

تطبيق تجريف الويب web scraping على موقع wuzzuf

الدكتور مهند عيسى *

رنا محمد خزامه **

رشا نوفل جزعة**

بهية محمد هواش**

ملخص:

سنقوم بتطبيق *web scraping* على موقع *wuzzuf* لاستخلاص كامل المعلومات المتعلقة بالبحث المطلوب وذلك باستخدام لغة بايثون .



المقدمة:-

أصبح هوس التقنية وتغير معطياتها بتسارع رهيب أحد سمات العصر الذي نعيشه ، كما تزايد الشغف لابتكار آليات تسهل الغوص في هذا العالم التقني بالغ الدقة والتجدد.

في لحظة قرأنا لأي مقال على أي موقع ويب فإننا نساهم في ترويج المحتوى الرقمي وتداوله على نطاق واسع ، ولا شك أننا سنصادف ملايين الروابط التي تتضمن صوراً لافقة ومعطيات رقمية قد نرغب بتخزينها في حواسيبنا .



من أبرز التحديات التي ستواجهنا : كيفية تخزين هذه المعلومات

فما هي الطرق التي سننتهجها للوصول إلى تخزين هذه المعطيات بسرعة وسلاسة بطريقة عملية ذكية وآمنة ومن دون التأثير على محتوى الملفات ؟؟

الإجابة على السؤال السابق هي بالضبط المهمة الأساسية التي تقوم بها كاشطات الويب!

مفهوم كشط الويب:

هو طريقة بسيطة لاستخراج كافة البيانات والمعلومات الموجودة بأي موقع متاح في شبكة الانترنت العالمية على شكل صور وبيانات وجداول ... جاهزة للاستعمال بدون الحاجة إلى أكواد معقدة لتحويلها من بيانات على شكل HTML إلى بيانات على شكل Excel أو XML-Csv أو Json .

• الحاجة لاستخدام كاشطات الويب Web Scrapers

قبل أن نغوص في شرح بدائل التخزين المتاحة بدقة واستفاضة لابد أن نستعرض مجموعة من الآليات المتاحة للحفظ والتخزين والتي أثبتت محدوديتها واحتواءها على عدد كبير من النقائص وأهمها : النسخ واللصق ، ميزة لقطة الشاشة ScreenShot ، وغيرها ..

في الواقع كل هذه الطرق تقليدية وغير موجهة لمحترفي الويب ، إضافة إلى كون بعض المعلومات غير قابلة للنسخ واللصق دون تشويه للهيكل العام للبيانات ، نحن نحتاج بلا شك إلى طريقة معاصرة كتقنية تتيح لك استخراج المعلومات كما هي منشورة بالضبط في الموقع المختار ، نحن نتحدث الآن عن تقنية web scraping والتي يقابلها مصطلح "تجريف الويب" بلغتنا العربية.

آلية عمل كاشطات الويب Web Scrapers

يمكن لكاشط الويب استخراج جميع البيانات الموجودة في موقع معين أو البيانات المحددة التي يريدها المستخدم ، من الأفضل تحديد البيانات التي نريدها بحيث يستخرج كاشط الويب البيانات المرغوبة فقط وبسرعة .

يتم أولاً توفير عناوين URL لمواقع الويب المستهدفة ثم يقوم بتحميل جميع أكواد HTML للموقع بهدف نسخ كافة المعلومات من

السيرفرات التي تخزن هذا الموقع و إذا كان كاشط الويب أكثر تقدما فيمكن له أن يستخرج ملفات CSS و JAVASCRIPT و بالتالي يحصل الكاشط على البيانات المطلوبة وتخرج هذه البيانات بالتنسيق المحدد من قبل المستخدم و غالبا ما يكون في شكل جدول بيانات .

