Configuration Manual

MSc Research Project
Cybersecurity

Marcello D’Angelone
Student ID: x21113777

School of Computing
National College of Ireland

Supervisor: Mark Monaghan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Marcello D'Angelone		
Student ID:	X21113777		
Programme:	MSc Cybersecurity	**Year:**	2022/23
Module:	Research Project		
Supervisor:	Mark Monaghan		
Submission Due Date:	Aug. 14 2023		
Project Title:	Email spoofing defence techniques: a comprehensive review and development of new measurement tool		
Word Count:	……1936……… **Page Count**…………15………………………		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Marcello D'Angelone

Date: 06/08/2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Marcello D’Angelone
X21113777

1 Introduction

This configuration manual outlines the requirements and steps to reproduce the environment and execute the Domain Crawler tool.

2 Hardware Configuration

The research project has been developed on a System with the hardware configuration represented in Table 1.

Table 1: Hardware Configuration

System Manufacturer	LENOVO
System Model	20L8S2Y100
Processor	Intel(R) Core(TM) i7-8650U CPU @ 1.90GHz, 2112 Mhz, 4 Core(s), 8 Logical Processor(s)
Installed Physical Memory (RAM)	24.0 GB
Hard Disk (SSD)	512 GB

The Domain Crowler tool was executed on the same hardware configuration for testing and datasets analysis purposes. A separate Amazon EC2 instance was created[1] to execute longer-running large-scale scans. Its hardware specifications are reported in Table 2.

Table 2: EC2 Instance details

System Manufacturer	AWS
Instance type	t2.micro
Processor	Intel(R) Xeon(R) CPU E5-2686 v4 @ 2.30GHz
Installed Physical Memory (RAM)	1 GB
Hard Disk (SSD)	16 GB

3 Software Configuration

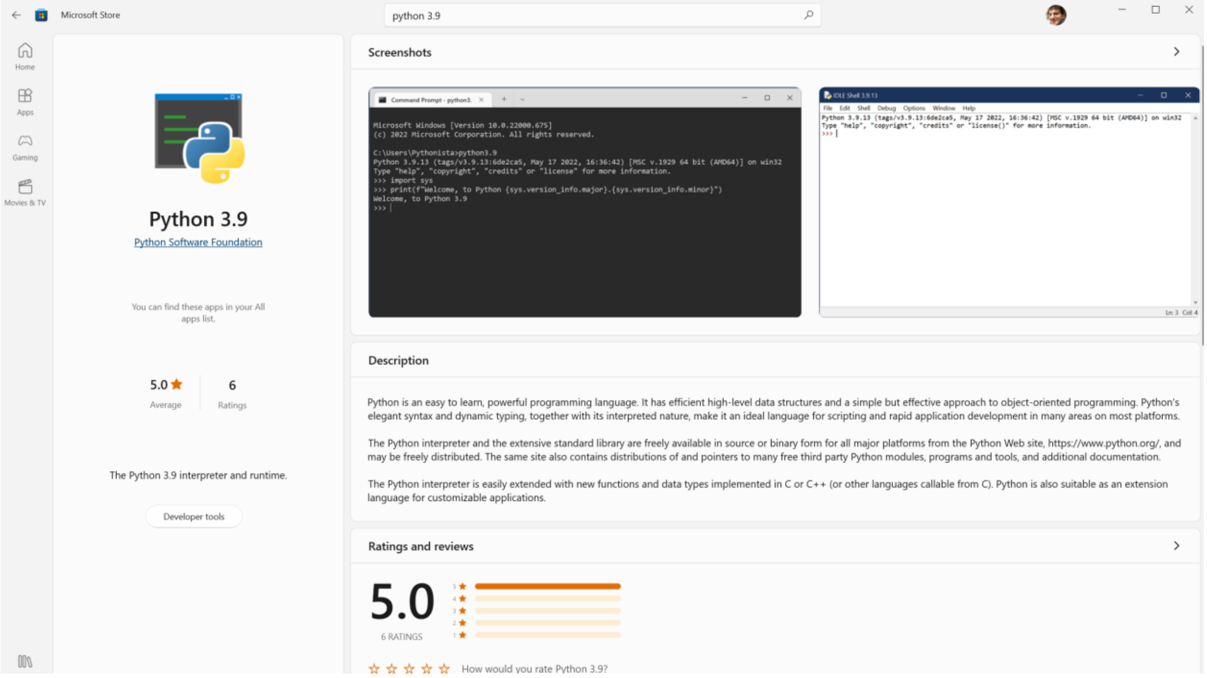
The minimum software required to reproduce the research findings is listed below:

