**Appreciation**

At the end of this report, I would like to express our deep respect and gratitude to the management of our institute: Higher Institute of Technological Studies of Kasserine who offered us the chance to carry out this internship in order to enrich our theoretical knowledge in SMARTICA SOLUTIONS through practical experience and discovering the professional environment.

Likewise, we extend our sincere thanks to our supervisors, Mrs. Manel Aloui, who helped us enormously and devoted a lot of assistance and time to carry out this work.

**Summary**

[Introduction General 1](#_Toc99557993)

[Chapter 1: Presentation of the company 2](#_Toc99557994)

[I- Presentation of the Smartica solutions company 2](#_Toc99557995)

[1- Identification of the company: 2](#_Toc99557996)

[II- Conclusion 2](#_Toc99557997)

[Chapter 2: Bibliography 3](#_Toc99557998)

[I- Web scraping: 3](#_Toc99557999)

[1- Definition : 3](#_Toc99558000)

[2- Beautiful Soup: 3](#_Toc99558001)

[3- Scrapy : 4](#_Toc99558002)

[II- Differences between scrapy and beautiful 5](#_Toc99558003)

[III- Top Web scraping libraries 6](#_Toc99558004)

[IV- Top used languages for web scraping 7](#_Toc99558005)

[1- Python : 7](#_Toc99558006)

[2- Node.js / js : 7](#_Toc99558007)

[3- Ruby : 8](#_Toc99558008)

[4- C++/C : 8](#_Toc99558009)

[5- PHP : 8](#_Toc99558010)

[V- Conclusion: 9](#_Toc99558011)

[Chapter 3: Realization 10](#_Toc99558012)

[I- Introduction : 10](#_Toc99558013)

[II- Beautiful soup : 10](#_Toc99558014)

[1- Implementation : 10](#_Toc99558015)

[2- Result : 10](#_Toc99558016)

[III- Scrapy : 11](#_Toc99558017)

[1- Implementation : 11](#_Toc99558018)

[2- Result 11](#_Toc99558019)

[IV- Conclusion 11](#_Toc99558020)

[Chapter 4: Methodological 12](#_Toc99558021)

[I- Introduction 12](#_Toc99558022)

[II- Git and GitHub: 12](#_Toc99558023)

[1- What is the Git: 12](#_Toc99558024)

[2- What is the GitHub: 12](#_Toc99558025)

[3- The relationship between them: 12](#_Toc99558026)

[III- The basic command-line of Git: 13](#_Toc99558027)

[IV- Result 13](#_Toc99558028)

[V- Conclusion 13](#_Toc99558029)

[General conclusion 14](#_Toc99558030)

**List of Figures**

[Figure 1 : Web Scraping 3](#_Toc99558031)

[Figure 2 : Library BeautifulSoup 4](#_Toc99558032)

[Figure 3 : Library Scrapy 4](#_Toc99558033)

[Figure 4 : Web scraping languages for web scraping 7](#_Toc99558034)

[Figure 5 : Python 7](#_Toc99558035)

[Figure 6 : Node.js 7](#_Toc99558036)

[Figure 7 : Ruby 8](#_Toc99558037)

[Figure 8 : C++ 8](#_Toc99558038)

[Figure 9 : PHP 8](#_Toc99558039)

[Figure 10 : Result 10](#_Toc99558040)

[Figure 11 : Covid \_data.csv 11](#_Toc99558041)

**List of Tables**

[Tableau 1 : Differences between scrapy and beautiful 5](#_Toc99558042)

[Tableau 2 : Top Web scraping libraries 6](#_Toc99558043)

[Tableau 3 : Command line of get 13](#_Toc99558044)

**The abbreviations**

**API:** Application Programming Interface.

**Cmd:**command prompt.

**Csv:** Comma-Separated Values.

**Curl:** Client URL.

**FTP:** File Transfert Protocol.

**HTML:**Hypertext markup language.

**HTTP:**  HyperText Transfer **Protocol.**

**IT:**  Information Technology.

**ISET :**institut Supérieur des Etudes Technologiques.

**Js:** JavaScript.

**TD:**table data.

**URL:**Uniform Resource Locator.

**Vs:** visual studio**.**

**Xml:**Extensible Markup Language file.

# Introduction General

As part of my second internship, advanced training within the company "Smartica solutions" from January 17th to February 13th, 2022.

It is a period of practical and theoretical lessons which helps me to integrate into professional life and to see the progress of the work.

Indeed, this internship also allows me to apply my knowledge acquired during my two years of university.

In this internship, we will study web scraping, its different frameworks and libraries and practice our theoretical studies with real world projects.

For this we have written this internship report to explain this experience.

In this report, we first present a presentation of the company, then explain the different aspects of the work during these few weeks and finally we summarize the contributions of this internship.

# Chapter 1: Presentation of the company

## Presentation of the Smartica solutions company

In this first chapter we have tried to develop a general presentation of the Smartica solutions company by citing its organization chart

### Identification of the company:

Smartica Solutions is an IT company founded in 2019 by engineers graduated from top universities.

It specializes in digital transformation solutions dedicated to business.

Its main services are:

Website development

Mobile app development

social media marketing

Web scraping

Chatbot development

Machine learning models development

## Conclusion

In this first chapter I tried to make a general presentation of the company Smartica Solution by citing our team, our services... Then I will pass to the 2nd chapter which describes the internship

