ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl

WYDZIAŁ
ELEKTROTECHNIKI
I INFORMATYKI
POLITECHNIKI RZESZOWSKIEJ

Pracownia Problemowa II

Projekt

Gromadzenie danych z Internetu dotyczących ofert pracy w oraz ich analiza.

Przemysław Skubel, 160466

Inżynieria i analiza danych, rok V, sem. IX

Rzeszów 2024r.

ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl

WYDZIAŁ
ELEKTROTECHNIKI
I INFORMATYKI
POLITECHNIKI RZESZOWSKIEJ

Spis treści

1. Czym jest Web Scraping

Web scraping to proces automatycznego zbierania danych z witryn internetowych. Jest to technika, która polega na ekstrakcji informacji z stron internetowych poprzez analizę ich struktury HTML, CSS lub innych formatów danych. Web scraping jest szeroko stosowany w różnych dziedzinach, takich jak analiza danych, zbieranie informacji marketingowej, monitorowanie cen online, badania konkurencji i wiele innych.

2. Opis Projektu

W projekcie do zbierania danych wykorzystam bibliotekę BeautifullSoup, jest to biblioteka w języku Python do parsowania dokumentów HTML i XML. Jest często używana do ekstrakcji danych z witryn internetowych, przetwarzania dokumentów HTML/XML i manipulowania strukturą drzewa dokumentu. BeautifullSoup umożliwia łatwe nawigowanie po drzewie dokumentu HTML, wyszukiwanie, filtrowanie i modyfikowanie danych.

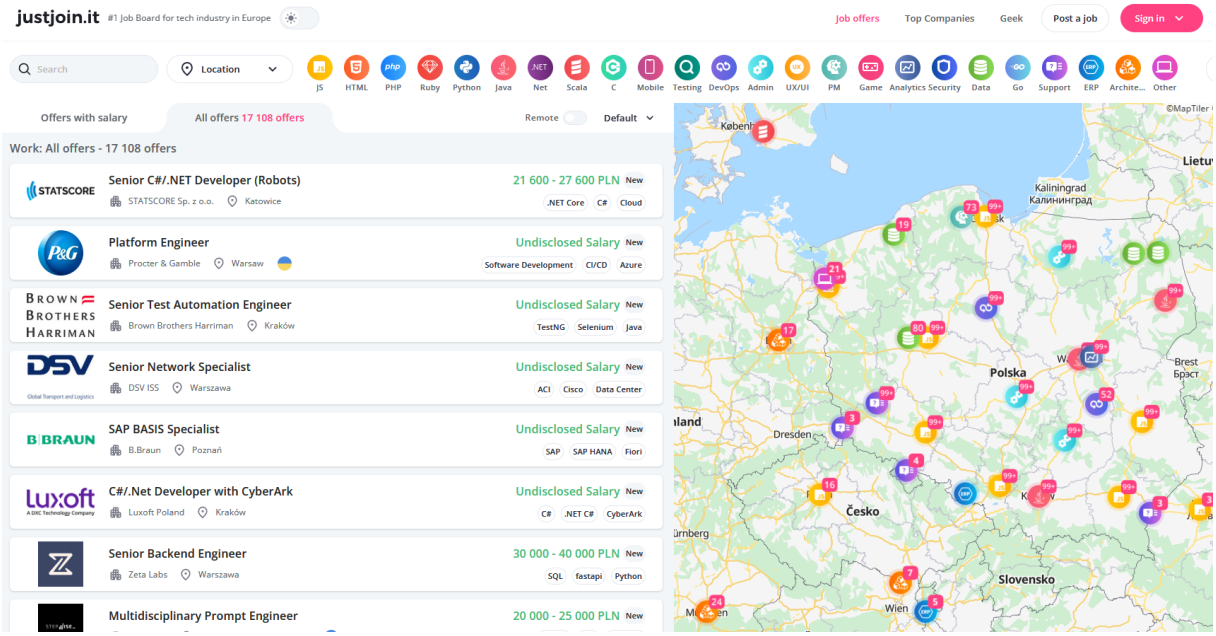
3. Zbieranie danych

Każda ze stron, z której będę zbierać informacje o dostępnych ofertach pracy w IT jest tego samego typu. Na stronie tytułowej znajdują się oferty, które możemy filtrować za pomocą dostępnych filtrów poprzez wybranie technologii, lokalizacji, widełek zarobków i tym podobnych.

3.1. Just Join IT

Pierwszą stroną z jakiej pobiorę oferty będzie strona Job Offers IT - Just Join IT. Oto jak przedstawia się strona główna pierwszej wyszukiwarki

ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl



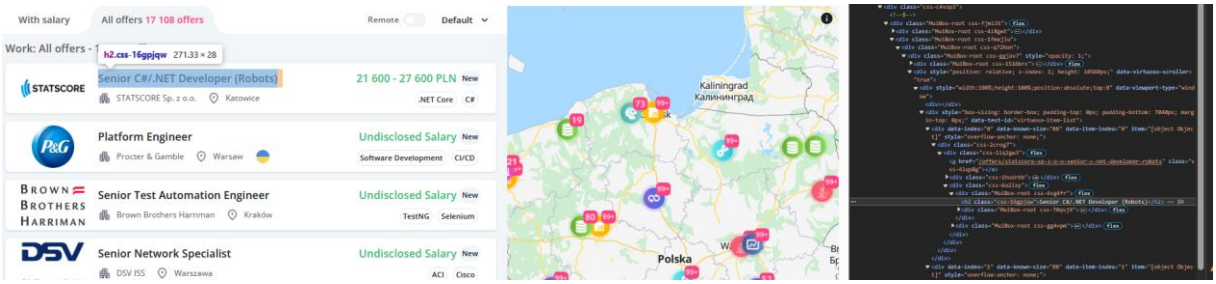
Na początku potrzebujemy tylko adres URL naszej strony.

```
url = 'https://justjoin.it/'
page = requests.get(url)
soup = BeautifulSoup(page.text, "html.parser")
print(soup.prettify())
```

```
<!DOCTYPE html>
<html>
 <head>
  <meta charset="utf-8"/>
  <link href="https://api.justjoin.it" rel="preconnect"/>
  <link href="https://public.justjoin.it" rel="preconnect"/>
  <link as="font" crossorigin="" href="https://public.justjoin.it/fonts/open-sans/open-sans-v34-latin_latin-ext-regular.woff2" rel="p
  <link as="font" crossorigin="" href="https://public.justjoin.it/fonts/open-sans/open-sans-v34-latin_latin-ext-600.woff2" rel="prelc
  <link as="font" crossorigin="" href="https://public.justjoin.it/fonts/open-sans/open-sans-v34-latin_latin-ext-700.woff2" rel="prelc
  <meta content="The first map of the labor market in the IT sector. We want to simplify the search process to minimum." property="de
  <meta content="https://justjoin.it/gfx/justjoinit/og-fallback-image.png" property="og:image"/>
  <meta content="1200" property="og:image:width"/>
  <meta content="630" property="og:image:height"/>
  <meta content="https://justjoin.it" property="og:url"/>
  <meta content="Just Join IT" property="og:title"/>
  <meta content="website" property="og:type"/>
  <meta content="Just Join IT: #1 Job Board for tech industry in Europe" name="description" property="og:description"/>
  <meta content="summary_large_image" name="twitter:card"/>
  <meta content="https://justjoin.it/gfx/justjoinit/og-fallback-image.png" property="twitter:image"/>
  <meta content="https://justjoin.it" property="twitter:site"/>
  <meta content="Just Join IT" property="twitter:title"/>
  <meta content="Just Join IT: #1 Job Board for tech industry in Europe" property="twitter:description"/>
  <meta content="width=device-width,initial-scale=1,maximum-scale=1,user-scalable=0,shrink-to-fit=no,interactive-widget=resizes-conte
  <link href="/_next/static/media/favicon_jjit_16x16.17c0797f.png" rel="icon" sizes="16x16" type="image/png"/>
  <link href="/_next/static/media/favicon_jjit_32x32.5fa64152.png" rel="icon" sizes="32x32" type="image/png"/>
 ...
   {"props":{"pageProps":{"tenantConfig":{"privacyPolicyUrl":{"rjNew":[{"url":"https://public.justjoin.it/Polityka%2BPrywatności%2BJ
  </script>
 </body>
</html>
```

ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl

Następnie na naszej stronie włączamy opcję „Zbadaj” i wyszukujemy interesujące nas treści. W tym przypadku na początku będziemy chcieli zebrać informacje na temat stanowiska.



