL SKOPJE Кс	18/07/2019	20/08/2019	Скопски	Минекс Гроуп дооел Скопје	334
10	Референт за продажба		Пет Шоп ТИНА ВЕТ има потреба с	17/07/2019	31/07/2019	Скопски	ТИНА ВЕТ	522
11	Агент за недвижности		За наши интерни потреби барам	16/07/2019	31/07/2019	Скопски	ДОРА ИНВЕСТ	333
12	Работник за обезбедување		Наплаќач на паркинг	15/07/2019	15/08/2019	Скопски	TRANS INTER	541
13	Delphi програмер		Join the world's largest virtual com	13/07/2019	10/08/2019	Цела Македонија	Црјгбр Џјандфкт	251
14	Лимар		Градежна фирна има потреба од	13/07/2019	08/08/2019	Скопски	Орбита Констракшн	721
15	Доставувач		Пица деливери има потреба од л	13/07/2019	27/07/2019	Скопски	Про логистик мах	962

Source: print screen from Excel

Although initial expectations were to obtain country regional data in order to assess skills needs at the regional and local level, in the pre-processing phase, I found that there were too many missing fields in the cells assigned to describe job location, and that some vacancies had more than one job location entry. So, I skipped this functionality in hope that maybe there is somewhere better structured data from where these information can be obtained.

After aggregating the results on ISCO 3-digit level, it was easy to analyze the data on various levels, from 1 digit -Main ISCO groups, 2 digit Submajor groups to 3 digit ISCO groups. Publishing data vacancies also helped me to reconstruct monthly demand by occupation groups. For the visualization of the results, I used Data visualization tool Tableau, which allowed me to publish the results on a free server and create interactive dashboards for data presentation. Below are some print screens of the dashboards created in Tableau.


Number of OJV by month by ISCO Main groups (1 digit)
August 2018: 32, 21, 78, 22, 37, 77
September 2018: 48, 28, 46, 151, 46, 77, 175
October 2018: 50, 29, 64, 132, 49, 65, 166
November 2018: 36, 23, 55, 145, 35, 68, 130
December 2018: 48, 30, 31, 119, 46, 77, 138
January 2019: 49, 27, 45, 125, 39, 90, 145
February 2019: 53, 26, 48, 134, 39, 73, 181
March 2019: 46, 28, 57, 158, 45, 95, 178
April 2019: 52, 32, 44, 156, 66, 88, 145
May 2019: 64, 25, 43, 159, 66, 109, 150
June 2019: 50, 42, 28, 117, 51, 71, 114
July 2019: 25, 14, 38, 25, 29, 60
August 2019

Select a group to see details:

Main groups in % of total

1. Managers	0.17%
2. Professionals	27.87%
3. Technicians and associate professionals	14.76%
4. Clerical support workers	8.90%
5. Service and sales workers	25.40%
6. Skilled agricultural, forestry and fishery workers	0.07%
7. Craft and related trades workers	8.33%
8. Plant and machine operators and assemblers	5.20%
9. Elementary occupations	9.32%