# Chapter 2: Bibliography

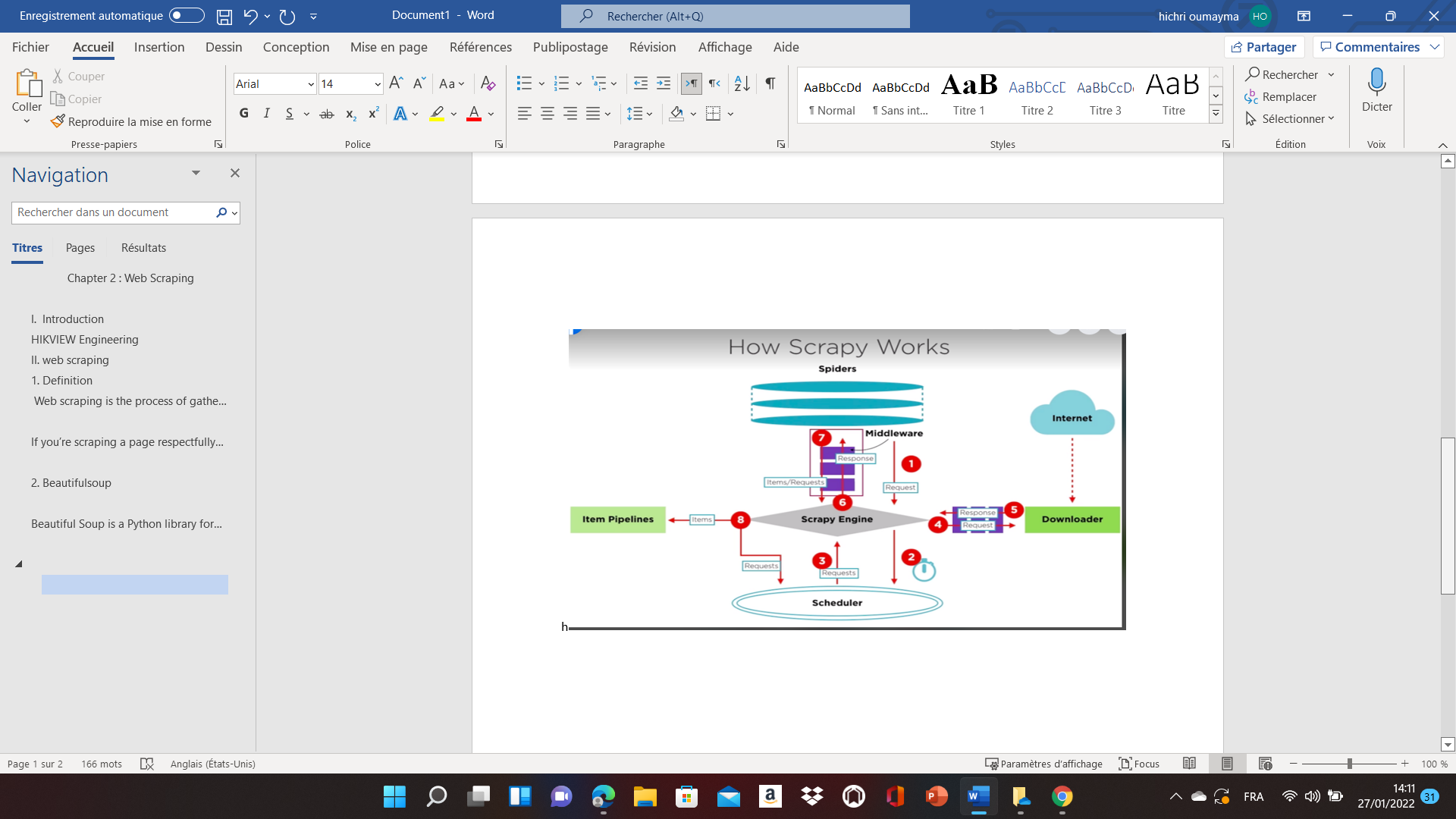
## Web scraping:

### Definition :

Web scraping is the process of gathering information from the Internet. Even copying and pasting the lyrics of your favorite song is a form of web scraping! However, the words “web scraping” usually refer to a process that involves automation. Some websites don’t like it when automatic scrapers gather their data, while others don’t mind.

If you’re scraping a page respectfully for educational purposes, then you’re unlikely to have any problems. Still, it’s a good idea to do some research on your own and make sure that you’re not violating any Terms of Service before you start a large-scale project

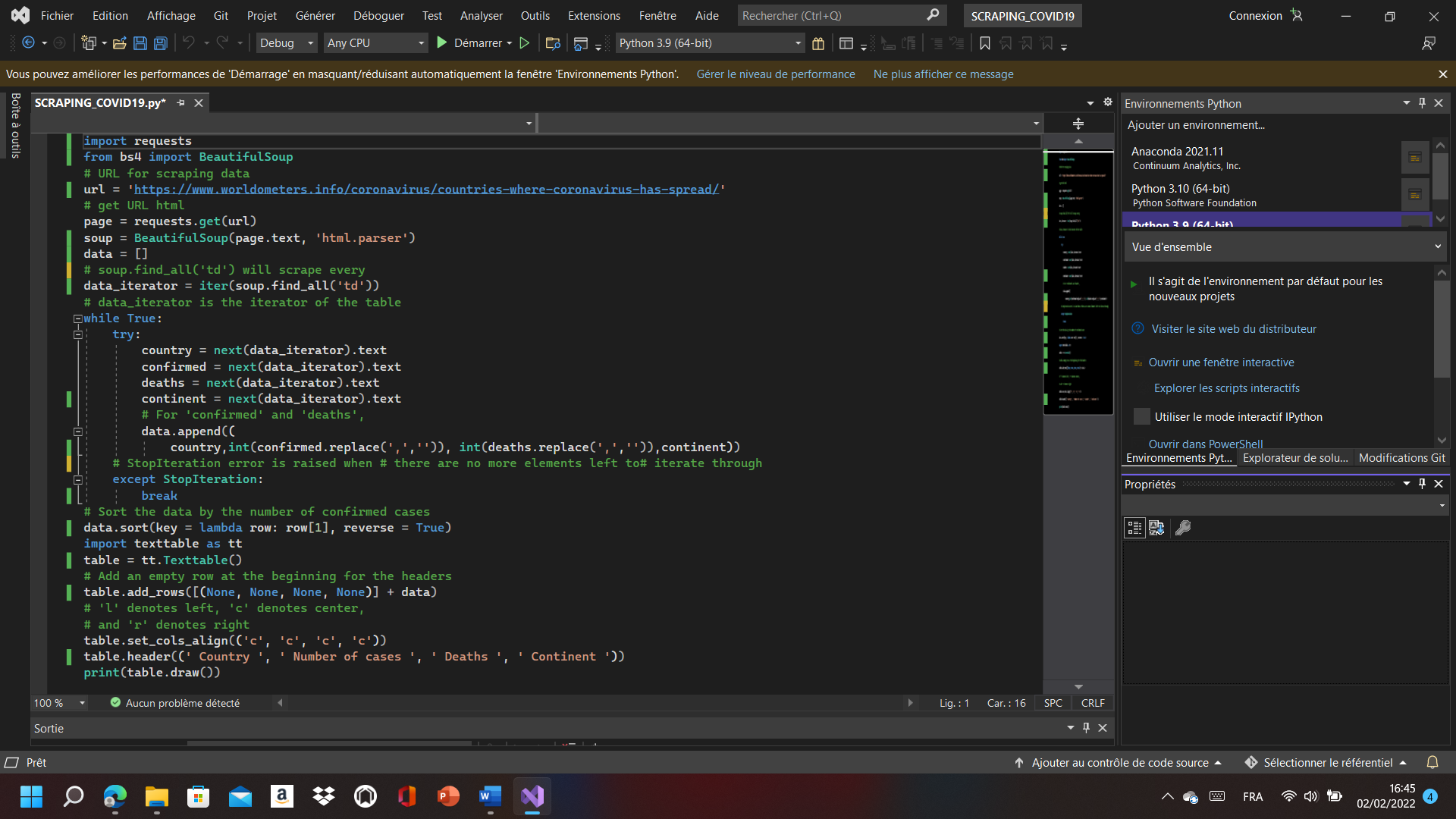
Figure 1 : Web Scraping



### Beautiful Soup:

Beautiful Soup is a Python library for parsing structured data. It allows you to interact with HTML in a similar way to how you interact with a web page using developer tools. The library exposes a couple of intuitive functions you can use to explore the HTML you received