Możemy zauważyć, że nazwy stanowisk są wpisane w akapicie <h2> w kodzie HTML, zatem możemy wykorzystać to do ich wyciągnięcia z tak zwanej „Zupy”:

```
nazwa = soup.find_all('h2')
nazwa
```

0.0s

Python

```
[<h2 class="css-16gpjqw">Senior C#/.NET Developer (Robots)</h2>,
 <h2 class="css-16gpjqw">Platform Engineer</h2>,
 <h2 class="css-16gpjqw">Senior Test Automation Engineer</h2>,
 <h2 class="css-16gpjqw">Senior Network Specialist</h2>,
 <h2 class="css-16gpjqw">SAP BASIS Specialist</h2>,
 <h2 class="css-16gpjqw">C#/.Net Developer with CyberArk</h2>,
 <h2 class="css-16gpjqw">Senior Backend Engineer</h2>,
 <h2 class="css-16gpjqw">Multidisciplinary Prompt Engineer</h2>,
 <h2 class="css-16gpjqw">Starszy Inżynier systemowy wirtualizacja Linux</h2>,
 <h2 class="css-16gpjqw">Linux/Windows BI Platform Administrator</h2>,
```

Następnie naszym celem jest wydobycie informacji o nazwie stanowiska z poszczególnego elementu listy za pomocą Regexu. Regex jest to wzorzec znaków, który służy nam do wyciągania potrzebnych nam znaków z łańcucha znaków. Do stworzenia Regexu pomogła mi strona regex101: build, test, and debug regex.

```
tekst = str(nazwa)
pattern = r'>\s*(.*?)\s*<'
mateches = re.findall(pattern, tekst)
mateches
stanowiska = []
for indeks, i in enumerate(mateches):
    if indeks % 2 == 0:
        stanowiska.append(i)
stanowiska
```

0.0s Python

```
['Manager ds. Zarządzania',
 'SASE Engineer',
 'PHP Software Developer',
 'Python Developer - TALENT POOL',
 'IT Project Manager - Banking 💰',
 'Java Tech Lead',
 'Microsoft Power Platform Engineer',
 'Entry ServiceNow Specialist',
 'Administrator ds. Utrzymania Systemów Linux',
```

Za pomocą prostej pętli i regexu wyciągamy interesujące nas informacje. Identycznie postępujemy z kolejnymi informacjami, które będziemy chcieli zawrzeć w naszym zbiorze.

Za każdym razem jednak zmieniamy sposób wyszukiwania tj. czy informacje znajdują się w nagłówku lub , jeżeli znajdują się w jednym z nich, będziemy musieli zawęzić zbiór informacji, do najczęściej klasy, w której znajdują się potrzebne dane.

```
lokalizacja = soup.find_all('span', class_ = 'css-1o4wo1x')
lokalizacja
```

0.2s

```
[<span class="css-1o4wo1x">Łódź</span>,
 <span class="css-1o4wo1x">Warszawa</span>,
 <span class="css-1o4wo1x">Gdańsk</span>,
 <span class="css-1o4wo1x">Warszawa</span>,
 <span class="css-1o4wo1x">Warsaw</span>,
```

ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl

WYDZIAŁ
ELEKTROTECHNIKI
I INFORMATYKI
POLITECHNIKI RZESZOWSKIEJ

```
tekst = str(lokalizacja)
pattern = r'>\s*(.*?)\s*<'
loc = re.findall(pattern, tekst)
lokalizacja = []
for indeks, i in enumerate(loc):
    if indeks % 2 == 0:
        lokalizacja.append(i)
lokalizacja
```

0.1s

```
['Łódź',
 'Warszawa',
 'Gdańsk',
 'Warszawa',
 'Warsaw',
 'Warszawa',
```

```
widelki = soup.find_all('div', class_ = 'css-1b2ga3v')
widelki
```

```
pattern = r'\b(\d{1,3}(?: \d{3})*(?:\.\d{2})?)\b'
```

0.0s

```
widly = []
for indeks, i in enumerate(widelki):
    w1 = str(widelki[indeks])
    w2 = re.findall(pattern, w1)
    widly.append(w2)
widly
```

```
Widełki = []
```

0.0s

```
for indeks, i in enumerate(widly):
    l = []
    for j in range(2):
        if len(widly[indeks]) == 2:
            a = widly[indeks][j]
            b = a.replace(" ","")
            # print(b)
            b = int(b)
            l.append(b)
        else:
            continue
    Widełki.append(l)
```

0.0s

```
Widełki
```

0.0s

```
[[7670, 11850],
 [27000, 32000],
 [],
```

```
# currency (waluta)
waluty = soup.find_all('span', class_ = 'css-jmy9db')
waluty
```

```
waluta = []
for indeks, i in enumerate(waluty):
    txt = str(waluty[indeks])
    pattern = r'<span class="css-jmy9db">(.*?)<\/span>'
    w = re.findall(pattern, txt)
    waluta.append(w)
waluta
```

```
waluta1 = []
for indeks, i in enumerate(waluta):
    k = waluta[indeks][0]
    waluta1.append(k)
waluta1
```

0.0s

```
['pln',
 'pln',
 'pln',
 'pln',
 'pln',
```

W przypadku waluty, niestety nie pojawiają się w kodzie z powodu braku w podanych widełek. Z tego powodu pobierana jest zdecydowanie mniejsza ilość danych niż w pozostałych kategoriach. Z powodu braku danych (w tym przypadku brakuje aż 72 rekordów), jestem zmuszony do pominięcia tej kategorii w analizie. Jednak z moich obserwacji wynika, że każde widełki są w walucie PLN.

```
data = {
    'Firma' : firma,
    'Stanowisko' : stanowiska,
    'Lokalizacja' : lokalizacja,
    'Widełki' : widly,

}
```

[92] ✓ 0.0s

```
df = pd.DataFrame(data) # utworzenie DataFrame
df.to_excel('JustJoinIT.xlsx', index = False) # zapis tabeli do Excel
df
```