[1] Tutorial: Get started with Amazon EC2 Linux instances: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EC2_GetStarted.html

Table 3: Software Configuration

Software	Version	Type
Microsoft Windows 10 Enterprise (64bit)	10.0.19045 Build 19045	Operating System on Lenovo Hardware
Debian	debian-11-amd64-20230515-1381	Operating System on EC2 instance
Python	3.9.2	Programming Language
Pycharm	Community Edition 2023.1	IDE
R for Windows	4.3.0	Programming Language for statistical analysis
Rstudio	2023.06	R Integrated Developer Environment
Excel	2304	Spreadsheet
Putty	0.74	SSH access to the AWS EC2 instance
WinSCP	6.1.1	Windows Secure copy to copy the tool and datasets to the AWS EC2 instance

Python can be installed on Windows 10 directly from the Microsoft store by typing Python 3.9 in the search bar, as depicted in Figure 1.


Microsoft Store
python 3.9
Home
Apps
Gaming
Movies & TV
Python 3.9
Python Software Foundation
You can find these apps in your All apps list.
5.0 ★
Average
6
Ratings
The Python 3.9 interpreter and runtime.
Developer tools
Screenshots
Command Prompt - python3.
IDLE Shell 3.9.13
File Edit Shell Debug Options Window Help
Ln: 3 Col: 4
Description
Ratings and reviews
5.0
6 RATINGS
How would you rate Python 3.9?

```
Microsoft Windows [Version 10.0.22000.675]
(c) 2022 Microsoft Corporation. All rights reserved.

C:\Users\Pythonista>python3.9
Python 3.9.13 (tags/v3.9.13:6de2ca5, May 17 2022, 16:36:42) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import sys
>>> print(f"Welcome, to Python {sys.version_info.major}.{sys.version_info.minor}")
Welcome, to Python 3.9
>>>
```

```
Python 3.9.13 (tags/v3.9.13:6de2ca5, May 17 2022, 16:36:42) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
```

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms.

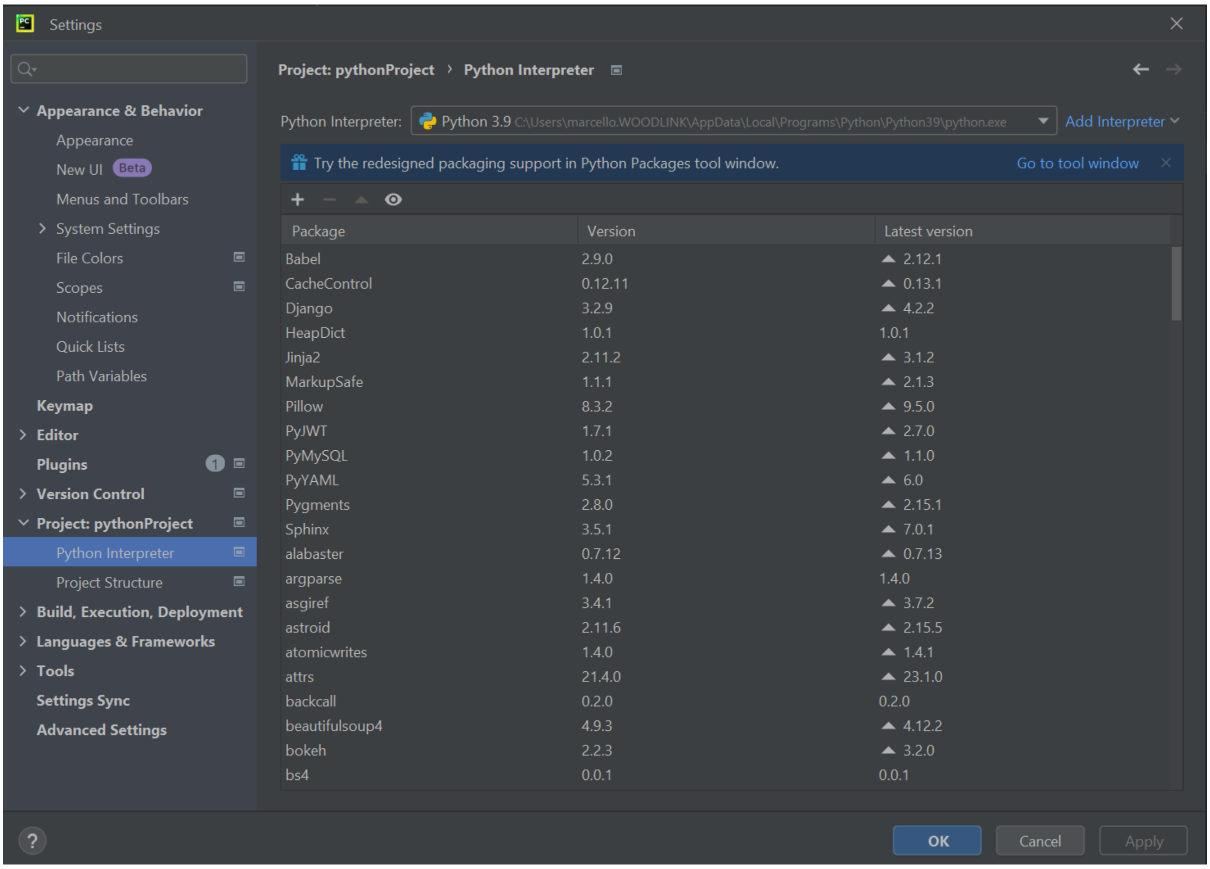
The Python interpreter and the extensive standard library are freely available in source or binary form for all major platforms from the Python Web site, https://www.python.org/, and may be freely distributed. The same site also contains distributions of and pointers to many free third party Python modules, programs and tools, and additional documentation.