Figure 2 : Library BeautifulSoup

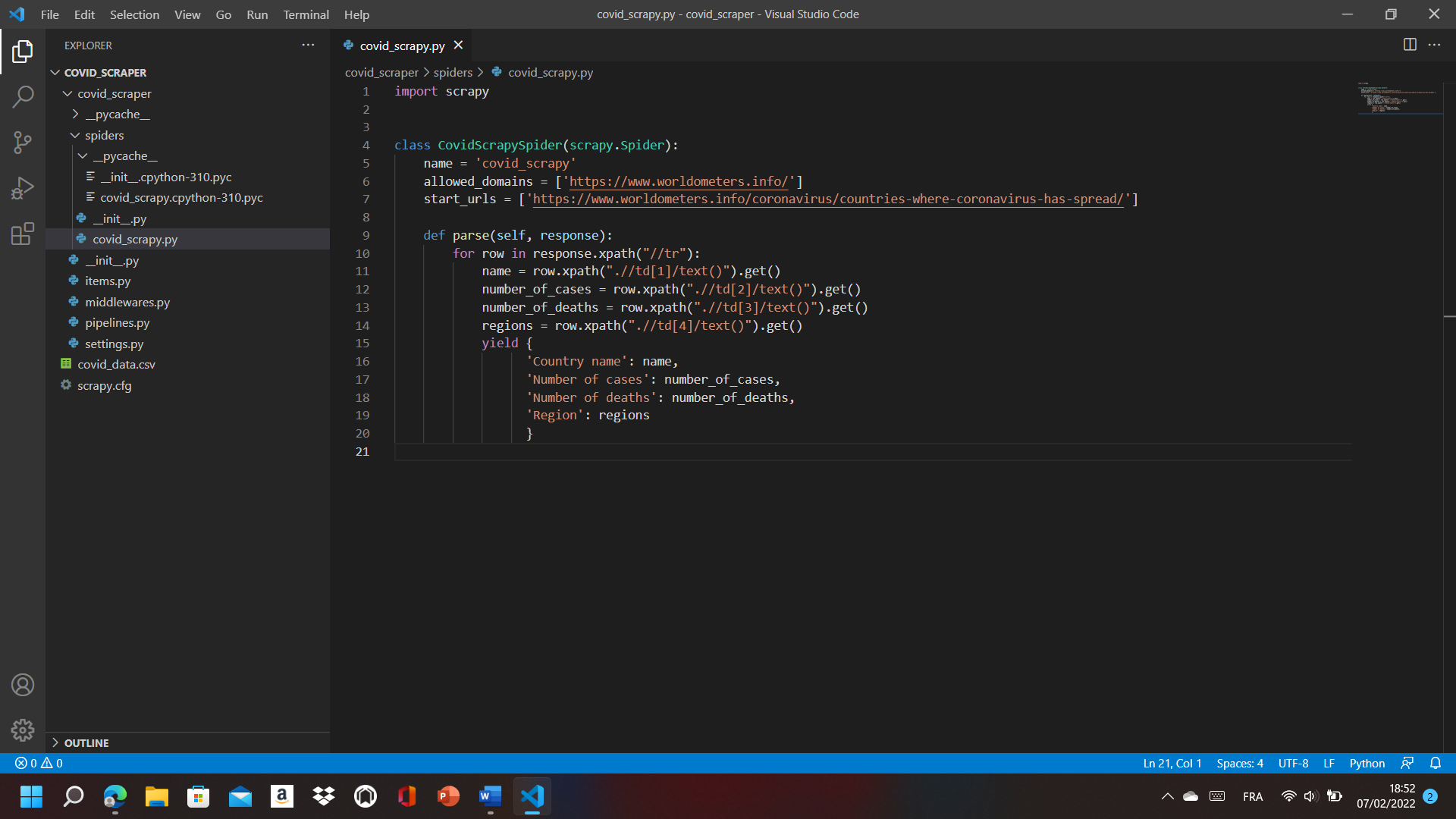


### Scrapy :

Scrapy is a free and open-source web-crawling framework written in Python. Originally designed for web scraping, it can also be used to extract data using APIs or as a general-purpose web crawler.

The Scrapy framework provides you with powerful features allowing you to scrape virtually undetected across the net. Scrapy also provides a web-crawling shell, which can be used by developers to test their assumptions on a site’s behavior

Figure 3 : Library Scrapy



## Differences between scrapy and beautiful

Tableau 1 : Differences between scrapy and beautiful

|  |  |
| --- | --- |
|  | Differences between them |
| structure | Scrapy is an open-source framework, whereas Beautiful Soup is a Python library designed for quick turnaround projects.  A framework inverts the control of the program and informs the developer what they need. Whereas in the case of a library, the developer calls the library where and when they need it. |
| performance | the performance of Scrapy can be said to be faster than Beautiful Soup. Working with Beautiful Soup can speed up with the help of Multithreading process |
| Extensibility | Beautiful Soup works best when working on smaller projects. On the other hand, Scrapy may be the better choice for larger projects with more complexities, as this framework can add custom functionalities and can develop pipelines with flexibility and speed. |
| Beginner-friendly | For a beginner who is trying hands-on web scraping for the first time, Beautiful Soup is the best choice to start with. Scrapy can be used for scraping, but it is comparatively more complex than the former. |
| Community | The developer’s community of Scrapy is stronger and vast compared to that of Beautiful Soup.  Developers can use Beautiful Soup for parsing HTML responses in Scrapy callbacks by feeding the response’s body into a BeautifulSoup object and extracting whatever data they need from it. |

In short, if you have not much experience in programming, the job is a very simple project, then BeautifulSoup can be your choice. If you want a more powerful and flexible web crawler, or you indeed have some experience in programming, then Scrapy is definitely the winner here