ISCO Main groups (1 digit)- Total

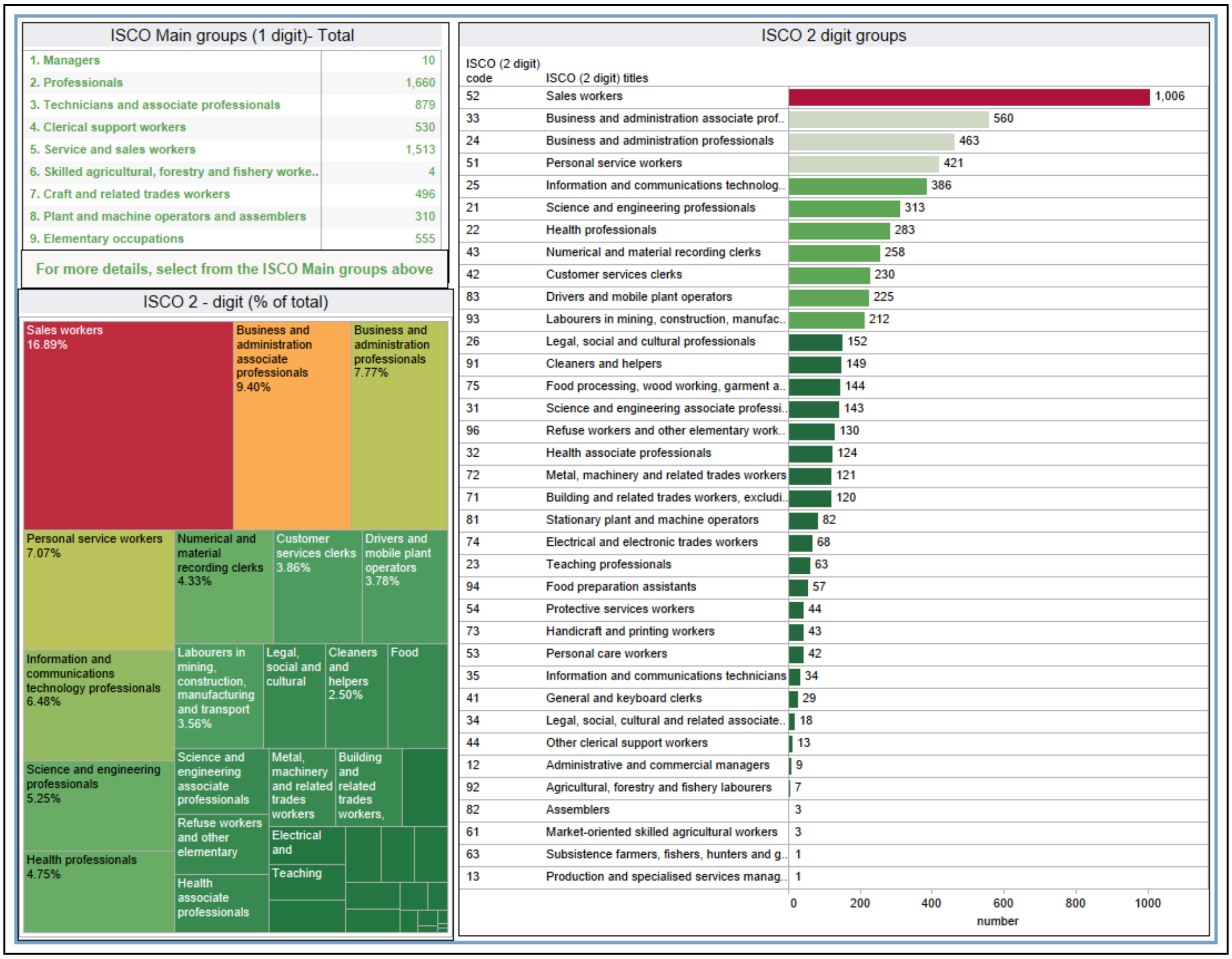
1. Managers	10
2. Professionals	1,660
3. Technicians and associate professionals	879
4. Clerical support workers	530
5. Service and sales workers	1,513
6. Skilled agricultural, forestry and fishery worke..	4
7. Craft and related trades workers	496
8. Plant and machine operators and assemblers	310
9. Elementary occupations	555

Legend

- 🟦 1. Managers
- 🟦 2. Professionals
- 🟧 3. Technicians and associate professionals
- 🟧 4. Clerical support workers
- 🟩 5. Service and sales workers
- 🟩 6. Skilled agricultural, forestry and fishery workers
- 🟥 7. Craft and related trades workers
- 🟥 8. Plant and machine operators and assemblers
- 🟪 9. Elementary occupations

Source: www.najdirabota.com.mk, Data analysis: Aleksandar Kostadinov

Note: The print screen can be seen here: https://public.tableau.com/profile/aleksandarkostadinov#!/vizhome/Tableau_15891784046170/ISCO