The Python interpreter is easily extended with new functions and data types implemented in C or C++ (or other languages callable from C). Python is also suitable as an extension language for customizable applications.

Figure 1: Python installation on Windwos 10

The next step requires installing an Integrated Developer Environment (IDE) for Python. For this research, PyCharm Community Edition 2023.1 has been used. This free IDE can be downloaded from the JetBrains website[2].

Once the installation is completed, the Python interpreter should be selected from File > Settings and Python Interpreter, as represented in Figure 2.



Python Interpreter: Python 3.9 C:\Users\marcello.WOODLINK\AppData\Local\Programs\Python\Python39\python.exe

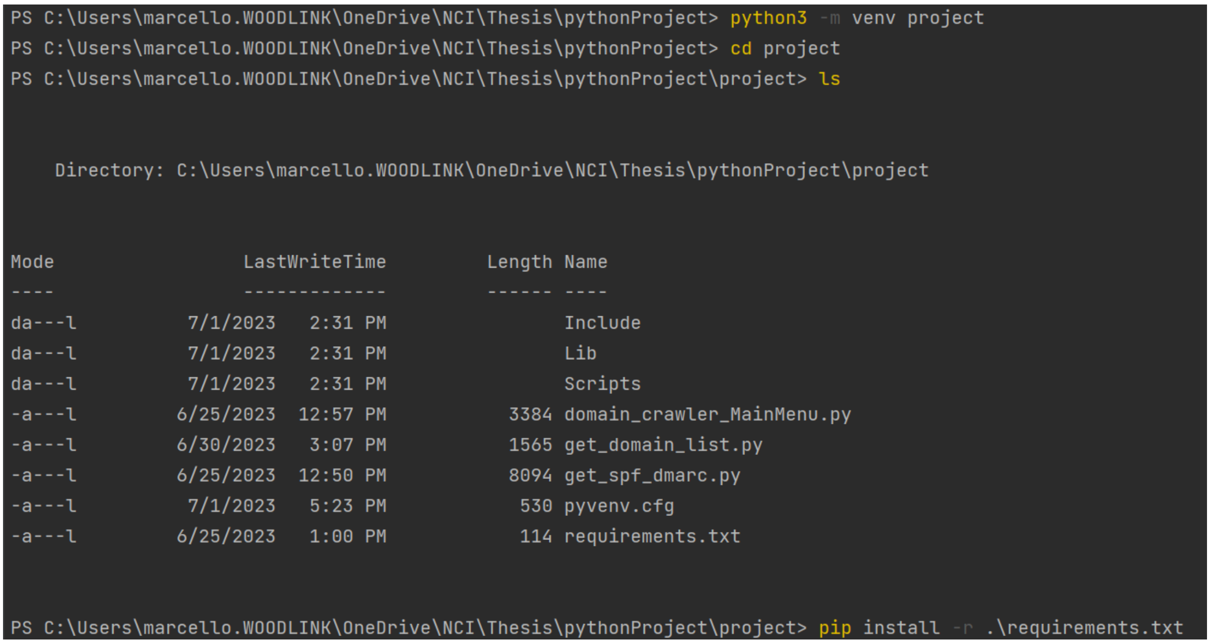
Try the redesigned packaging support in Python Packages tool window.