• لماذا تعتبر *Python* لغة شائعة في *Web*

Scraping ؟

يبدو أن بايثون هي الموضوعة في لغات البرمجة هذه الأيام !
إنها اللغة الأكثر شيوعا لكشط الويب حيث يمكنها التعامل مع معظم العمليات بسهولة كما أن لديها مجموعة من المكتبات التي تم إنشاؤها خصيصا لكشط الويب .

SCRAPY هو إطار عمل كشط ويب مفتوح المصدر شائع جدا ومكتوب بلغة بايثون ويعتبر مثاليا لكشط الويب و استخراج البيانات باستخدام واجهة برمجة التطبيقات .

BAEUTIFUL SOUP هو طريقة أخرى مناسبة جدا لكشط الويب و أيضا باستخدام لغة بايثون التي تقوم بإنشاء شجرة تحليل يمكن استخدامها لاستخراج البيانات من على موقع ويب يحتوي BEAUTIFUL SOUP على ميزات متعددة للتنقل والبحث وتعديل أشجار التحليل هذه.

بعض استخدامات Web Scraping

كشط الويب له تطبيقات متعددة في مختلف الصناعات والقطاعات:

1-مراقبة الأسعار :

يمكن للشركات استخدام تقنية كشط الويب لسحب بيانات المنتج الخاصة بمنتجاتها والمنتجات المنافسة أيضا لمعرفة مدى تأثيرها على استراتيجيات التسعير الخاصة بها
ومن ثم استخدام هذه البيانات لاصلاح الأسعار والحصول على اقصى عائد.

2-أبحاث السوق :

يمكن استخدام تجريف الويب لاجبات السوق من قبل الشركات ويمكن ان تكون البيانات المجمعة والتي تم الحصول عليها بأحجام كبيرة مفيدة جدا بالشركات في تحليل اتجاهات المستهلكين وفهم الاتجاه الذي يجب أن تتحرك فيه الشركة في المستقبل.

3-مراقبة الأخبار:

يمكن لمواقع كشط الأخبار على الويب تقديم تقارير مفصلة عن الأخبار الحالية للشركة ويعد هذا أكثر أهمية بالنسبة للشركات التي تظهر بشكل متكرر على الأخبار أو التي تعتمد على الأخبار اليومية لأداء عملها.

4-تحليل المشاعر :

إذا أرادت الشركات فهم الشعور العام اتجاه منتجاتها بين المستهلكين فإن تحليل المشاعر أمر لا بد منه ، حيث يمكن للشركات استخدام تجريف الويب لجمع البيانات من مواقع التواصل الاجتماعي مثل فيسبوك وتويتر لمعرفة المشاعر العامة حول منتجاتها .

5-التسويق عبر البريد الإلكتروني :

يمكن للشركات أيضا استخدام تجريف الويب لجمع معرفات البريد الإلكتروني من مواقع مختلفة ثم ارسال رسائل ترويجية وتسويقية مجمعة الى جميع الأشخاص الذين يملكون معرفات البريد الإلكتروني هذه.

تطبيق ال web scraping على موقع wuzzuf:

سنقوم بتطبيق تجريف الويب على مثال البحث عن عمل وذلك عن طريق الدخول إلى موقع الويب wuzzuf ومحاولة إيجاد عمل في مجال البرمجة بلغة بايثون .

1-الخطوة الأولى:

نقوم بتنزيل المكتبات الآتية:

beautiful soup ,lxml,requests

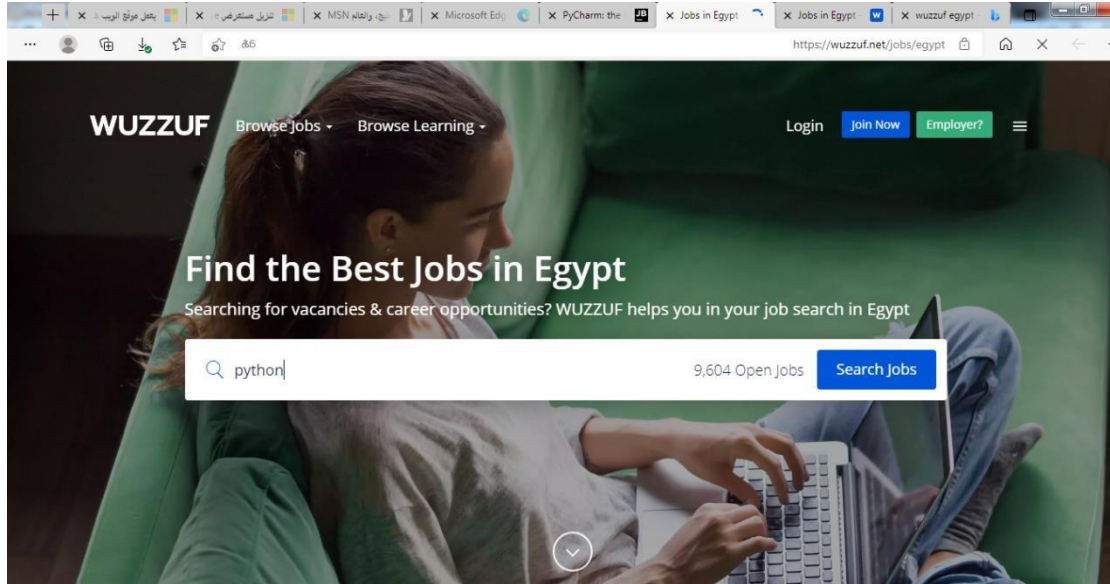
```
Microsoft Windows [Version 10.0.10240]
(c) 2015 Microsoft Corporation. All rights reserved.

C:\Users\B\>pip list
Package Version
-----
beautifulsoup4 4.11.1
bs4 0.0.1
certifi 2022.5.18.1
charset-normalizer 2.0.12
idna 3.3
lxml 4.9.0
pip 22.1.2
requests 2.28.0
setuptools 41.2.0
soupsieve 2.3.2.post1
urllib3 1.26.9

C:\Users\B\>
```

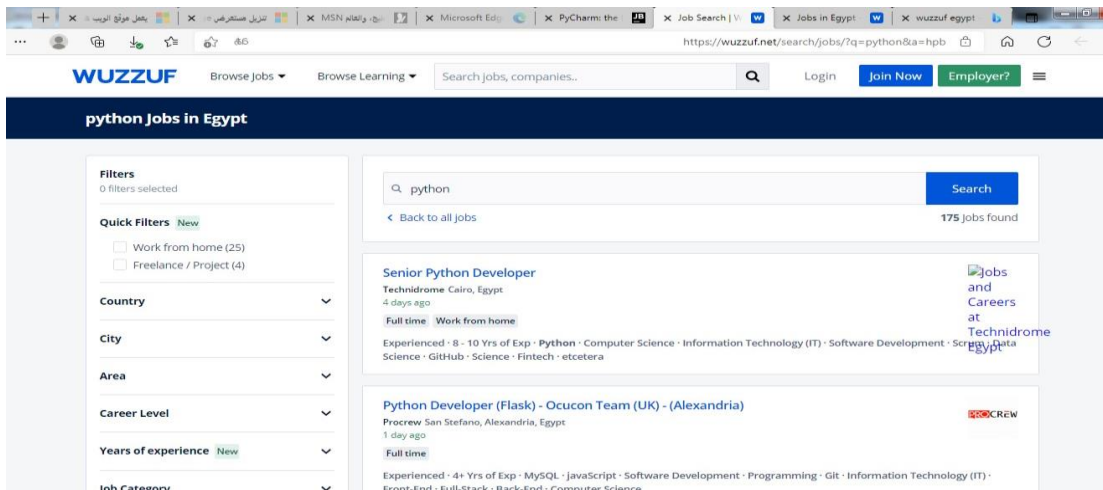

2-الخطوة الثانية :-

ندخل إلى الموقع ونكتب بالبحث python



عند الدخول الى الموقع نرى ضمنه عدة أقسام وكل قسم يتعلق بوظيفة معينة.