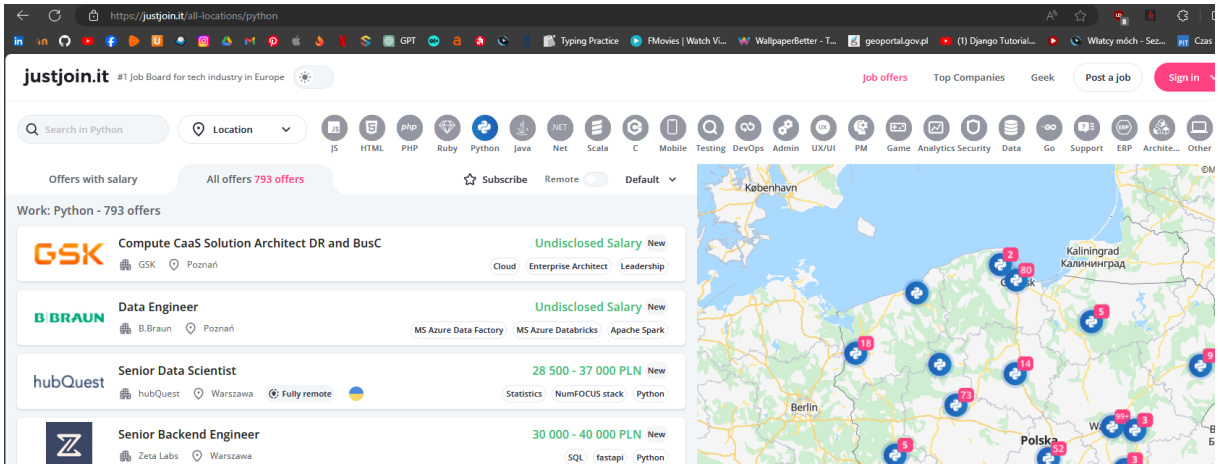
[94] ✓ 0.2s

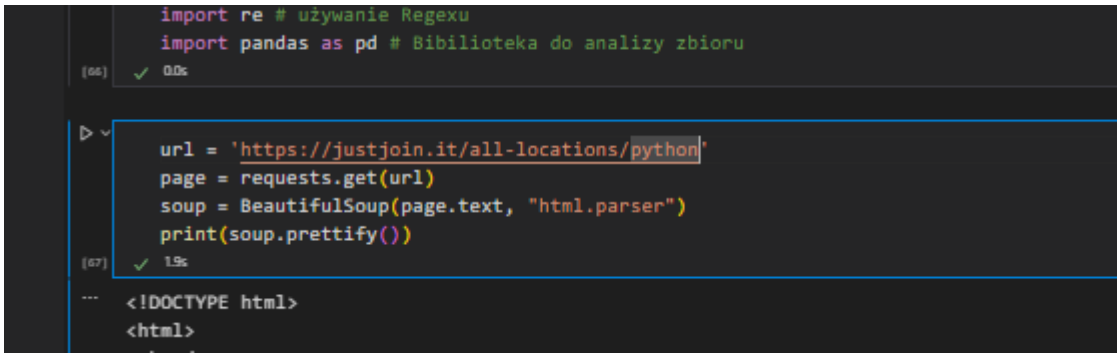
	Firma	Stanowisko	Lokalizacja	Widełki
0	GSK	Product Manager - PAT System	Poznań	[]
1	Human4Human Recruitment	C++ Engineering Lead (Medical industry)	Kraków	[]
2	Human4Human Recruitment	Odoo and Big Data Architect	Kraków	[(32 000, 36 000)]
3	Code Poets	Developer C++	Wrocław	[(17 000, 23 500)]
4	Human4Human Recruitment	Senior Google Analyst	Kraków	[(20 000, 24 000)]
...	...	...	...	...
95	DSV ISS	Senior Back-end Java Developer	Warszawa	[]
96	Schibsted	Engineering Manager .Net	Kraków	[]
97	Schibsted	Engineering Manager	Kraków	[]
98	DRÄXLMAIER	V&V Engineer - Cybersecurity Testing	Gliwice	[]
99	ITDS	GoLang Developer - Analytics Consulting Team -...	Warsaw	[(19 000, 22 000)]

100 rows × 4 columns

Tak przedstawia się tabela zebranych przeze mnie danych ze strony JustJoinIT. Warto wspomnieć o tym, że strona oferuje filtrowanie ofert, które następnie jest zapisywane w adresie URL strony. Zatem, jeśli byśmy chcieli zebrać informacje tylko i wyłącznie informacje odnośnie ofert w np. języku Python, musimy tylko zaznaczyć na stronie technologię Python, a następnie zamienić początkowy adres URL na otrzymany po dodaniu filtra.

ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl

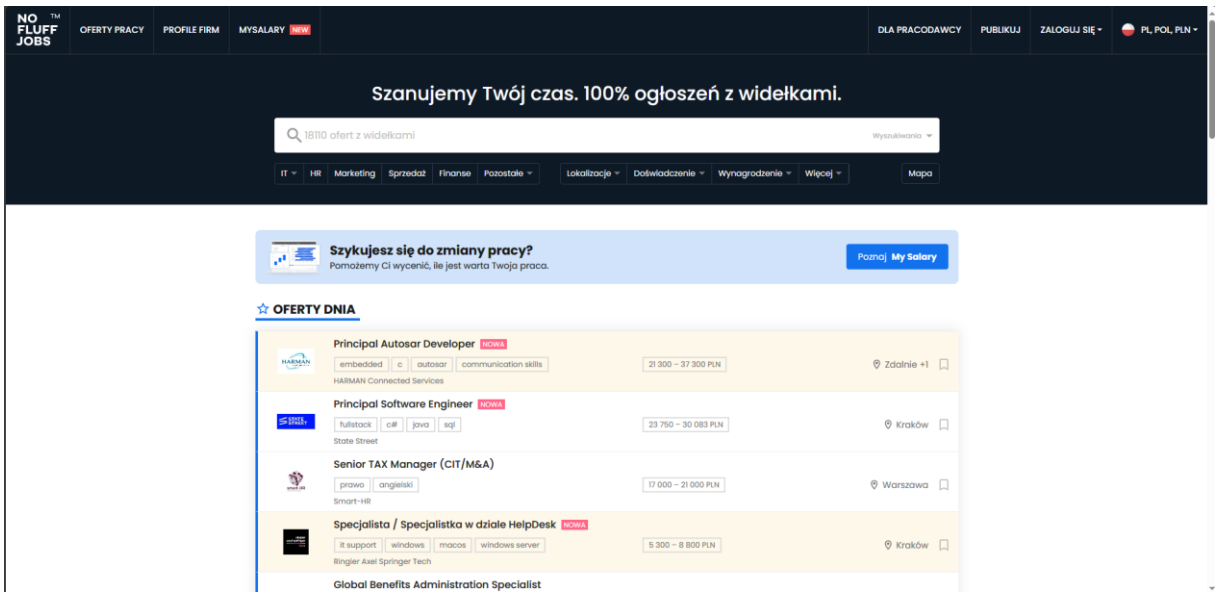




3.2. No Fluff Jobs

Drugą z kolei stroną będzie strona [Portal z ofertami pracy IT | Praca IT dla Ciebie | No Fluff Jobs] tej znajdują się poza ofertami pracy dla programistów również stanowiska w np. Human Resources lub stanowisk stricte dla inżynierów.

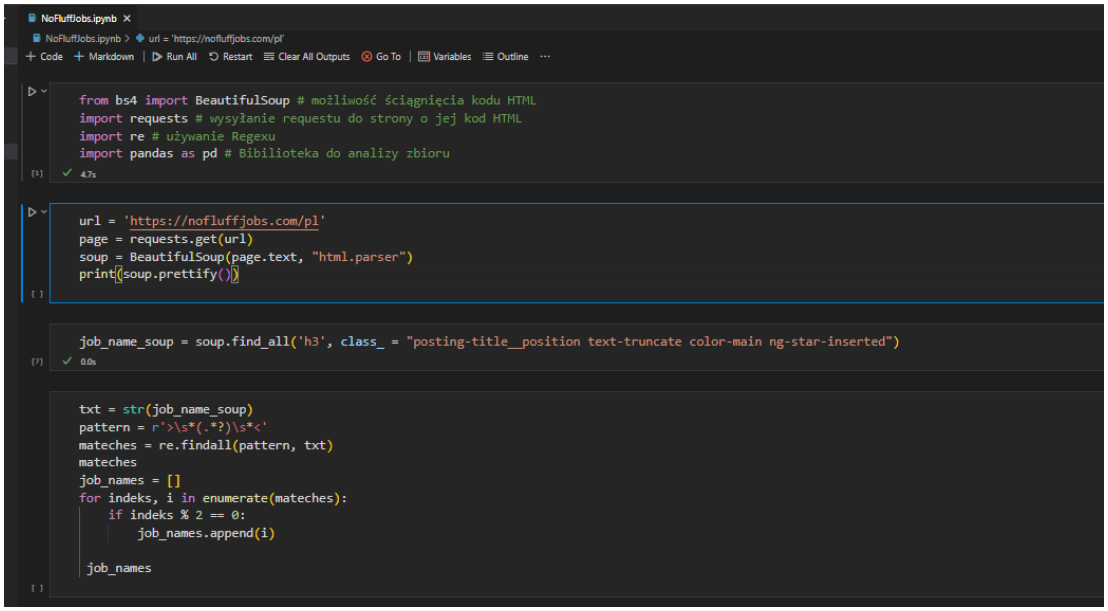
Oto jak przedstawia się strona tytułowa:

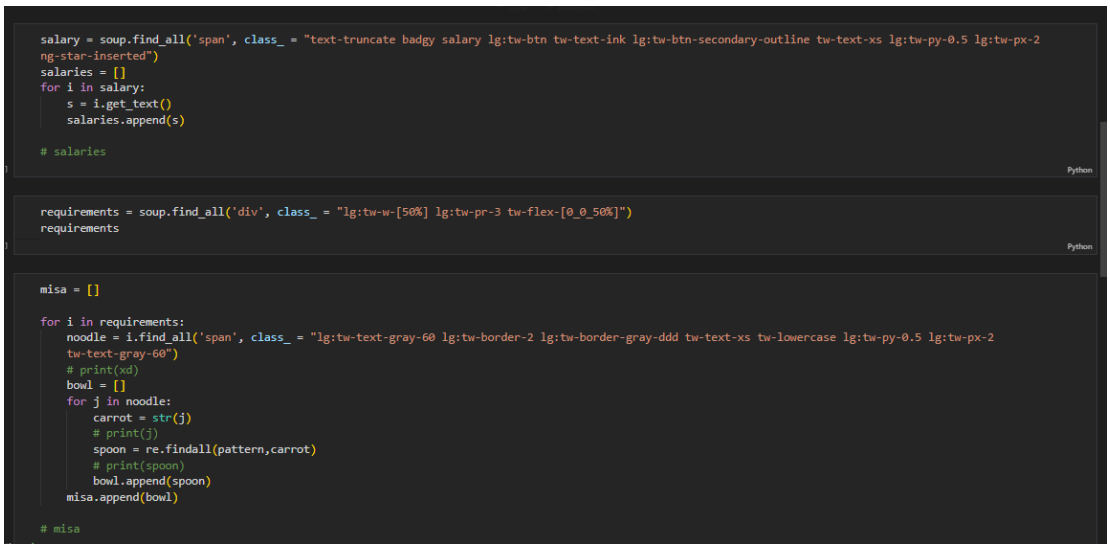


Wszystkie operacje zbierania danych są analogiczne jak przy zbieraniu informacji ze strony Just Join IT z małymi różnicami w kodach regexu. No Fluff Jobs wymagało również więcej operacji na pętlach i głębszego wyciągania danych z kodu HTML, ponieważ informacje na tej stronie są inaczej rozmieszczone i ułożone w bardziej skomplikowanym i złożonym formacie.

ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl

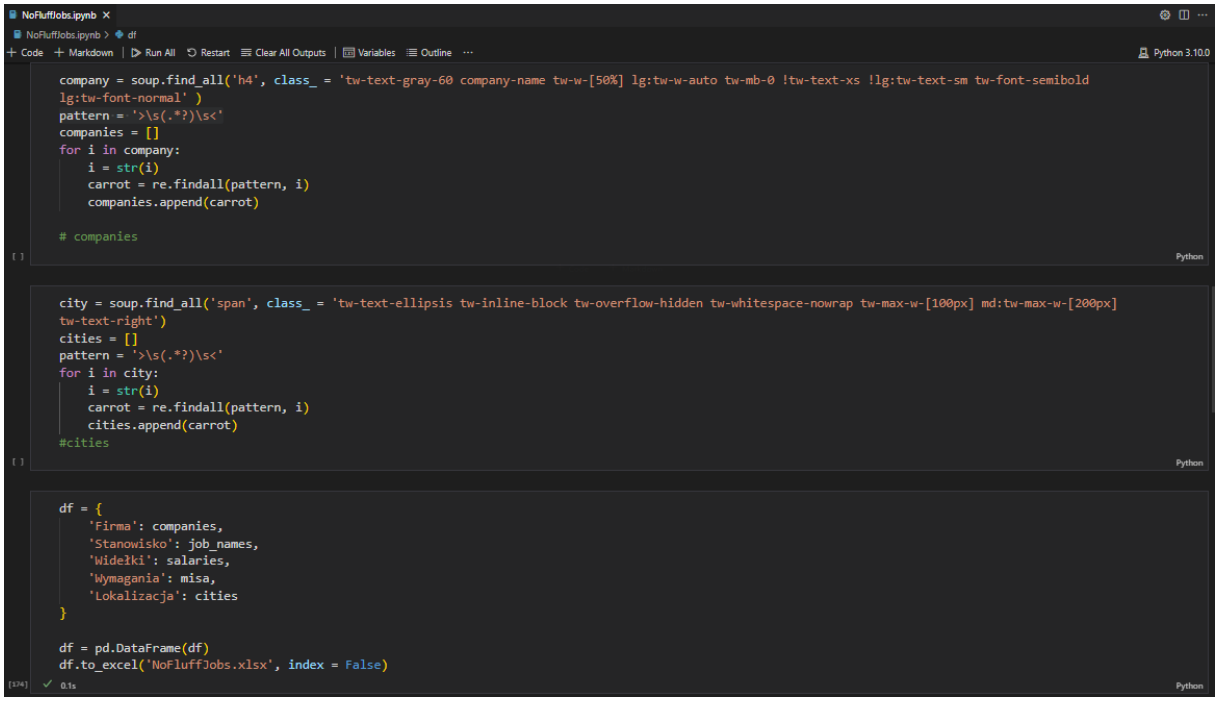
WYDZIAŁ
ELEKTROTECHNIKI
I INFORMATYKI
POLITECHNIKI RZESZOWSKIEJ





ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl

WYDZIAŁ
ELEKTROTECHNIKI
I INFORMATYKI
POLITECHNIKI RZESZOWSKIEJ



```
company = soup.find_all('h4', class_ = 'tw-text-gray-60 company-name tw-w-[50%] lg:tw-w-auto tw-mb-0 !tw-text-xs !lg:tw-text-sm tw-font-semibold
lg:tw-font-normal' )
pattern = '>\s(.*?)\s<'
companies = []
for i in company:
    i = str(i)
    carrot = re.findall(pattern, i)
    companies.append(carrot)

# companies
```

```
city = soup.find_all('span', class_ = 'tw-text-ellipsis tw-inline-block tw-overflow-hidden tw-whitespace-nowrap tw-max-w-[100px] md:tw-max-w-[200px]
tw-text-right')
cities = []
pattern = '>\s(.*?)\s<'
for i in city:
    i = str(i)
    carrot = re.findall(pattern, i)
    cities.append(carrot)
#cities
```

```
df = {
    'Firma': companies,
    'Stanowisko': job_names,
    'Widełki': salaries,
    'Wymagania': misa,
    'Lokalizacja': cities
}

df = pd.DataFrame(df)
df.to_excel('NoFluffJobs.xlsx', index = False)
```

```
df
```