ISCO Main groups (1 digit)- Total

1. Managers	10
2. Professionals	1,660
3. Technicians and associate professionals	879
4. Clerical support workers	530
5. Service and sales workers	1,513
6. Skilled agricultural, forestry and fishery worke..	4
7. Craft and related trades workers	496
8. Plant and machine operators and assemblers	310
9. Elementary occupations	555

For more details, select from the ISCO Main groups above

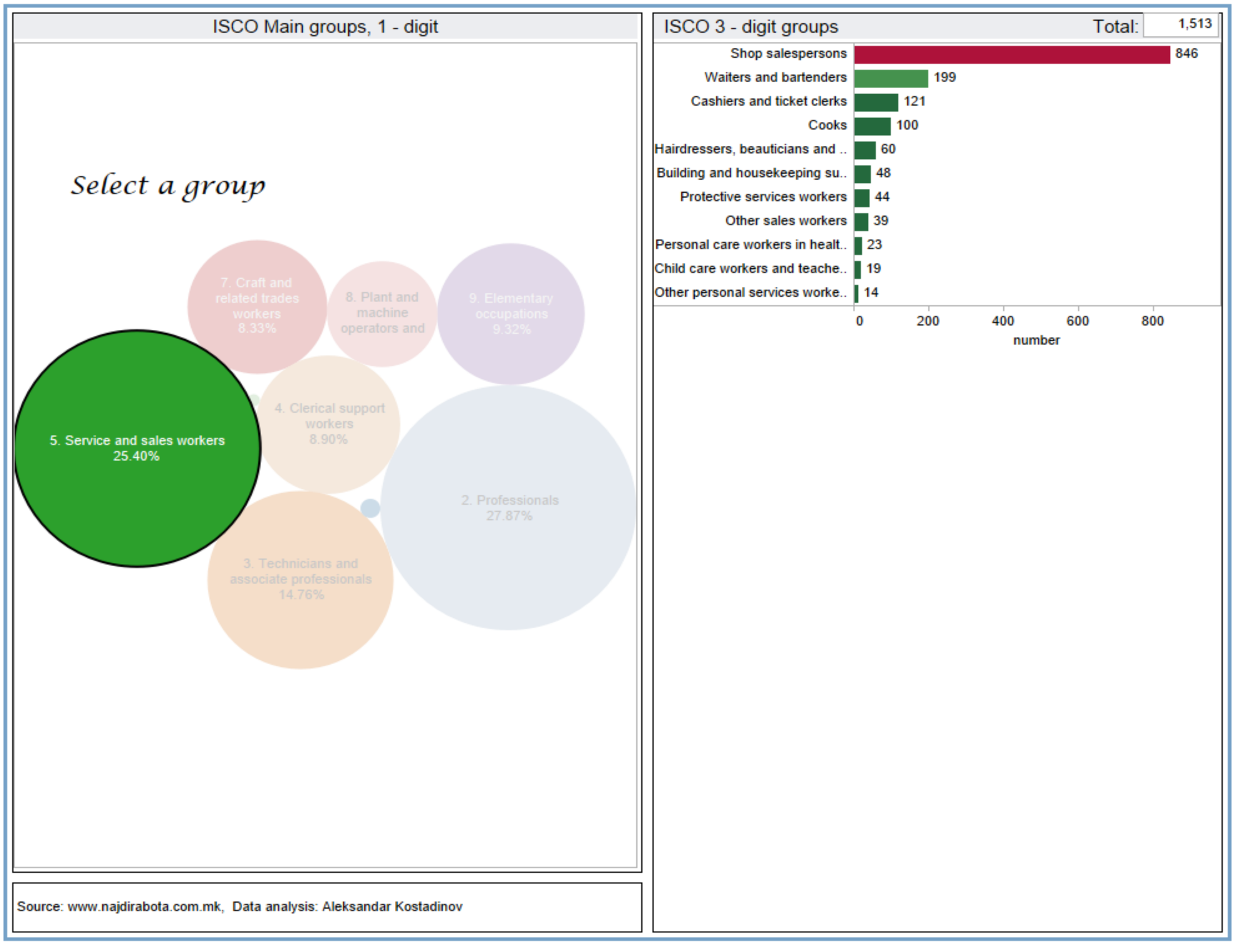
ISCO 2 - digit (% of total)