مثلا اول وظيفة عنوانها senior python Developer موجود فيها معلومات عن اسم الشركة ومكانها وتاريخ وضع الوظيفة على الموقع والخبرات والمهارات المطلوبة من أجل الوظيفة



*دكتور في قسم هندسة الاتصالات والالكترونيات-جامعة تشرين.
**طالبات في قسم هندسة الاتصالات والالكترونيات –السنة الخامسة.

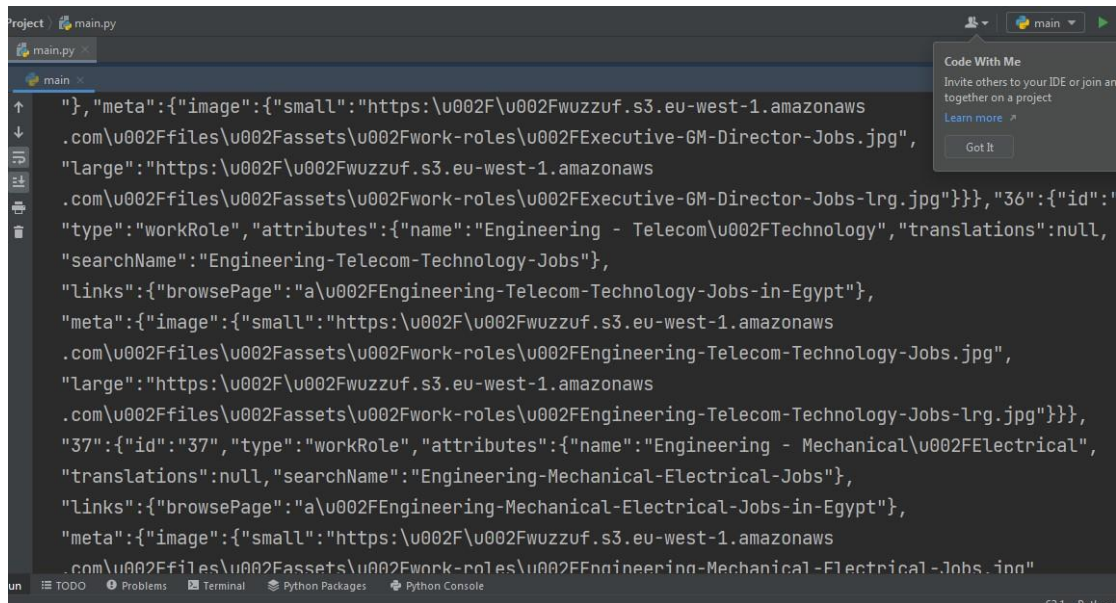
كل هذه المعلومات سنقوم بتجميعها ضمن ال csv file ونحفظه ضمن ملف بحيث يكون لدينا مكان نقرأ منه المعلومات المهمة بالنسبة لنا في هذا الملف بدلا من الدخول الى كل قسم على حدى. بهذه الطريقة نقوم بتوفير الوقت والجهد

```
main.py x project.py x
1 import requests
2 from bs4 import BeautifulSoup
3 import csv
4 from itertools import zip_longest
5
```

ثم نقوم بإنشاء متغير اسمه result نخزنه ضمنه رابط الصفحة التي سنستخرج منها المعلومات سننشأ متغير اسمه src يقوم بإعادة محتوى الصفحة

```
result=requests.get("https://wuzzuf.net/search/jobs/?q=python&a=hpb")
src=result.content
```