	Firma	Stanowisko	Widełki	Wymagania	Lokalizacja
0	[State Street]	Senior UI/UX Developer	17 416 – 30 083 PLN\n	[[Design], [ui], [ui design], [ux design]]	[Kraków]
1	[Alsendo Sp. z o. o.]	Remote Data/ Business Analyst (data provisioning)	21 840 – 23 520 PLN\n	[[Data], [data mapping], [csv], [business anal...	[Warszawa]
2	[LeasingTeam Professional]	Logisztikai SAP koordinátor-6355	10 315 – 12 608 PLN\n	[[Logistyka], [sap], [węgierski], [angielski]]	[Warszawa]
3	[State Street]	Principle DBA Engineer	23 750 – 30 083 PLN\n	[[Backend], [relational database], [operating ...	[Kraków]
4	[Neontri]	HR Specialist with German	10 500 – 14 280 PLN\n	[[HR], [niemiecki]]	[Zdalnie]
5	[7N]	Senior DevOps Developer	23 750 – 30 083 PLN\n	[[DevOps], [python], [powershell], [perl]]	[Zdalnie]
6	[State Street]	Specjalista / Specjalistka w dziale HelpDesk	5 300 – 8 800 PLN\n	[[IT Support], [windows], [macos], [windows se...	[Gdańsk]
7	[Ringier Axel Springer Tech]	Senior .NET Developer	16 000 – 20 000 PLN\n	[[Fullstack], [.net], [.net core], [c#]]	[Kraków]
8	[Inter Cars S.A.]	Remote Senior PHP Developer (Magento)	10 000 – 20 000 PLN\n	[[Backend], [php], [magento], [html]]	[Warszawa]
9	[Forward Thinking Systems Polska]	Senior .NET Developer	16 000 – 23 000 PLN\n	[[Backend], [.net], [c#], [azure devops]]	[Zdalnie]
10	[CTHINGS.CO]	Software Development Engineer	20 000 – 25 000 PLN\n	[[Backend], [golang], [java], [sql]]	[Warsaw]
11	[Exorigo-Upos S. A.]	Senior PHP Developer	15 000 – 18 000 PLN\n	[[Backend], [php], [oop], [symfony]]	[Zdalnie]
12	[Hitachi Energy]	Mid Java/Kotlin/Scala Developer	11 424 – 17 136 PLN\n	[[Backend], [java], [spring], [git]]	[Krakow]
13	[HL Tech]	Frontend Engineer (React, Next.js, AWS)	15 000 – 21 000 PLN\n	[[Frontend], [react], [html], [css]]	[Zdalnie]
14	[LegalZoom]	Senior Frontend Developer	20 000 – 30 000 PLN\n	[[Frontend], [react], [javascript], [typescript]]	[Kraków]

Podobnie jak w przypadku strony Just Join IT, aby pobierać informacje tylko o stanowiskach z filtrami, które nas interesują, musimy wybrać te filtry i następnie wprowadzić zmieniony URL do programu.

Jeżeli będziemy chcieli zbierać dane masowo, niestety nie jesteśmy w stanie tego zrobić. Klikając na dole strony „załaduj więcej ofert", kod URL zmienia się na: https://nofluffjobs.com/pl/?page=2, co mogłoby nam umożliwić ustawienie pętli i zapisywanie bardzo dużej ilości danych. Niestety nie jesteśmy w stanie tego zrobić. W pliku „proba1.py", zastosowałem taki kod, który zapisuje dane z pierwszych 5 stron, ale dane są takie same w każdym arkuszu. Powodem tego jest to, że strona nie generuje więcej kodu HTML, dopóki nie klikniemy banneru „Załaduj więcej", a przejście bezpośrednio z linku do strony np. 4, przekierowuje nas na sam początek strony.

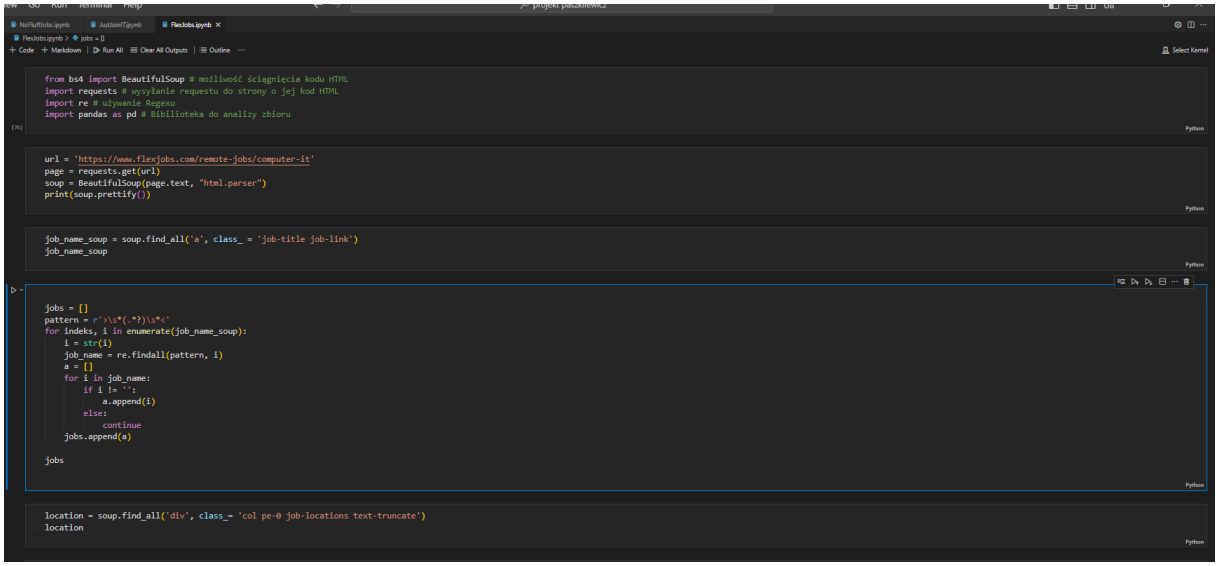
3.3. FlexJobs

Trzecią stroną, z której zbiorę informację jest Computer & IT Jobs - Remote Work From Home & Online | FlexJobs która zawiera oferty z branży IT jak i innych branż inżynieryjnych. Strona jest anglojęzyczna (oferty USA) i oferuje stanowiska w większości pracy zdalnej. Wybrałem tą stronę, ponieważ przy pobieraniu danych, widnieje jej opis, co pozwala nam na lepszą analizę tych danych.

Podobnie jak przy scrapowaniu poprzednich stron, cała technika jest analogiczna.

ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl

WYDZIAŁ
ELEKTROTECHNIKI
I INFORMATYKI
POLITECHNIKI RZESZOWSKIEJ



```
from bs4 import BeautifulSoup # możliwość ściągnięcia kodu HTML
import requests # wysyłanie requestu do strony o jej kod HTML
import re # używanie Regexu
import pandas as pd # Bibilioteka do analizy zbioru
```

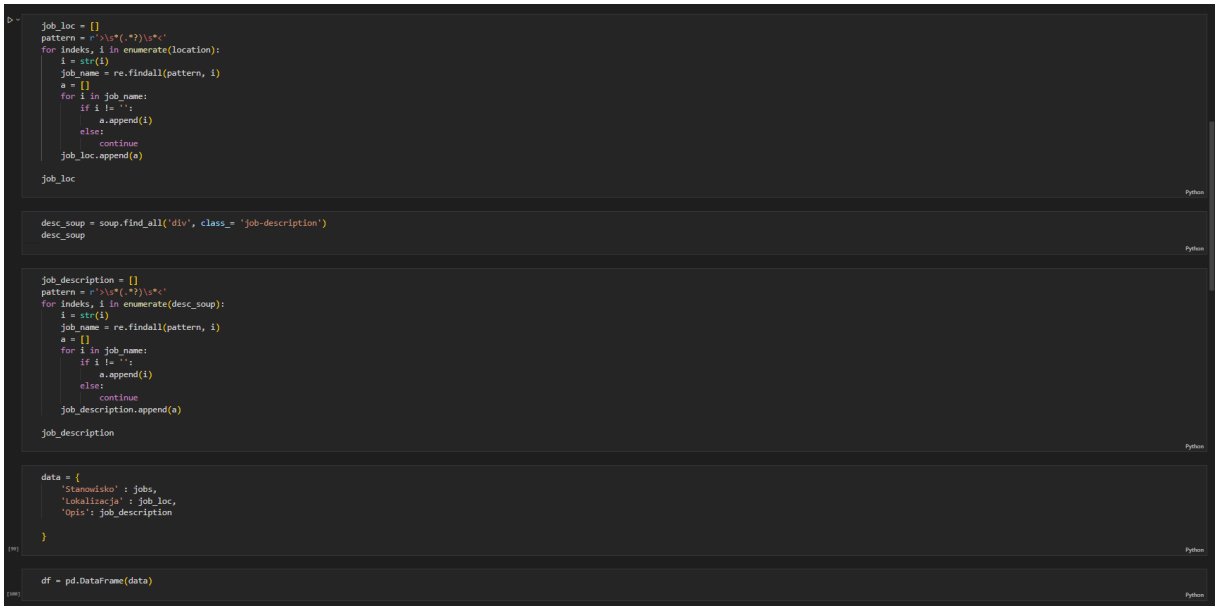
```
url = 'https://www.flexjobs.com/remote-jobs/computer-it'
page = requests.get(url)
soup = BeautifulSoup(page.text, "html.parser")
print(soup.prettify())
```

```
job_name_soup = soup.find_all('a', class_ = 'job-title job-link')
job_name_soup
```

```
jobs = []
pattern = r'>\s*(.*?)\s*<'
for indeks, i in enumerate(job_name_soup):
    i = str(i)
    job_name = re.findall(pattern, i)
    a = []
    for i in job_name:
        if i != '':
            a.append(i)
        else:
            continue
    jobs.append(a)

jobs
```

```
location = soup.find_all('div', class_= 'col pe-0 job-locations text-truncate')
location
```



```
job_loc = []
pattern = r'>\s*(.*?)\s*<'
for indeks, i in enumerate(location):
    i = str(i)
    job_name = re.findall(pattern, i)
    a = []
    for i in job_name:
        if i != '':
            a.append(i)
        else:
            continue
    job_loc.append(a)

job_loc
```

```
desc_soup = soup.find_all('div', class_= 'job-description')
desc_soup
```

```
job_description = []
pattern = r'>\s*(.*?)\s*<'
for indeks, i in enumerate(desc_soup):
    i = str(i)
    job_name = re.findall(pattern, i)
    a = []
    for i in job_name:
        if i != '':
            a.append(i)
        else:
            continue
    job_description.append(a)

job_description
```

```
data = {
    'Stanowisko' : jobs,
    'Lokalizacja' : job_loc,
    'Opis': job_description

}
```