## Top Web scraping libraries

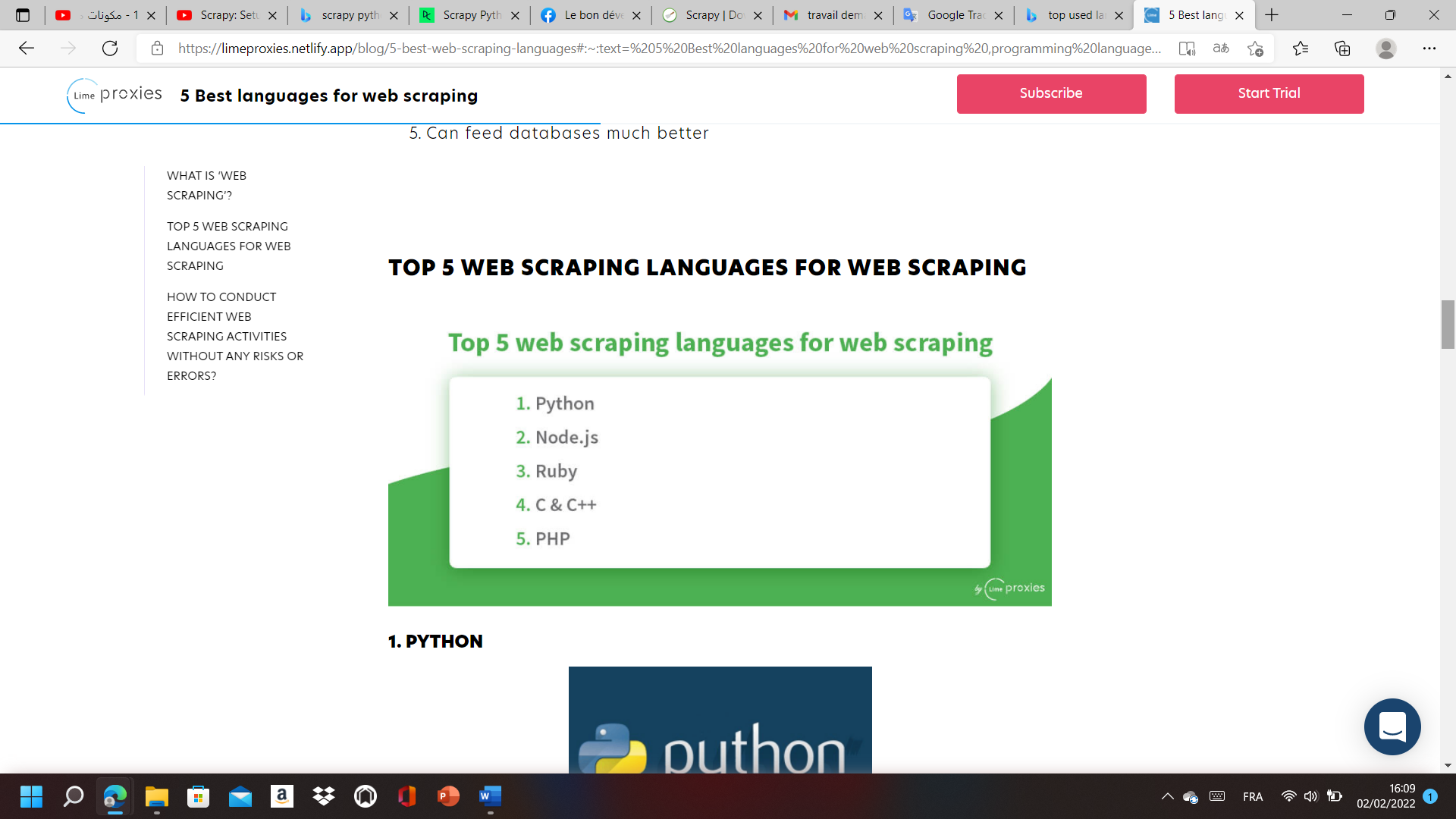
Tableau 2 : Top Web scraping libraries

|  |  |
| --- | --- |
| libraries | Definition |
| Requests | The Requests library is vital to add to your data science toolkit.  It’s a simple yet powerful HTTP library, which means you can use it to access web pages. |
| Beautifulsoup4 | Beautiful Soup (BS4) is a parsing library that can use different parsers. A parser is simply a program that can extract data from HTML and XML documents. |
| lxml | Lxml is a high-performance, production-quality HTML and XML is also a parsing library. |
| selenium | Selenium is a tool that automates browsers, also known as a web-driver. With it, you can actually open a Google Chrome window, visit a site, and click on links. |
| scrapy | Scrapy is technically not even a library… it’s a complete web scraping framework. That means you can use it to manage requests, preserve user sessions, follow redirects, and handle output pipelines |

* There are a whole host of Python libraries available to perform web scraping. But how do you decide which the best one to choose for your particular project, it depends on which Python library holds the most flexibility?