Package	Version	Latest version
Babel	2.9.0	▲ 2.12.1
CacheControl	0.12.11	▲ 0.13.1
Django	3.2.9	▲ 4.2.2
HeapDict	1.0.1	1.0.1
Jinja2	2.11.2	▲ 3.1.2
MarkupSafe	1.1.1	▲ 2.1.3
Pillow	8.3.2	▲ 9.5.0
PyJWT	1.7.1	▲ 2.7.0
PyMySQL	1.0.2	▲ 1.1.0
PyYAML	5.3.1	▲ 6.0
Pygments	2.8.0	▲ 2.15.1
Sphinx	3.5.1	▲ 7.0.1
alabaster	0.7.12	▲ 0.7.13
argparse	1.4.0	1.4.0
asgiref	3.4.1	▲ 3.7.2
astroid	2.11.6	▲ 2.15.5
atomicwrites	1.4.0	▲ 1.4.1
attrs	21.4.0	▲ 23.1.0
backcall	0.2.0	0.2.0
beautifulsoup4	4.9.3	▲ 4.12.2
bokeh	2.2.3	▲ 3.2.0
bs4	0.0.1	0.0.1

Figure 2: IDE interpreter configuration

Once the PyCharm IDE is installed, creating a virtual environment where the code is executed with all the required libraries is recommended. The *python3 -m venv project* command must be executed from the IDE terminal or an SSH shell to do so. Then the *cd* command allows to change the directory to the project folder. Finally, the Python files *domain_crawler_MainMenu.py*, *get_domain_list.py* and *get_spf_dmar.py* user-defined functions and the requirements.txt file must be copied to the same project folder. To install the required libraries, it is necessary to type *pip install -r requirements.txt*. These steps are illustrated in Figure 3.

[2] https://www.jetbrains.com/pycharm/download/?section=windows



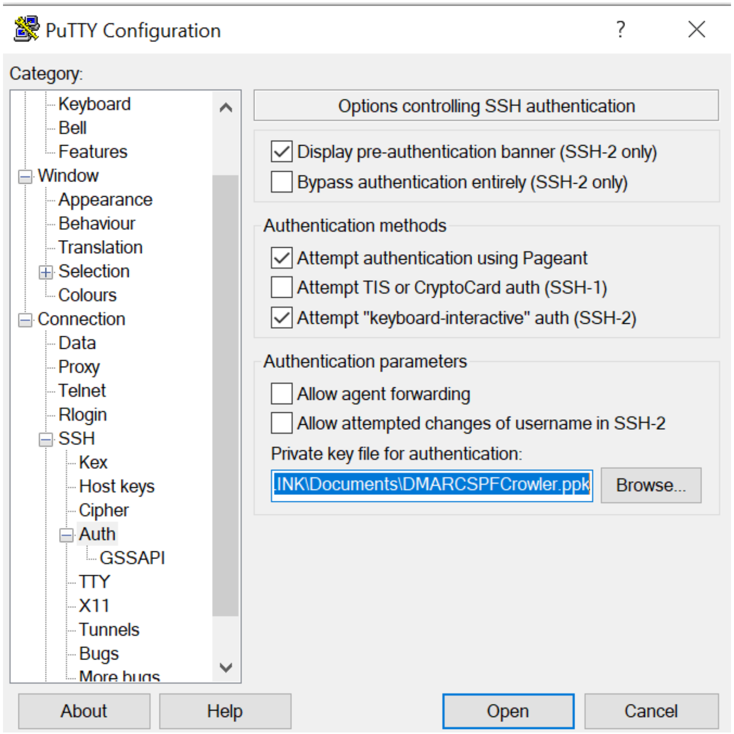
```
PS C:\Users\marcello.WOODLINK\OneDrive\NCI\Thesis\pythonProject> python3 -m venv project
PS C:\Users\marcello.WOODLINK\OneDrive\NCI\Thesis\pythonProject> cd project
PS C:\Users\marcello.WOODLINK\OneDrive\NCI\Thesis\pythonProject\project> ls


    Directory: C:\Users\marcello.WOODLINK\OneDrive\NCI\Thesis\pythonProject\project


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
da---l          7/1/2023   2:31 PM                Include
da---l          7/1/2023   2:31 PM                Lib
da---l          7/1/2023   2:31 PM                Scripts
-a---l         6/25/2023  12:57 PM           3384 domain_crawler_MainMenu.py
-a---l         6/30/2023   3:07 PM           1565 get_domain_list.py
-a---l         6/25/2023  12:50 PM           8094 get_spf_dmarc.py
-a---l          7/1/2023   5:23 PM            530 pyvenv.cfg
-a---l         6/25/2023   1:00 PM            114 requirements.txt


PS C:\Users\marcello.WOODLINK\OneDrive\NCI\Thesis\pythonProject\project> pip install -r .\requirements.txt
```