نقوم بعمل RUN. تظهر لدينا نتيجة تعبر عن محتوى هذه الصفحة



```
{
  "meta": {
    "image": {
      "small": "https://u002F\u002Fwuzzuf.s3.eu-west-1.amazonaws.com\u002Ffiles\u002Fassets\u002Fwork-roles\u002FExecutive-GM-Director-Jobs.jpg",
      "large": "https://u002F\u002Fwuzzuf.s3.eu-west-1.amazonaws.com\u002Ffiles\u002Fassets\u002Fwork-roles\u002FExecutive-GM-Director-Jobs-lrg.jpg"
    }
  },
  "36": {
    "id": "36",
    "type": "workRole",
    "attributes": {
      "name": "Engineering - Telecom\u002FTechnology",
      "translations": null,
      "searchName": "Engineering-Telecom-Technology-Jobs"
    },
    "links": {
      "browsePage": "a\u002FEngineering-Telecom-Technology-Jobs-in-Egypt"
    },
    "meta": {
      "image": {
        "small": "https://u002F\u002Fwuzzuf.s3.eu-west-1.amazonaws.com\u002Ffiles\u002Fassets\u002Fwork-roles\u002FEngineering-Telecom-Technology-Jobs.jpg",
        "large": "https://u002F\u002Fwuzzuf.s3.eu-west-1.amazonaws.com\u002Ffiles\u002Fassets\u002Fwork-roles\u002FEngineering-Telecom-Technology-Jobs-lrg.jpg"
      }
    }
  },
  "37": {
    "id": "37",
    "type": "workRole",
    "attributes": {
      "name": "Engineering - Mechanical\u002FElectrical",
      "translations": null,
      "searchName": "Engineering-Mechanical-Electrical-Jobs"
    },
    "links": {
      "browsePage": "a\u002FEngineering-Mechanical-Electrical-Jobs-in-Egypt"
    },
    "meta": {
      "image": {
        "small": "https://u002F\u002Fwuzzuf.s3.eu-west-1.amazonaws.com\u002Ffiles\u002Fassets\u002Fwork-roles\u002FEngineering-Mechanical-Electrical-Jobs.jpg",
        "large": "https://u002F\u002Fwuzzuf.s3.eu-west-1.amazonaws.com\u002Ffiles\u002Fassets\u002Fwork-roles\u002FEngineering-Mechanical-Electrical-Jobs-lrg.jpg"
      }
    }
  }
}
```

سنستخرج المعلومات من هذه الصفحة من خلال استدعاء مكتبة Beautiful Soup ونعطيها بارامترين هما المتغير src الذي يعبر عن محتوى الصفحة و "lxml" الذي قمنا بتنزيله وهو يساعد في القيام بعمليات ومعالجة الصفحة

ونقوم بوضعها ضمن المتغير soup

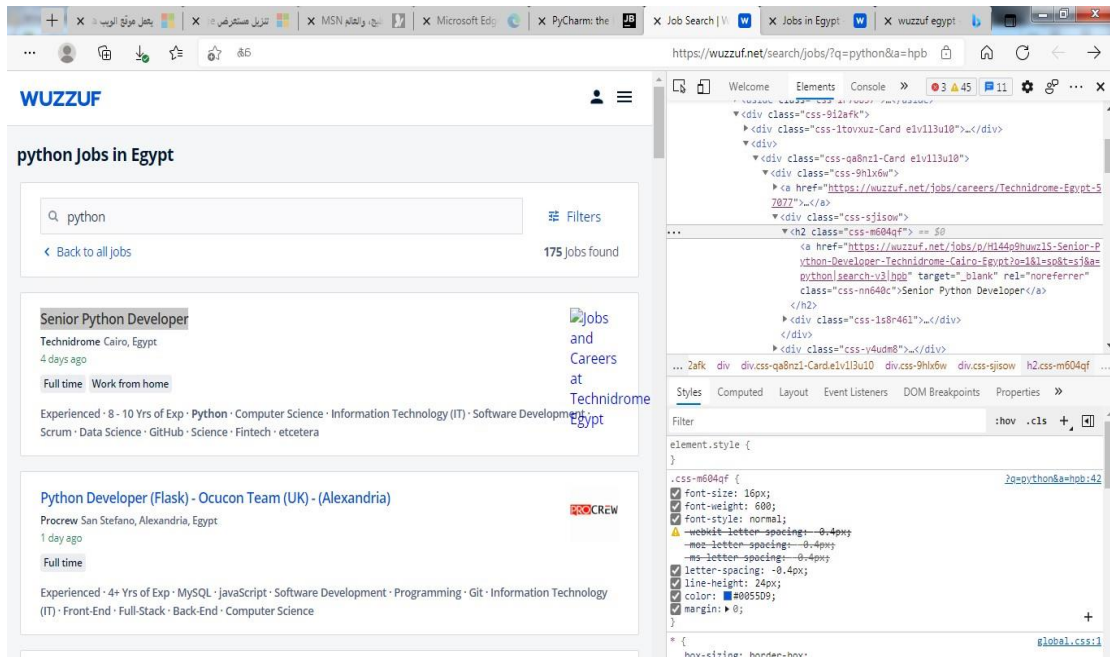
```
soup = BeautifulSoup(src, "lxml")
print(soup)
```