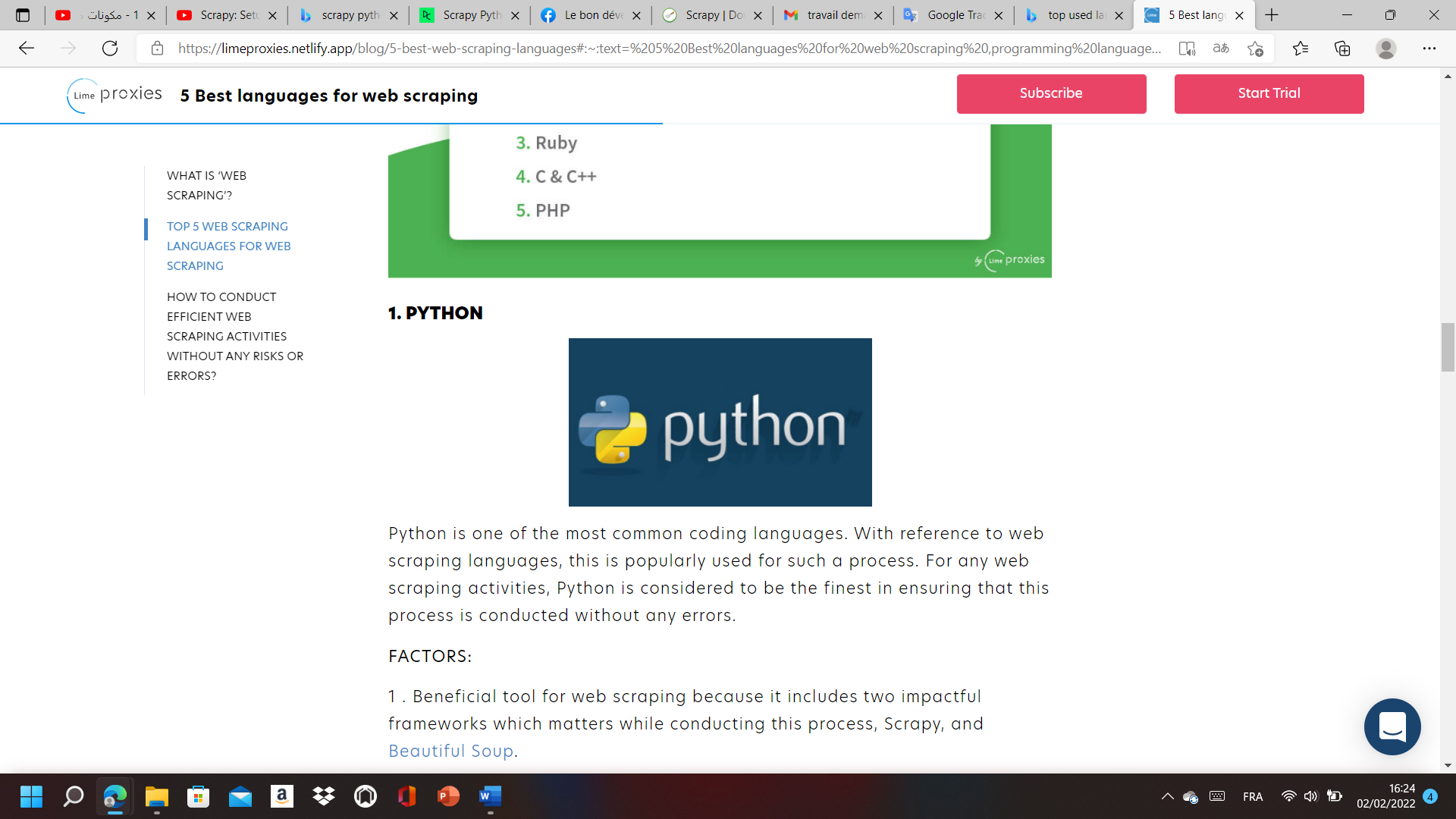
## Top used languages for web scraping

Figure 4 : Web scraping languages for web scraping



### Python :

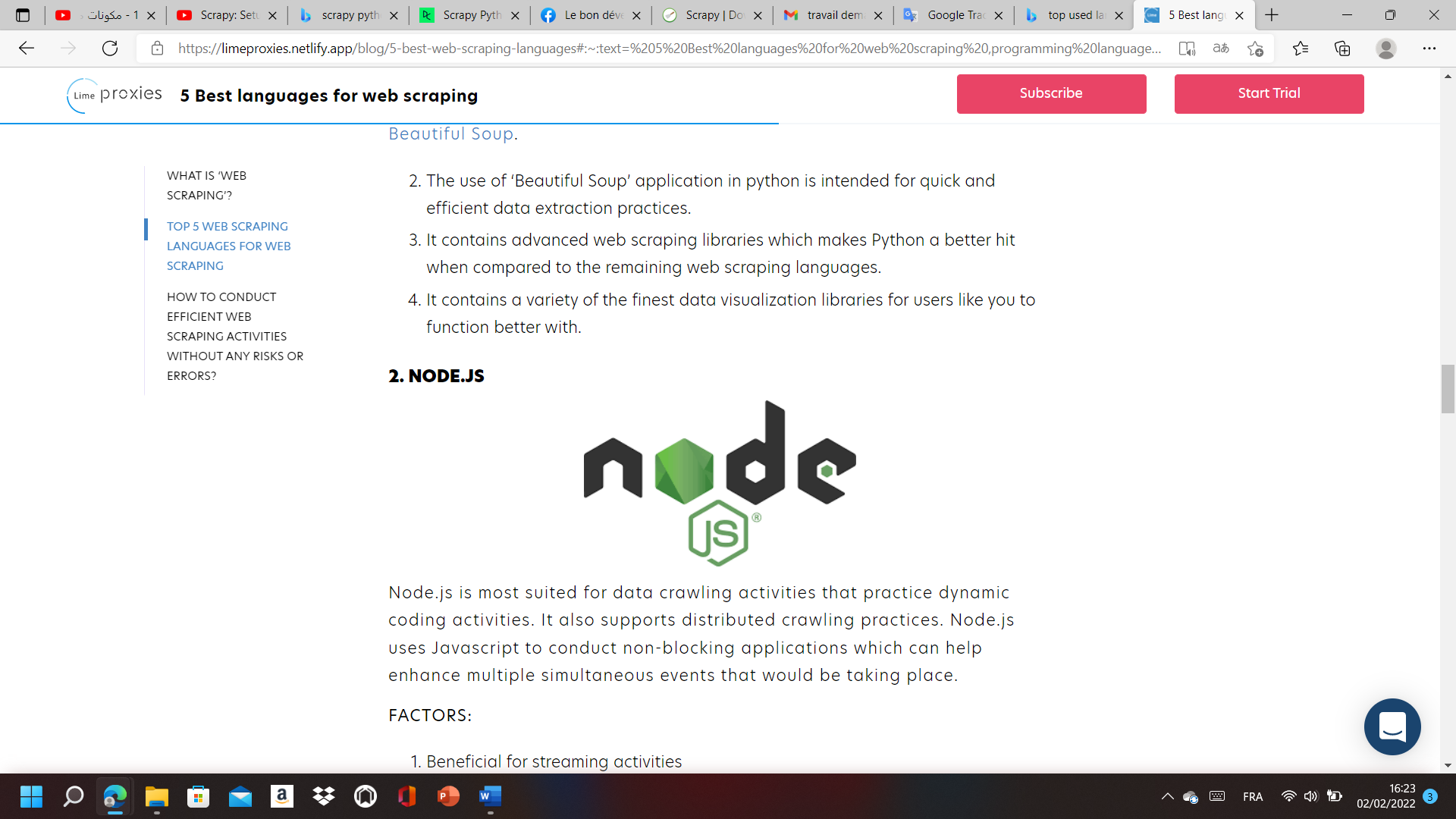
Figure 5 : Python



Python is one of the most common coding languages. With reference to web scraping languages, this is popularly used for such a process. For any web scraping activities, Python is considered to be the finest in ensuring that this process is conducted without any errors.

### Node.js / js :

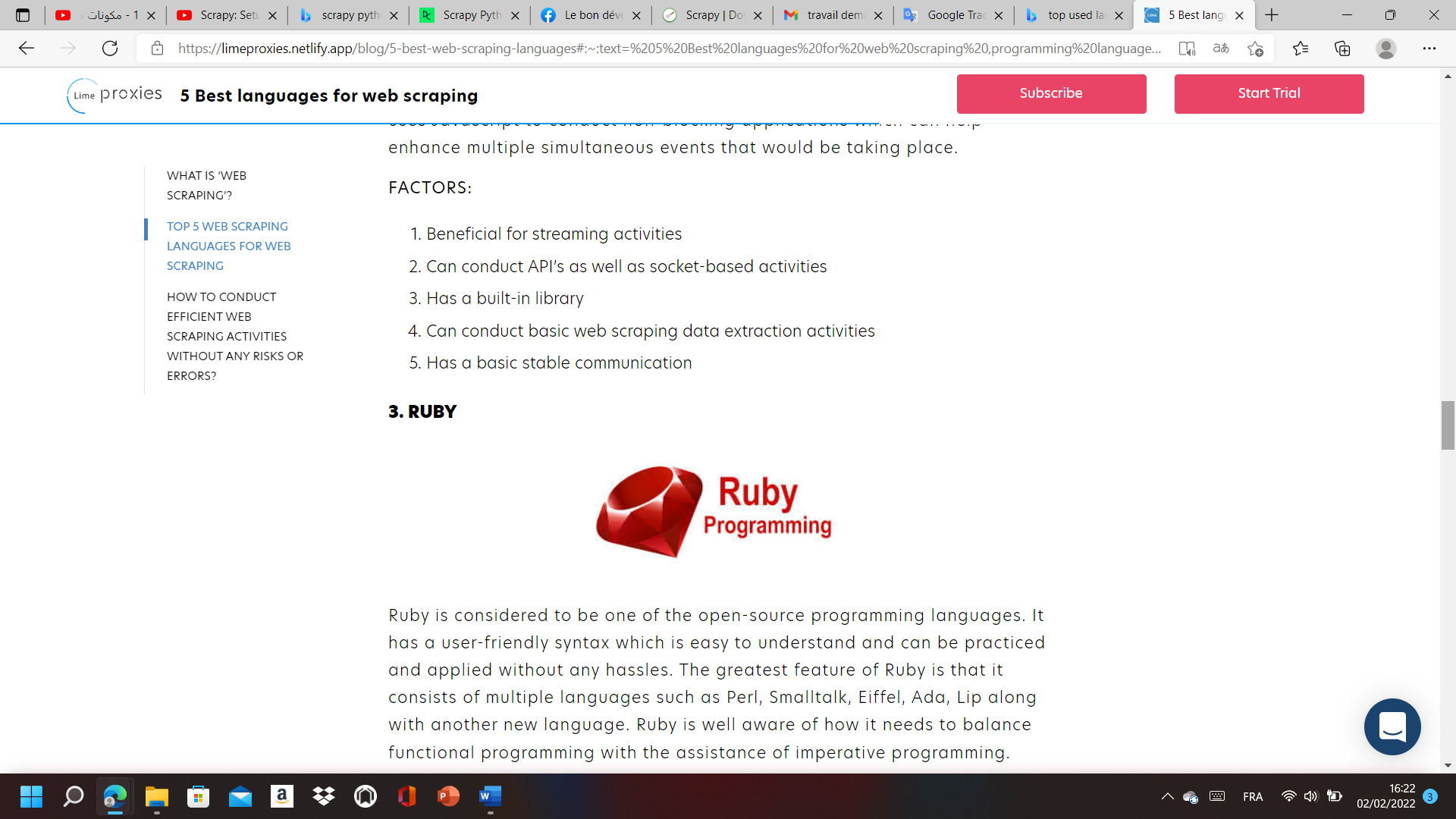
Figure 6 : Node.js



Node.js is most suited for data crawling activities that practice dynamic coding activities. It also supports distributed crawling practices. Node.js uses JavaScript to conduct non-blocking applications which can help enhance multiple simultaneous events that would be taking place.