Sales workers 16.89%
Business and administration associate professionals 9.40%
Business and administration professionals 7.77%
Personal service workers 7.07%
Numerical and material recording clerks 4.33%
Customer services clerks 3.86%
Drivers and mobile plant operators 3.78%
Information and communications technology professionals 6.48%
Labourers in mining, construction, manufacturing and transport 3.56%
Legal, social and cultural
Cleaners and helpers 2.50%
Food
Science and engineering professionals 5.25%
Science and engineering associate professionals
Metal, machinery and related trades workers
Building and related trades workers,
Refuse workers and other elementary
Electrical and
Teaching
Health professionals 4.75%
Health associate professionals

ISCO 2 digit groups

ISCO (2 digit) code	ISCO (2 digit) titles	number
52	Sales workers	1,006
33	Business and administration associate prof..	560
24	Business and administration professionals	463
51	Personal service workers	421
25	Information and communications technolog..	386
21	Science and engineering professionals	313
22	Health professionals	283
43	Numerical and material recording clerks	258
42	Customer services clerks	230
83	Drivers and mobile plant operators	225
93	Labourers in mining, construction, manufac..	212
26	Legal, social and cultural professionals	152
91	Cleaners and helpers	149
75	Food processing, wood working, garment a..	144
31	Science and engineering associate professi..	143
96	Refuse workers and other elementary work..	130
32	Health associate professionals	124
72	Metal, machinery and related trades workers	121
71	Building and related trades workers, excludi..	120
81	Stationary plant and machine operators	82
74	Electrical and electronic trades workers	68
23	Teaching professionals	63
94	Food preparation assistants	57
54	Protective services workers	44
73	Handicraft and printing workers	43
53	Personal care workers	42
35	Information and communications technicians	34
41	General and keyboard clerks	29
34	Legal, social, cultural and related associate..	18
44	Other clerical support workers	13
12	Administrative and commercial managers	9
92	Agricultural, forestry and fishery labourers	7
82	Assemblers	3
61	Market-oriented skilled agricultural workers	3
63	Subsistence farmers, fishers, hunters and g..	1
13	Production and specialised services manag..	1

https://public.tableau.com/profile/aleksandarkostadinov#!/vizhome/Tableau_15891784046170/ISCO


ISCO Main groups, 1 - digit
Select a group
7. Craft and related trades workers 8.33%
8. Plant and machine operators and
9. Elementary occupations 9.32%
4. Clerical support workers 8.90%
5. Service and sales workers 25.40%
2. Professionals 27.87%
3. Technicians and associate professionals 14.76%
ISCO 3 - digit groups
Total: 1,513
Shop salespersons 846
Waiters and bartenders 199
Cashiers and ticket clerks 121
Cooks 100
Hairdressers, beauticians and .. 60
Building and housekeeping su.. 48
Protective services workers 44
Other sales workers 39
Personal care workers in healt.. 23
Child care workers and teache.. 19
Other personal services worke.. 14
0
200
400
600
800
number

Source: www.najdirabota.com.mk, Data analysis: Aleksandar Kostadinov

https://public.tableau.com/profile/aleksandarkostadinov#!/vizhome/Tableau_15891784046170/ISCO

In order to check and compare the analysis of the results obtained from the web-scraped data, I compared the statistical results with the results of the State Statical Office - Job Vacancies report. The period chosen for the SSO's survey data was July 2018 to June 2019, which is the closest comparable period to that covered by the web-scraped data (15 August 2018 to 15 August 2019). As expected, most of the discrepancies in the statistics are among "Professionals", meaning that web-based demand is much higher for Professionals than the one reported in Job vacancies survey from the State Statistical Office.
On other hand, elementary occupations and agricultural occupations are underrepresented because it is less likely that employers would find those type of workers on web portals and use another channels for recruitment.