Figure 3: Python virtual environment and libraries installation

Like Python, R requires the programming language and the IDE to be installed locally on the machine. R can be downloaded from the Comprehensive R Archive Network[3] (CRAN), whereas the Rstudio desktop can be downloaded from the Posit website[4].
Excel is part of the Office 365 suite and is pre-installed on most Windows Laptops. If this is missing, or the CSV file containing the domain list is imported on a Linux machine, an open-source version called LibreOffice can be downloaded and installed[5].

Putty is a ssh client for Windows Operating systems which has been used to connect to the EC2 instance and execute the tool. This is freeware that can be downloaded and installed from the Putty website[6]. Putty has no Linux or Mac version since they both have a built-in ssh client. The connection is authenticated through SSH key pair created as described in the EC2 tutorial and configured in Figure 4.

[3] R: https://cran.rstudio.com/
[4] RStudio: https://posit.co/downloads/
[5] Libreoffice: https://www.libreoffice.org/download/download-libreoffice/
[6] Putty: https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html



Options controlling SSH authentication

☑ Display pre-authentication banner (SSH-2 only)
☐ Bypass authentication entirely (SSH-2 only)

Authentication methods

☑ Attempt authentication using Pageant
☐ Attempt TIS or CryptoCard auth (SSH-1)
☑ Attempt "keyboard-interactive" auth (SSH-2)

Authentication parameters

☐ Allow agent forwarding
☐ Allow attempted changes of username in SSH-2

Private key file for authentication:
INK\Documents\DMARCSPFCrowler.ppk

Figure 4: Putty SSH private key configuration

WinSCP is a secure file transfer utility for Windows used to transfer the code and CSV files to the EC2 instance, and it can be downloaded from the official WinSCP website[7]. The file transfer is performed via FTP over and SSH encrypted connection, which leverages Putty keys and credentials.

4 Datasets preparation

4.1 CSV datasets procedure

The datasets obtained in CSV format, along with their referenced download link, are listed below:

1. The Alexa Top 1 Million domain list[8].
2. Moz's list of the 500[9].
3. DomCop[10].
4. U.S. General Services Administration[11].

All the datasets in question have a column with a list of domains. The only requirement to parse these CSV files is that its column header must be called "Domain". As depicted in