الان نريد استخراج المعلومات التي نحتاجها (عنوان الوظيفة واسم الشركة وموقعها والمهارات المطلوبة)

المعلومات المطلوبة موجودة ضمن HTML المكون من html tags ،اي:

في حال نريد استخراج عنوان الوظيفة نحدد العنوان ثم بالزر اليمين نختار فحص عنصر فيظهر لدينا ال developer tools الذي يحوي معلومات الصفحة التي نبحث عنها

المتغير job_titles ندخل فيه اول بارمتر وهو عنصر ال html الذي ظهر معنا وثاني بارمتر وصف link tags الذي يحوي عنوان الوظيفة class له قيمة نحصل عليها من الصفحة



*دكتور في قسم هندسة الاتصالات والالكترونيات-جامعة تشرين.
**طالبات في قسم هندسة الاتصالات والالكترونيات –السنة الخامسة.

ثم نكرر الأمر نفسه بالنسبة لاستخراج اسماء الشركات ومواقعها والمهارات.

```
job_titles=soup.find_all("h2",{ "class": "css-m604qf"})
company_names=soup.find_all("a",{ "class": "css-17s97q8"})
locations_names=soup.find_all("span",{ "class": "css-5wy50k"})
job_skills=soup.find_all("div",{ "class": "css-y4udm8"})
```

سنعرف بداية قوائم فارغة أخرى كي نضع ضمن كل منها القيم التي نحتاجها

```
job_title = []
company_name = []
location_name = []
skills = []
```