### Ruby :

Figure 7 : Ruby



Ruby is considered to be one of the open-source programming languages. It has a user-friendly syntax which is easy to understand and can be practiced and applied without any hassles. The greatest feature of Ruby is that it consists of multiple languages such as Perl, Smalltalk, Eiffel, Ada, Lip along with another new language. Ruby is well aware of how it needs to balance functional programming with the assistance of imperative programming.

### C++/C :

Figure 8 : C++



C and C++ are a great execution solution but it can be costly when it comes to conducting web scraping. [Prowebscraper](https://www.prowebscraper.com/blog/best-programming-language-for-web-scraping/)recommends, it is not advisable to use these languages to set up a crawler unless it’s a specialized organization that you have in mind, focusing only on extracting data.

### PHP :

Figure 9 : PHP



PHP may not be able to be the ideal choice when it comes to the creation of a crawler program. In order to extract information such as graphics, images, videos, and other visual forms, using a CURL library is better.

The best thing about the curl library is that it can help to transfer files with the help of protocol lists which has HTTP and FTP in it. Having this can help you in the creation of web spiders which could be utilized to download any kind of information from the online platform.

## Conclusion:

After completing second chapter the internship we will move on To the penultimate stage, which isthe Realization phase.

# Chapter 3: Realization

## Introduction :

In the chapter number 3 I’m going to talk about the specific stages of my work during this month concerned the framework “scrapy” and the library “beautiful soup”

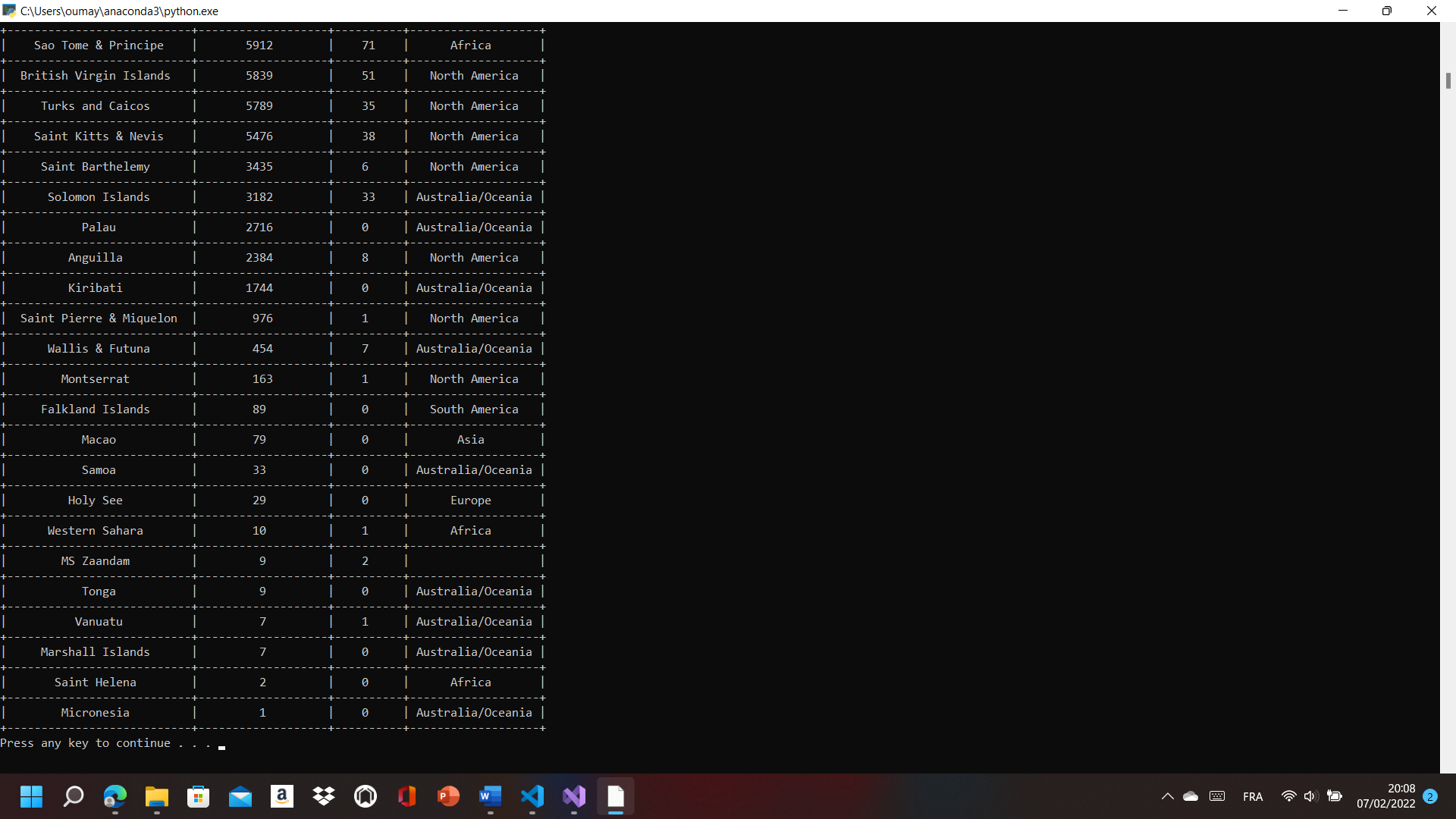
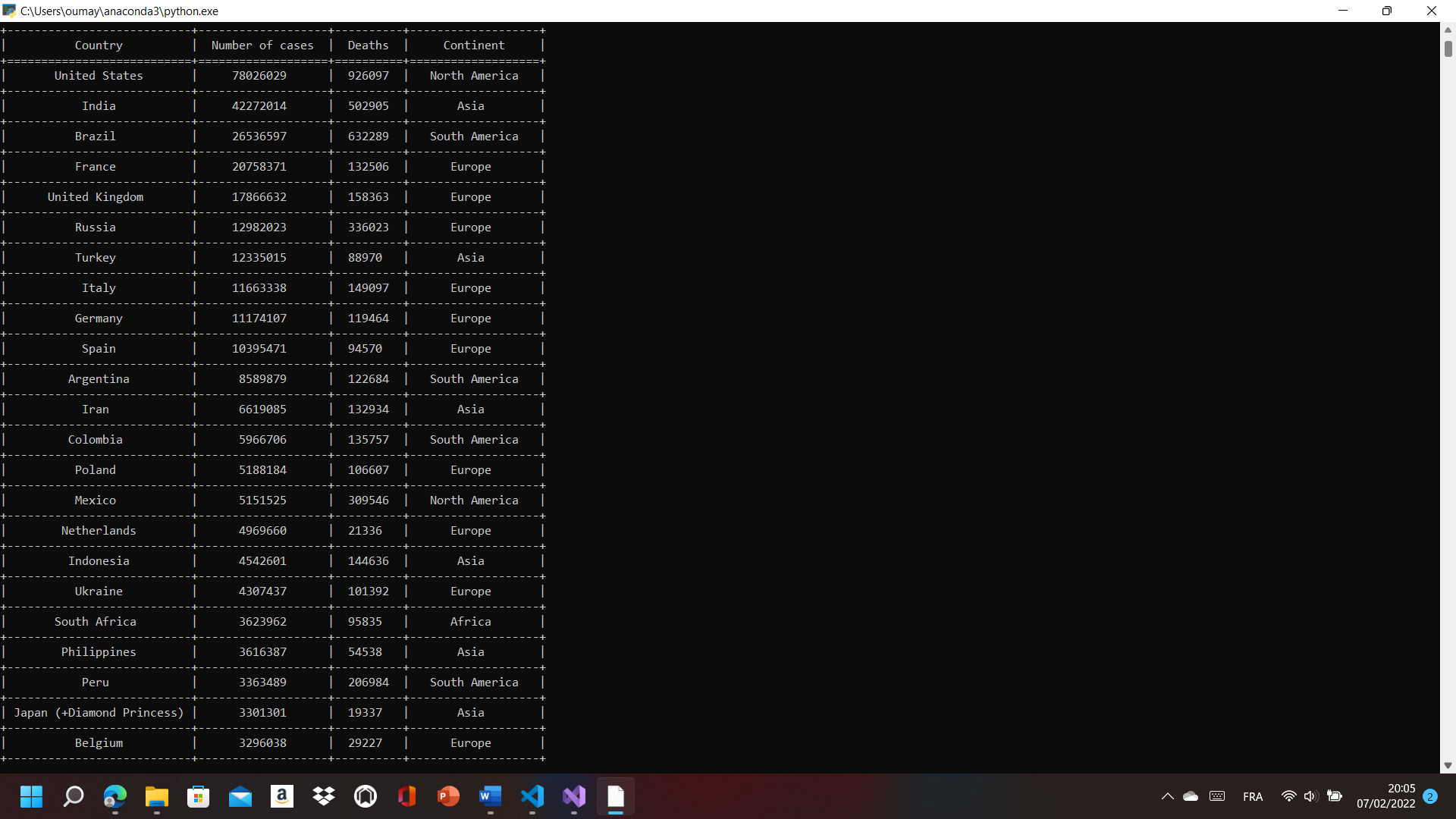
## Beautiful soup :