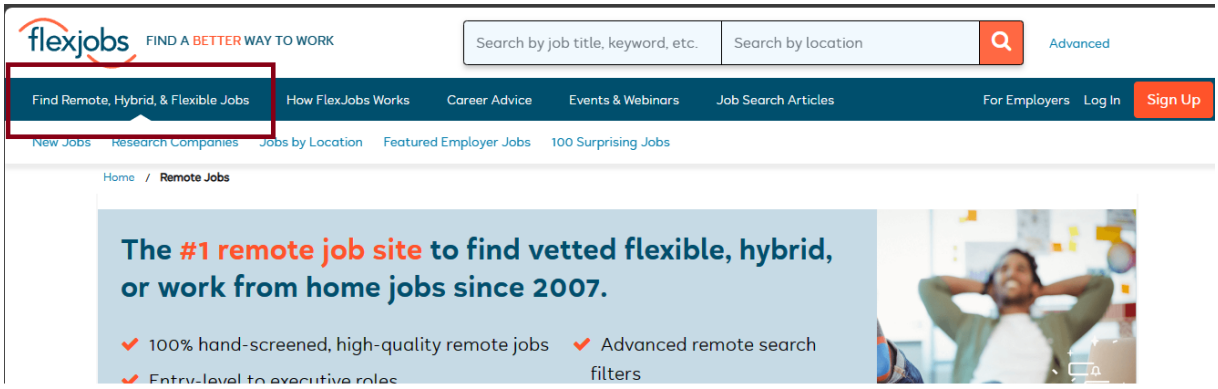
```
df = pd.DataFrame(data)
```

```
df = pd.DataFrame(data)
df.to_excel('FlexJobs.xlsx', index = False)
```

```
df
```

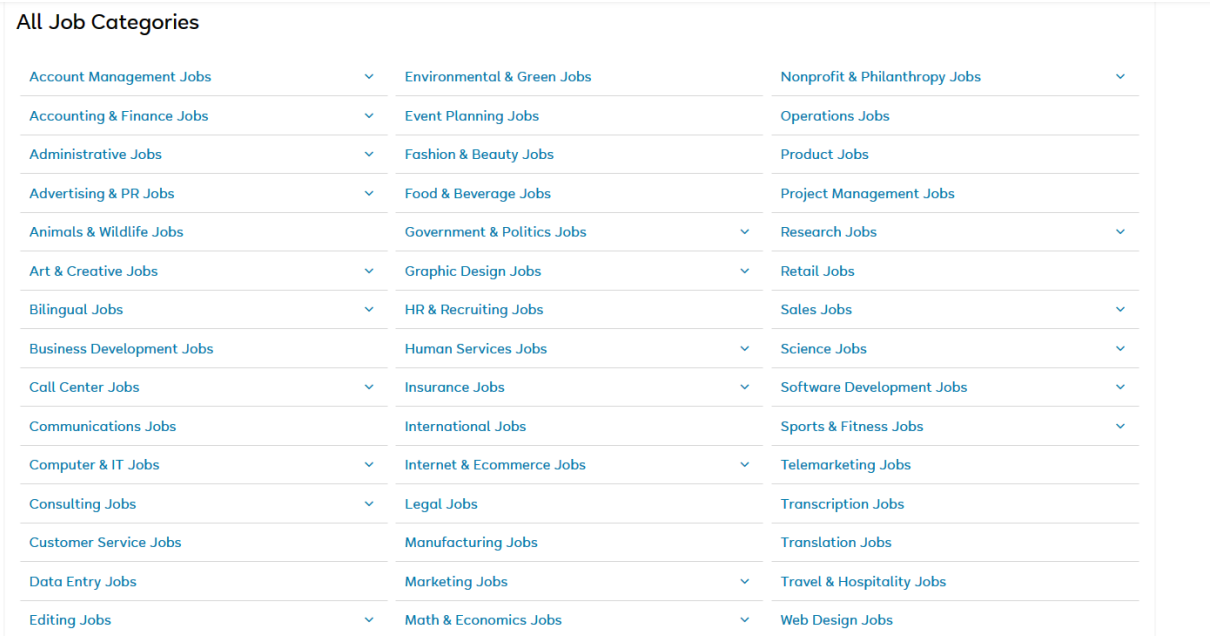
	Stanowisko	Lokalizacja	Opis
0	[Senior JDE E1 Developer]	[Thornton, CO]	[Develop, modify and maintain application syst...
1	[Golang Core Developer]	[Bloomfield, CT]	[Design, develop, and maintain Golang-based mi...
2	[Senior Salesforce Developer]	[Mexico]	[Architect and develop complex technical solut...
3	[Application Developer, Workday Integration]	[Mexico]	[Design, build, deliver, and support technical...
4	[Senior Software Development Engineer, Big Data]	[Mexico]	[Design, implement, and take ownership of miss...
5	[Director, IT Product Management]	[Oakland, CA, Long Beach, CA, El Dorado Hills,...	[Strategically define and design technical sol...
6	[Senior Manager, Machine Learning Engineering]	[US National]	[Lead team of engineers in building scalable i...
7	[Customer Data Platform Technologist]	[US National]	[Play a key role in configuring, integrating, ...
8	[Business Systems Analyst II - Remote]	[Raleigh, NC]	[Perform business systems analysis for Core ba...
9	[Senior Staff Technical Product Manager]	[MA, or Work from Anywhere]	[Responsible for leading, implementing, and ma...
10	[Intune Administrator]	[US National]	[An SCCM/Intune Administrator is needed for a ...
11	[Lead Configuration Management Specialist]	[Chantilly, VA]	[Responsible for developing and maintaining co...
12	[Manager, Infrastructure Core Engineering]	[Dublin, OH]	[Manage a team of Systems Engineers and Admini...
13	[Analytics Specialist - Claims Litigation - Le...	[Madison, WI]	[Analyze business results and external market ...
14	[Data Engineer II]	[Dublin, OH]	[Design and build new data integrations. Work ...
15	[Lead Data Engineer]	[Dublin, OH]	[Lead and deliver enterprise projects for the ...