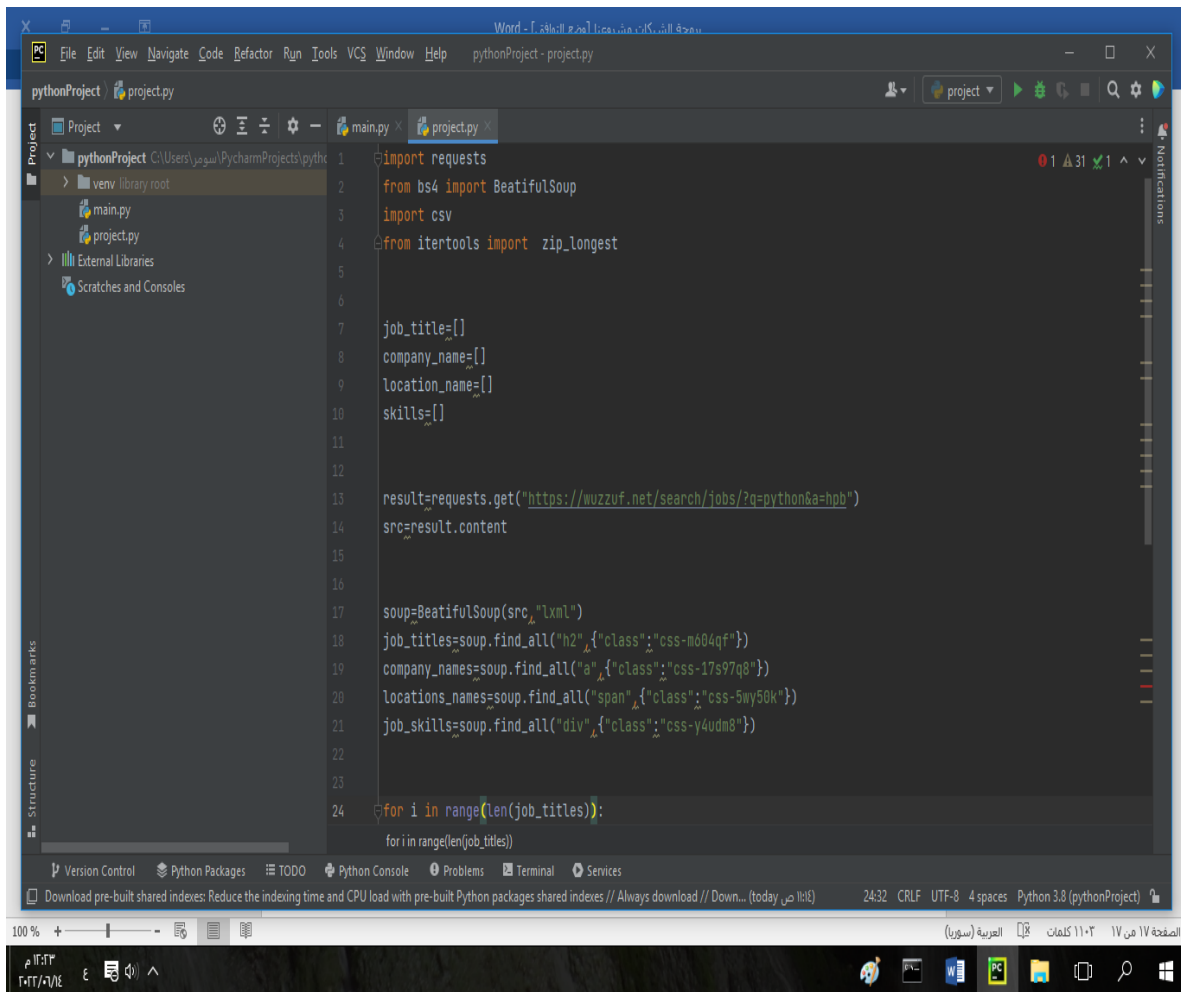
ضمن الحلقة نستخرج من كل قائمة (اسم الوظيفة ،اسم الشركة ،موقعها ،المهارات) النص الموجود ضمنها ونضعه ضمن متغير `job_titles` نخزن ضمنه كل القيم التي حصلنا عليها من القائمة `job_titles` التي نريد اضافتها إلى القائمة الفارغة `job_title` ونكرر الأمر نفسه بالنسبة لأسماء ومواقع الشركات والمهارات

```
for i in range(len(job_titles)):
    job_title.append(job_titles[i].text)
    company_name.append(company_names[i].text)
    location_name.append(location_names[i].text)
    skills.append(job_skills[i].text)
```