Comparison of the most demanded ISCO main occupational groups (Q3 and Q4 of 2018 + Q1 and Q2 of 2019)

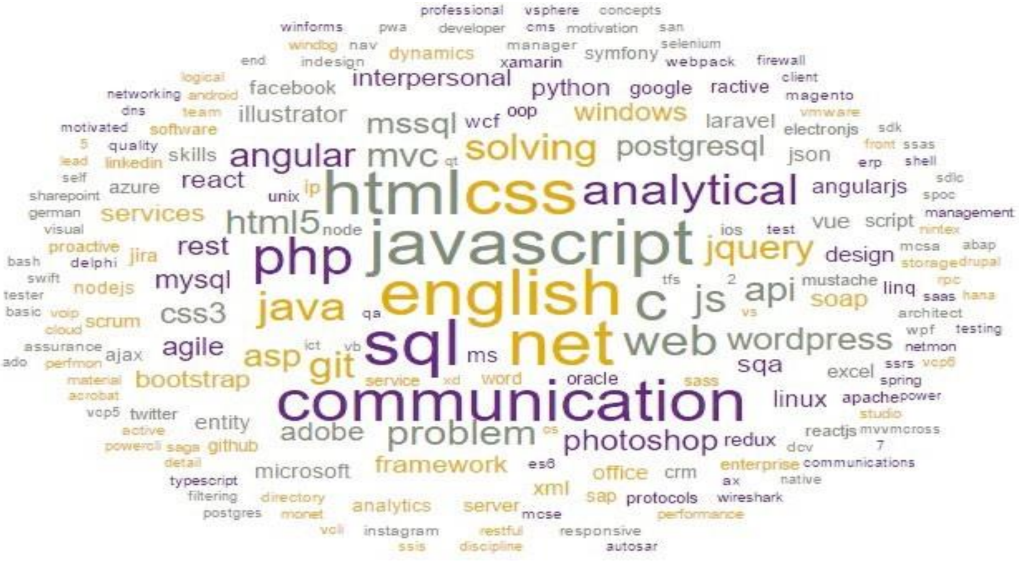
	% of total (SSO)*	Web % of total**
1. Managers	0.4	0.2
2. Professionals	7.3	27.9

3. Technicians and associate professionals	13.2	14.8
4. Clerical support workers	12.1	8.9
5. Service and sales workers	32.5	25.4
6. Skilled agricultural, forestry and fishery workers	1.2	0.1
7. Craft and related trades workers	7.1	8.3
8. Plant and machine operators, and assemblers	14.6	5.2
9. Elementary occupations	11.5	9.3

Source: *Data from the SSO news release, ‘Job Vacancies’. ** http://www.najdirabota.com.mk/. Author’s calculations.

Skills demand and real time on-line monitoring

Another use of obtained data was to analyze concrete ISCO occupational group **251- Software and applications developers and analysts**. After collecting and preprocessing the job description fields of the related ISCO group, I used Machine Learning tool Orange3 to perform textual analysis and to obtain Word cloud which represents the skills and qualifications most frequently demanded by employers looking for Software and Applications Developers.