### Implementation :

We used three steps to run our program. In the first step, to begin with we imported from libraries CSV, Texttable, request and beautiful soup.Then we installed the page’s link which contains all the data . using a variable called URL. added to that, we used two variables called page and soup that can be used to copy and paste the text that exists there. In the second step, we used a variable called data\_iterator where we put all the data that existed in the TDS using a function called Find All. Moreover, we installed 4 variables: country, confirmed, deaths, continent. Each variable contains one of the TDS. In the end, we added a line in the beginning that contains the names of the four columns. Then we displayed it with a function called print. In the last step we added all the information in a list called data using the function append. Then, we classified them row by row using a function called sort. Also, we put all the results in a table. Finally, we run the program.

### Result :

Figure 10 : Result



## Scrapy :

### Implementation :

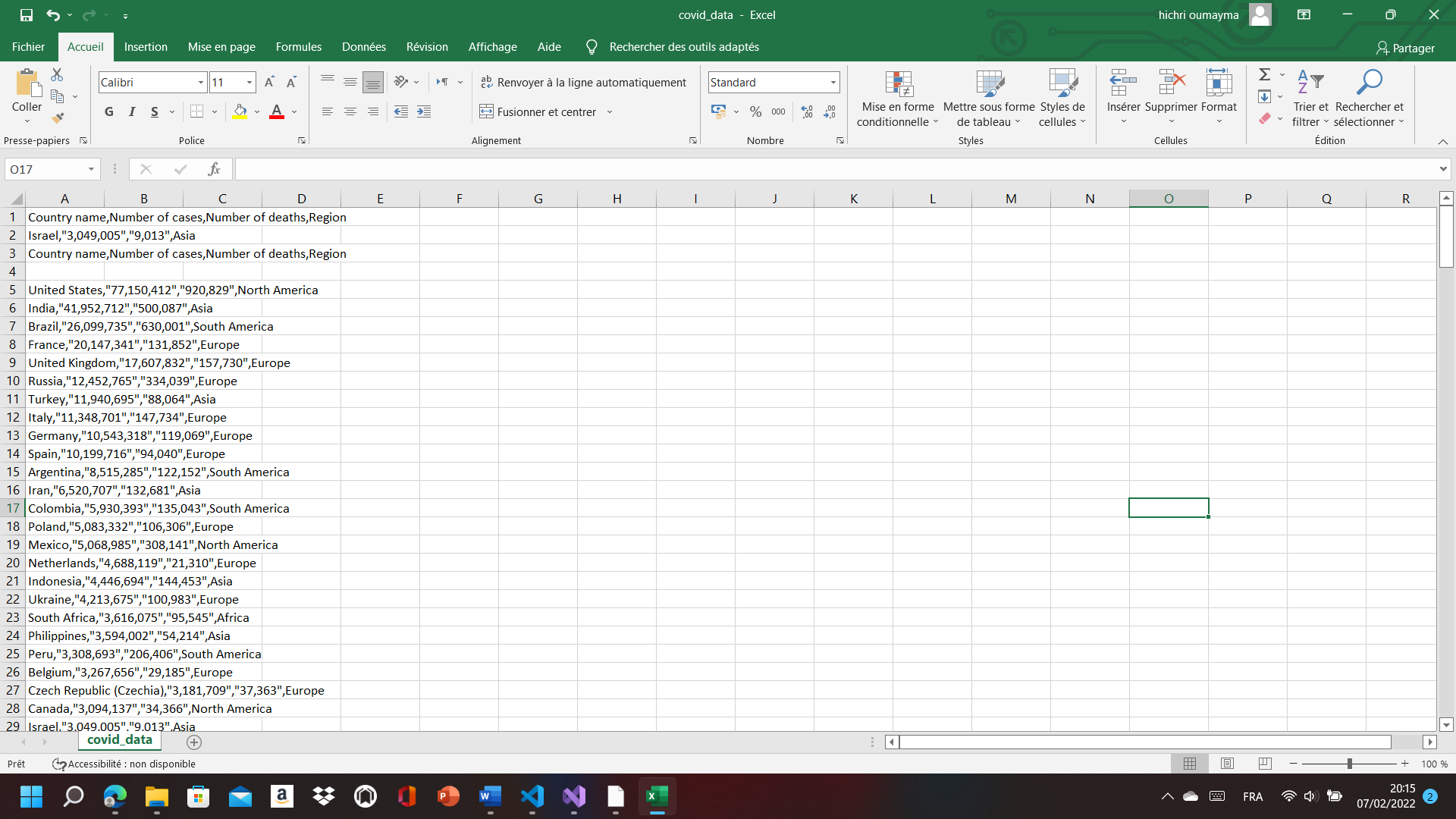
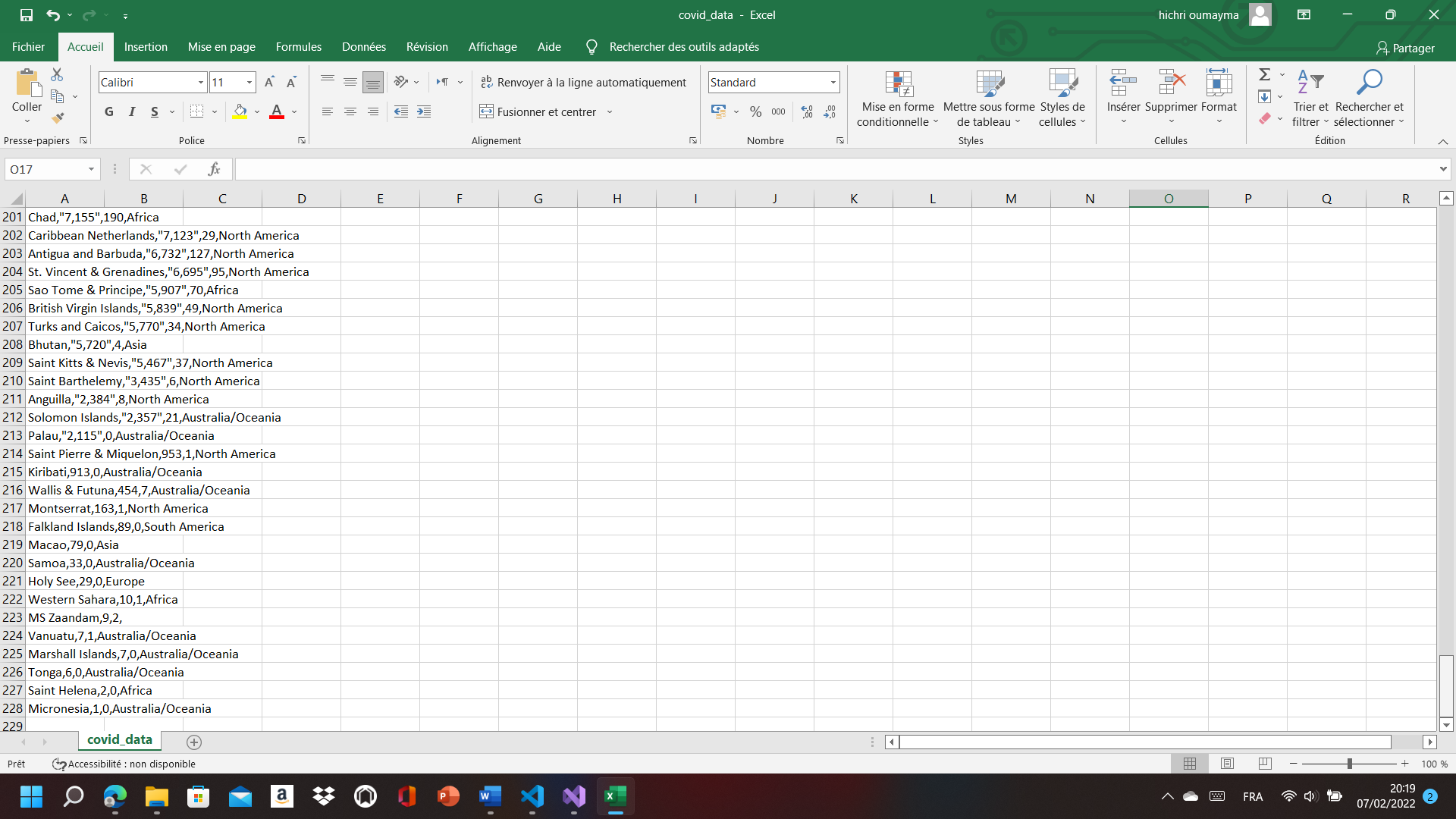
First of all, we installed the python application 3.10 then we used cmd commands to install scrapy by “PIP install scrapy”, we made a new project named by covid\_scrapy using “scrapy start project” command, then we put the website :