سننشأ ملف ال csv لنخزن ضمنه المعلومات التي استخرجناها

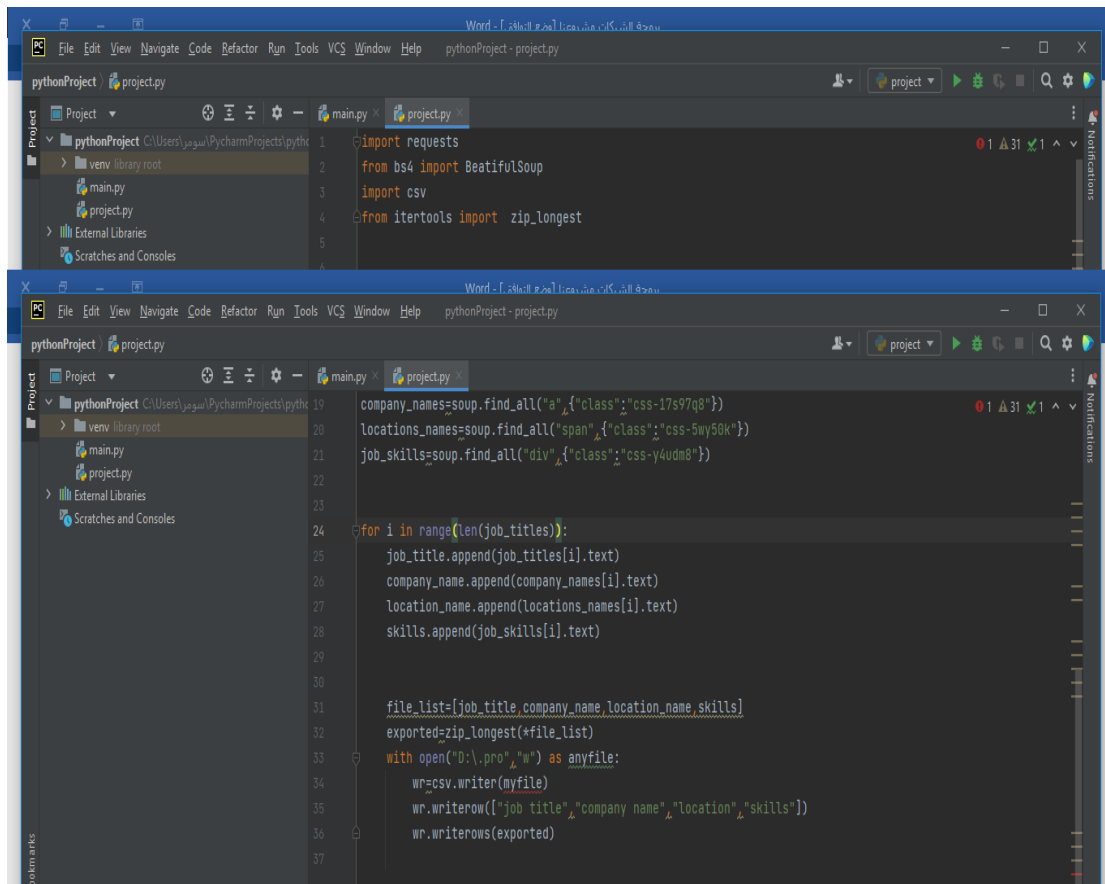
```
with open("D:\.pro", "w") as anyfile:
```

فيكون الكود النهائي هو :



```
1 import requests
2 from bs4 import BeautifulSoup
3 import csv
4 from itertools import zip_longest
5
6
7 job_titles=[]
8 company_name=[]
9 location_name=[]
10 skills=[]
11
12
13 result=requests.get("https://wuzzuf.net/search/jobs/?q=python&a=hpb")
14 src=result.content
15
16
17 soup=BeautifulSoup(src,"xml")
18 job_titles=soup.find_all("h2",{ "class": "css-m604qf"})
19 company_names=soup.find_all("a",{ "class": "css-17s97q8"})
20 locations_names=soup.find_all("span",{ "class": "css-5my50k"})
21 job_skills=soup.find_all("div",{ "class": "css-y4udm8"})
22
23
24 for i in range(len(job_titles)):
25     for i in range(len(job_titles))
```

*دكتور في قسم هندسة الاتصالات والالكترونيات-جامعة تشرين.
**طالبات في قسم هندسة الاتصالات والالكترونيات –السنة الخامسة.



عند تنفيذ الكود لم يكن هناك خطأ برمجي وحصلنا على النتيجة التالية:



ولكن لم تظهر النتائج ضمن ملف ال excel...

*دكتور في قسم هندسة الاتصالات والالكترونيات-جامعة تشرين.
 **طالبات في قسم هندسة الاتصالات والالكترونيات –السنة الخامسة.