Strona Flexjobs jest o wiele bardziej rozbudowana pod względem różnorodności dostępnych ofert pracy. Ja wybrałem zawody z branży IT jak widać po kodzie URL. Możemy jednak łatwo to zmienić przechodząc na stronie do zakładki „Find Remote, Hybrid & Flexible Jobs”



Aby następnie wybrać kierunek, który nas interesuje.

ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl



All Job Categories

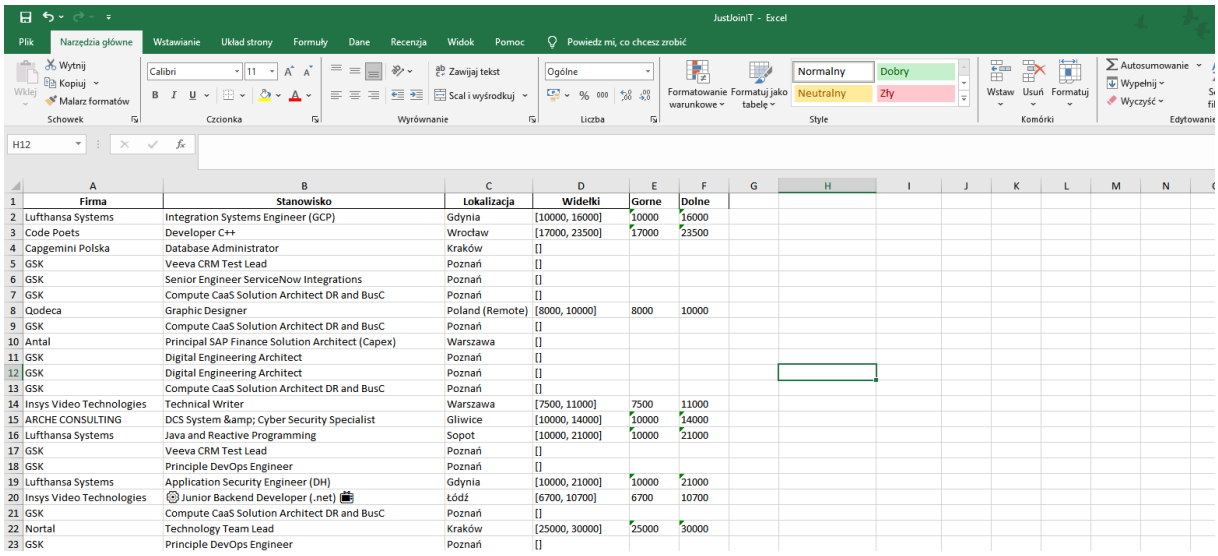
Account Management Jobs	Environmental & Green Jobs	Nonprofit & Philanthropy Jobs
Accounting & Finance Jobs	Event Planning Jobs	Operations Jobs
Administrative Jobs	Fashion & Beauty Jobs	Product Jobs
Advertising & PR Jobs	Food & Beverage Jobs	Project Management Jobs
Animals & Wildlife Jobs	Government & Politics Jobs	Research Jobs
Art & Creative Jobs	Graphic Design Jobs	Retail Jobs
Bilingual Jobs	HR & Recruiting Jobs	Sales Jobs
Business Development Jobs	Human Services Jobs	Science Jobs
Call Center Jobs	Insurance Jobs	Software Development Jobs
Communications Jobs	International Jobs	Sports & Fitness Jobs
Computer & IT Jobs	Internet & Ecommerce Jobs	Telemarketing Jobs
Consulting Jobs	Legal Jobs	Transcription Jobs
Customer Service Jobs	Manufacturing Jobs	Translation Jobs
Data Entry Jobs	Marketing Jobs	Travel & Hospitality Jobs
Editing Jobs	Math & Economics Jobs	Web Design Jobs

Po kliknięciu w wybrany obszar naszego zainteresowania, kopiujemy kod URL i zamieniamy go w naszym programie.

ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl

4. Analiza zebranych danych

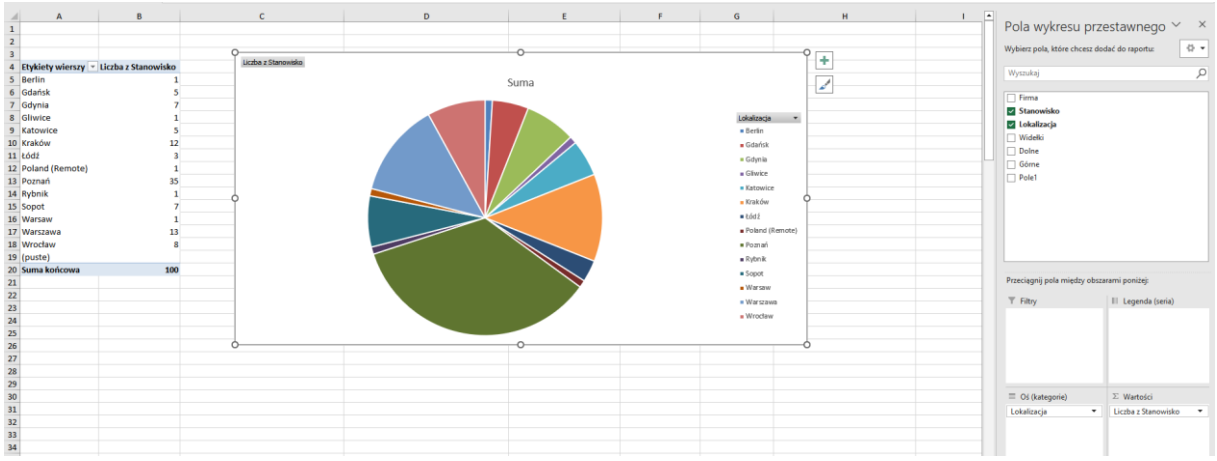
Analizę przeprowadzę na 100 próbkach uzyskanych ze strony JustJoinIT. Oto jak przedstawiają się dane w arkuszu Excel. Widełki podzieliłem na dwie górne i dolne.



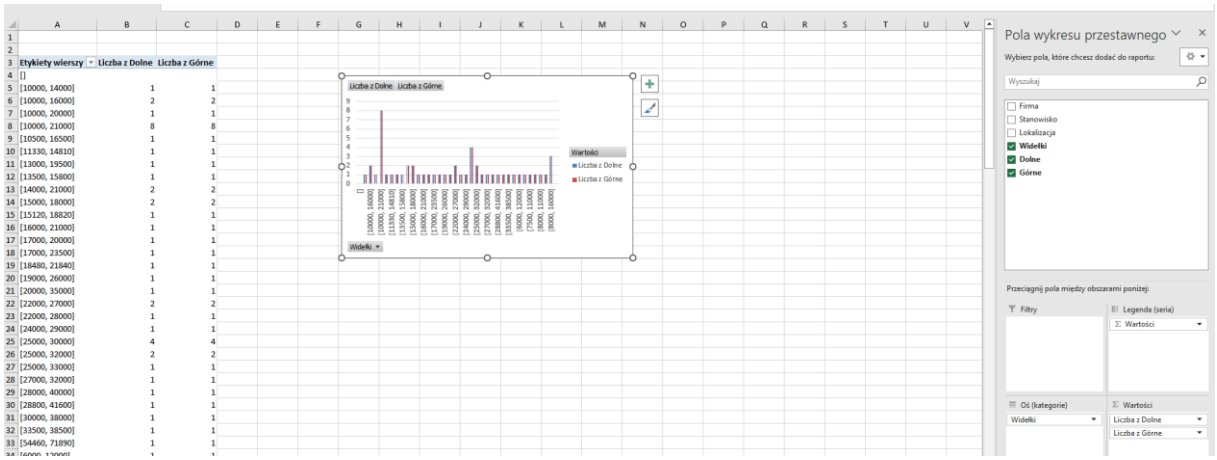
	A	B	C	D	E	F
1	Firma	Stanowisko	Lokalizacja	Widełki	Gorne	Dolne
2	Lufthansa Systems	Integration Systems Engineer (GCP)	Gdynia	[10000, 16000]	10000	16000
3	Code Poets	Developer C++	Wrocław	[17000, 23500]	17000	23500
4	Capgemini Polska	Database Administrator	Kraków	[]		
5	GSK	Veeva CRM Test Lead	Poznań	[]		
6	GSK	Senior Engineer ServiceNow Integrations	Poznań	[]		
7	GSK	Compute CaaS Solution Architect DR and BusC	Poznań	[]		
8	Qodeca	Graphic Designer	Poland (Remote)	[8000, 10000]	8000	10000
9	GSK	Compute CaaS Solution Architect DR and BusC	Poznań	[]		
10	Antal	Principal SAP Finance Solution Architect (Capex)	Warszawa	[]		
11	GSK	Digital Engineering Architect	Poznań	[]		
12	GSK	Digital Engineering Architect	Poznań	[]		
13	GSK	Compute CaaS Solution Architect DR and BusC	Poznań	[]		
14	Insys Video Technologies	Technical Writer	Warszawa	[7500, 11000]	7500	11000
15	ARCHE CONSULTING	DCS System & Cyber Security Specialist	Gliwice	[10000, 14000]	10000	14000
16	Lufthansa Systems	Java and Reactive Programming	Sopot	[10000, 21000]	10000	21000
17	GSK	Veeva CRM Test Lead	Poznań	[]		
18	GSK	Principle DevOps Engineer	Poznań	[]		
19	Lufthansa Systems	Application Security Engineer (DH)	Gdynia	[10000, 21000]	10000	21000
20	Insys Video Technologies	Junior Backend Developer (.net)	Łódź	[6700, 10700]	6700	10700
21	GSK	Compute CaaS Solution Architect DR and BusC	Poznań	[]		
22	Nortal	Technology Team Lead	Kraków	[25000, 30000]	25000	30000
23	GSK	Principle DevOps Engineer	Poznań	[]		