<https://www.worldometers.info/coronavirus/countries>

where coronavirus has spread which is full of the information about corona virus by the command “scrapy genspidercovid\_scrapy”. After that we opened the visual studio and we wrote a code where we made a function parse that brings the data from website Finally, we made a csv file by the “scrapy crawl covid\_scrapy -o covid\_data.csv “ where all the data are going to be preserved.

### Result

Figure 11 : Covid \_data.csv

 .... 

## Conclusion

In this chapter, we have described the phase of realization of all the functionalities of our system as well as the main interfaces of the application.

# Chapter 4: Methodological

## Introduction

In the chapter number 4 I’m going to talk about the specific stages of my work during this month concerned the Git and GitHub.

## Git and GitHub:

### What is the Git:

Git is software for tracking changes in any set of files, it was created by Linus Torvalds in 2005 for development of the Linux kernel, usually used for coordinating work among programmers collaboratively developing source code during software development. git is free and open-source distributed version control system designed to handle everything from small to large projects with speed.

### What is the GitHub:

the GitHubis a provider of internet hosting for software development and version control using Git It offers the distributed version control and source code management.

### The relationship between them:

Now git by itself is amazing but GitHub is the web service that allows git to be even more powerful. With GitHub you have team collaboration, you have extensions there’s so much you can do with it, they have their own desktop application, they have their own browser extension …

through the combination of git and GitHub, this term social

coding comes to be where you can combine these two   services and now you can work with anyone around the world on your own piece, on your own package, project you can hold each other accountable you can save you progress.

## The basic command-line of Git:

Tableau 3 : Command line of get

|  |  |
| --- | --- |
| Git init | The git init command creates a new Git repository |
| . Git | that contains all of your necessary repository files — a Git repository skeleton. |
| Git commit | : The git commit command captures a snapshot of the project's currently staged changes |
| Git add | The git add command adds a change in the working directory to the staging area, it tells Git that you want to include updates to a particular file in the next commit |
| Git clone | which is used to target an existing repository and create a clone, or copy of the target repository. |

Git is an important part of daily programming and is widely used in the software industry. Since there are many various commands you can use, mastering Git takes time. But those commands are used more frequently.

## Result

We register the code via git on GitHub. Therefore, it will now be available for people to see:

<https://github.com/OucemaGharsalli0509?tab=repositories>

<https://github.com/Hichri-oumayma?tab=repositories>

## Conclusion

In this chapter, we have described the phasemethodological and in the following chapter we conclude.

# General conclusion

According to this advanced course offered by "ISET Kasserine" in collaboration with the company "Smartica solution", whose main objective is to perform web scraping with the beautiful soup library and with the scrapy framework

In order to achieve this experience,we divided this work into 3 main parts:

For the first part, we started with the description of the organization of which we had the opportunity to carry out this project. In the second part, we indicated the internship.

And in the third part, was devoted to the realization of "web scraping" as well as the presentation of the graphic interfaces of the solution.

Finally, by carrying out this project, we learned to handle the following tools:

✓ Visual Studio code 2022✓ scrapy Framework

✓ Python 3.10✓ invite de command

✓ VS code✓GitHub

This internship allowed us to deepen my theoretical knowledge and improve our professional skills, and to have an idea of ​​the world of work and more precisely in the IT sector.Absolutely we will not stop at this point, we are continuing our progress. Because of the short of time that we had (only one month), we were obliged to stop for the moment.But we will definitely keep moving forward and maybe if we will be in front of you next time, our project will be better in terms of flexibility and visuals.