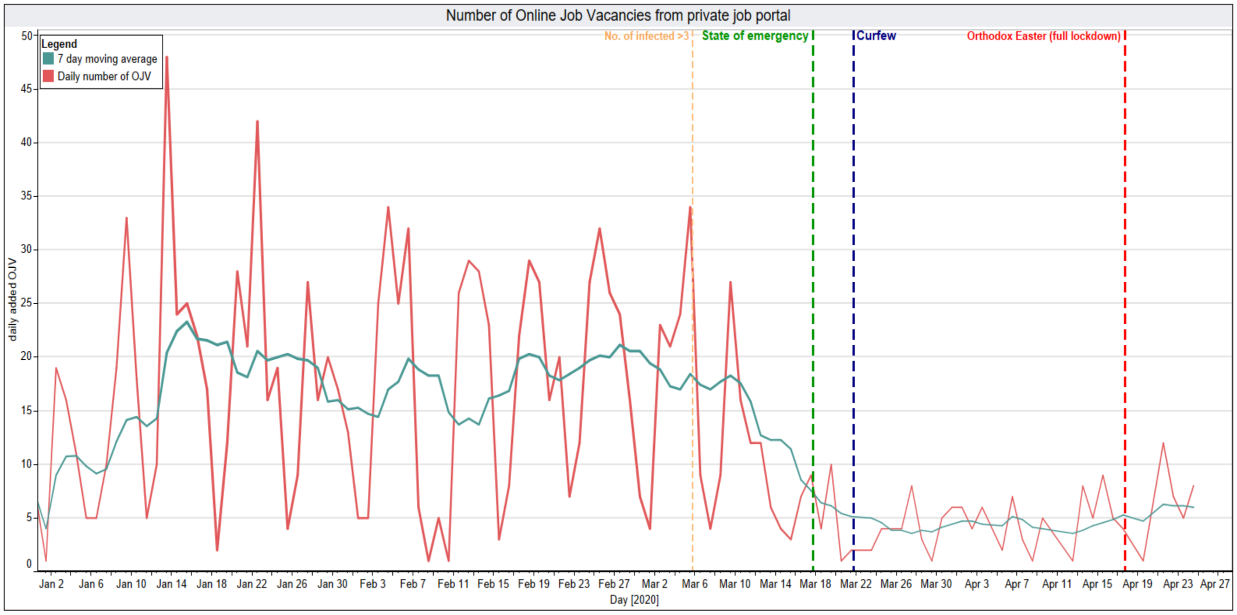
Real time on-line job monitoring

The most important benefit of this methodology is the ability to monitor real-time data from online job vacancy portals, which can then send signals regarding job demand trends to policy makers. Based on data of adverted vacancies, it can be analyzed daily volume of job vacancies by occupation and location, among other possibilities.

This discovery was really helpful, and its potentials were tested during the COVID-19 lockdowns in North Macedonia. As presented in the chart below, the number of daily vacancies had risen by mid-January 2020, reaching nearly 23 new job vacancies per day (on a 7-day average). The rising trend continued until the beginning of March 2020, at which point there was an increase in the cases of COVID19 detected in the country. A sharp decline in OJVs as of early April is clearly visible.

In the period after declaring a ‘state of emergency’ and introducing a police curfew, the number of new OJVs further decreased to a rate of between one and five new OJVs per day. Rumors about relaxing lockdown measures after the Orthodox Easter had a positive impact on employer’s expectations, with a short trend upwards showing more OJVs posted towards the end of April 2020.

Number of daily and weekly OJVs on the job portal, 1 January–25 April 2020



Source: Web data from job portal:www.najdirabota.com.mk. Author's analysis

Source: http://www.najdirabota.com.mk/. Author’s calculations.

Future considerations

This blog post offers new research methods and approaches on how to apply and use web scrapping methods for analyzing online job vacancies. Here, I must warn about the legal risks from web scrapping which may arise, and web portal owners do not always welcome and tolerate this activity.

The use of online platforms and social platforms has become increasingly popular during the Covid-19 and post covid period. This trend is expected to continue, with more employers embracing the convenience and accessibility of online job portals.

As digitalization continues to grow at a faster pace, there is a significant increase in demand for digital skills as well. Policy makers could assess and estimate this demand by occupation and location and offer educational and training courses which can meet the demand for such skills. Current traditional methods for collecting skills needs and occupations needs are time inefficient, expensive and not inclusive, meaning that use samples.

The web scraping methods despite convenient, it may challenge following challenges and potential setbacks:

- Quality and consistency of data entered on web portal;
- Lack of comparable classifications and methodologies;
- Geographical representation;
- Double input of jobs;
- Low level of participation of elementary, agriculture and administrative job occupations as usually these occupations hire through different channels.

These challenges are addressed in the full paper with Python code and can be accessed here :

European Training Foundation, Bardak, U., Rosso, F., Fetsi, *Changing skills for a changing world – Understanding skills demand in EU neighbouring countries*, Publications Office, 2021

Prepared by:
Aleksandar Kostadinov