Następnie nasze dane przedstawiłem w postaci tabeli przestawnej (PowerPivot), aby analiza była dobrze widoczna. Na samym początku sprawdzimy z jakiego miasta jest najwięcej ofert.

ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl

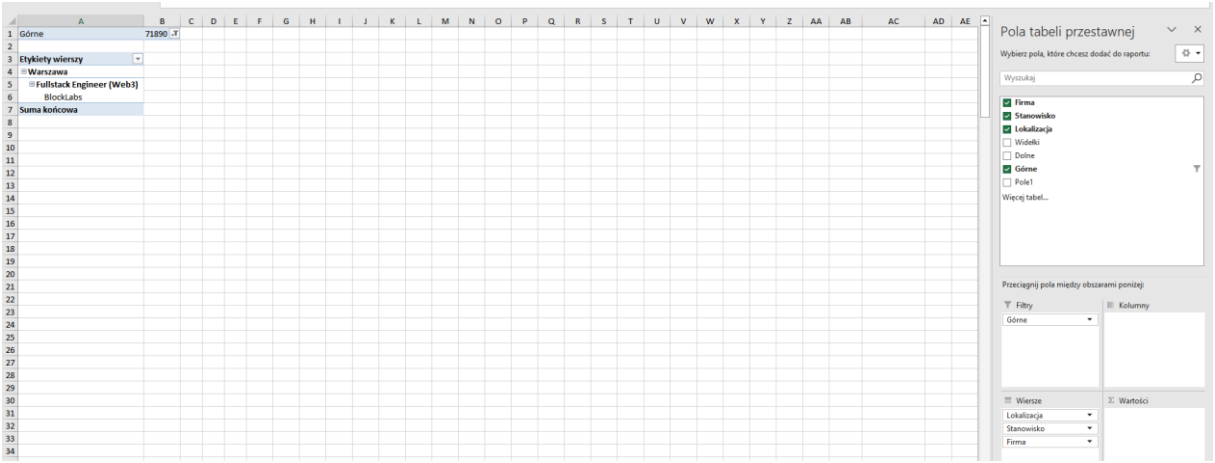


Przy pomocy wykresu kołowego, możemy zauważyć, że najwięcej ofert z aktualnie zescrapowanych pochodzi z Poznania, następna jest Warszawa i na trzecim miejscu Kraków. Popatrzmy na widełki

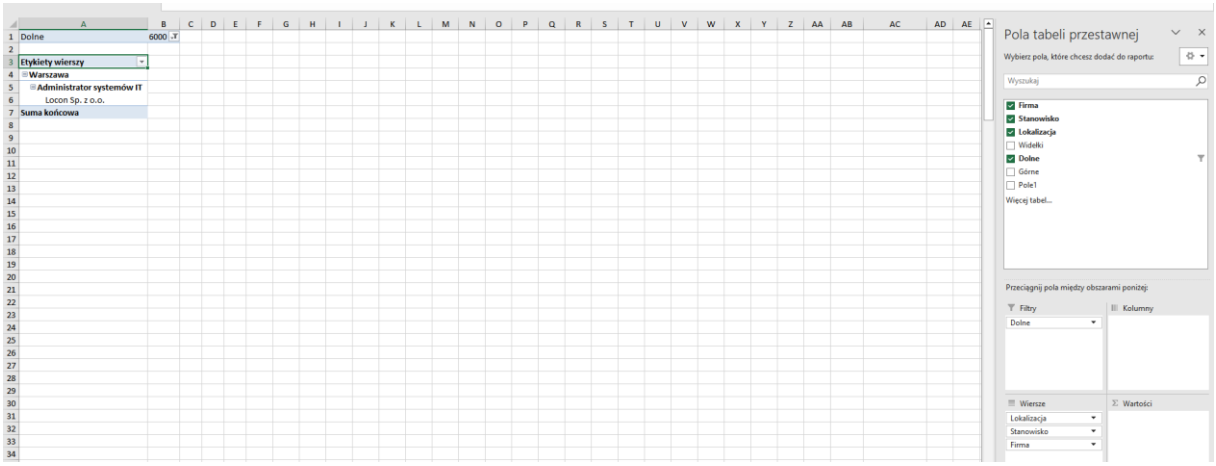


Ofert z widełkami na poziomie 10000 – 21000 PLN jest najwięcej.

ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl



Najwięcej zaś można zarobić w Warszawie. Największa górna granica zarobków znalazła się właśnie w Warszawie i jest na poziomie 71890 PLN miesięcznie na stanowisku FullStack Engineer w firmie BlockLabs .Najmniejsze widełki również zaczynają się w Warszawie na poziomie 6000 PLN na stanowisku Administratora Systemów IT w Locon Sp. z.o.o..



ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl

WYDZIAŁ
ELEKTROTECHNIKI
I INFORMATYKI
POLITECHNIKI RZESZOWSKIEJ

Podsumowanie i wnioski

Projekt zakładał zebranie danych ze stron oferujących zatrudnienie w różnych branżach. Udało mi się zescrapować i przeanalizować zebrane przeze mnie dane z trzech stron. Technologia w jakiej pracowałem o Python oraz Excel.

Po analizie ofert możemy łatwo wywnioskować, że dobrze płatną pracę w IT możemy znaleźć najłatwiej w stolicy w Warszawie oraz największych miastach Polski. Zdecydowałem się na branżę IT z powodu, że mnie najbardziej interesowała. Technologia jaką wykorzystałem niestety nie potrafi scrapować stron, w których zostało użyte bardzo dużo JavaScript, co zawęża zakres stron. Technika scrapowania jest bardzo przydatna, jest używana przez np. Urząd Statystyczny w Rzeszowie, co zostało nam pokazane na laboratoriach z pracownikiem tego urzędu.

ZAKŁAD SYSTEMÓW ZŁOŻONYCH
Wydział Elektrotechniki i Informatyki
ul. Wincentego Pola 2, 35-959 Rzeszów,
tel. 17 865 1340
zsz.prz.edu.pl

Źródła

- regex101: build, test, and debug regex
- https://nofluffjobs.com/
- justjoin.it
- Computer & IT Jobs - Remote Work From Home & Online | FlexJobs
- https://chat.openai.com/