[7] WinSCP: https://winscp.net/download/WinSCP-6.1.1-Setup.exe

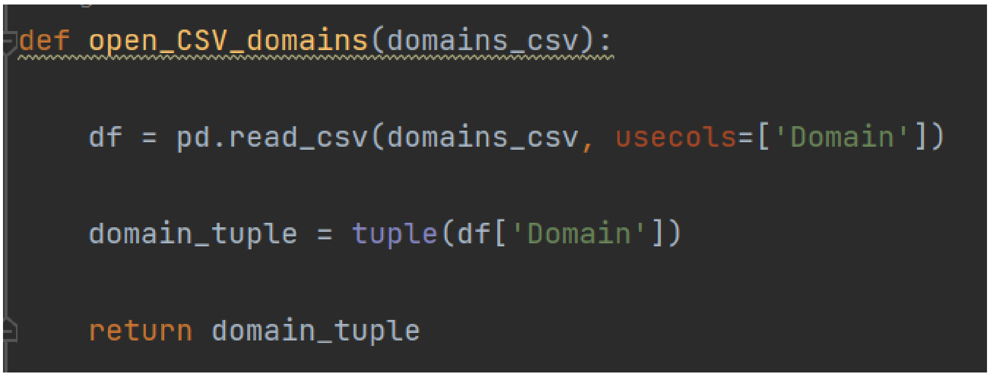
[8] Alexa Top 1 Million: http://s3.amazonaws.com/alexa-static/top-1m.csv.zip

[9] Moz: https://moz.com/top-500/download/?table=top500Domains

[10] DomCop: https://www.domcop.com/files/top/top10milliondomains.csv.zip

[11] U.S. General Services Administration: https://github.com/GSA/govt-urls/blob/main/1_govt_urls_full.csv

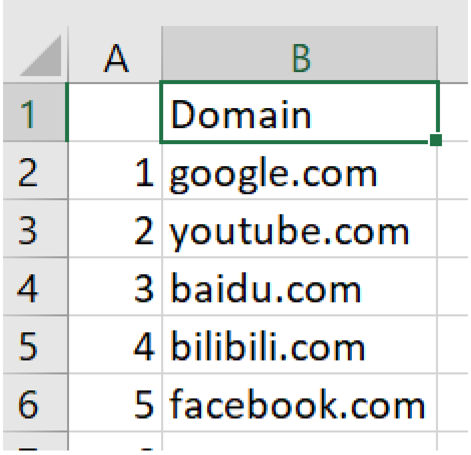
Figure 5, the function *open_CSV_domains* in *get_domain_list.py* that imports the CSV file as pandas dataframe expects a column header called “Domain”. This means that the column header must be edited if it is called with a different name.



```
def open_CSV_domains(domains_csv):

    df = pd.read_csv(domains_csv, usecols=['Domain'])

    domain_tuple = tuple(df['Domain'])

    return domain_tuple
```

Figure 5: Open CSV function

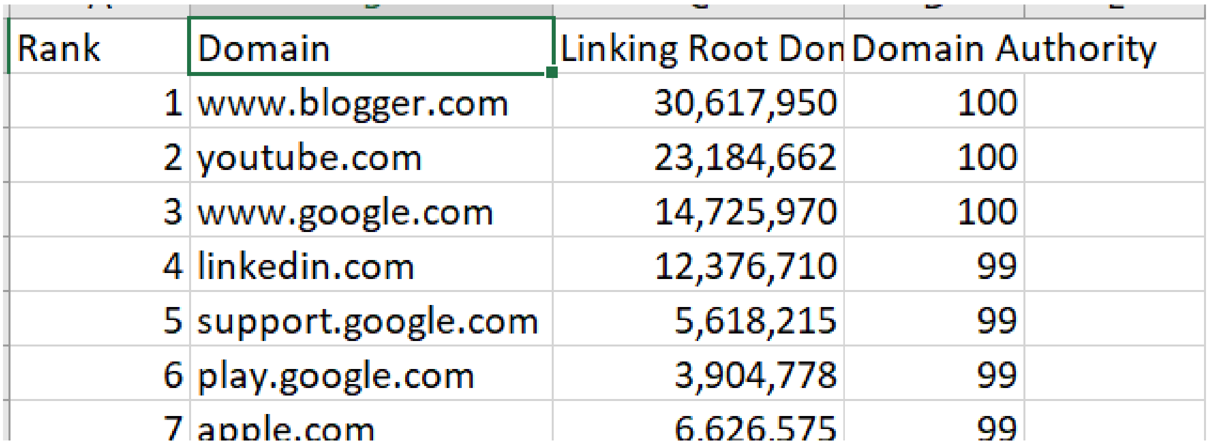
For instance, the Alexa Top 1 Million domain list contains a disclaimer and no Domain header. The disclaimer needs to be removed and added the Domain header in the column with the list of domains as displayed in Figure 6.



	A	B
1		Domain
2	1	google.com
3	2	youtube.com
4	3	baidu.com
5	4	bilibili.com
6	5	facebook.com

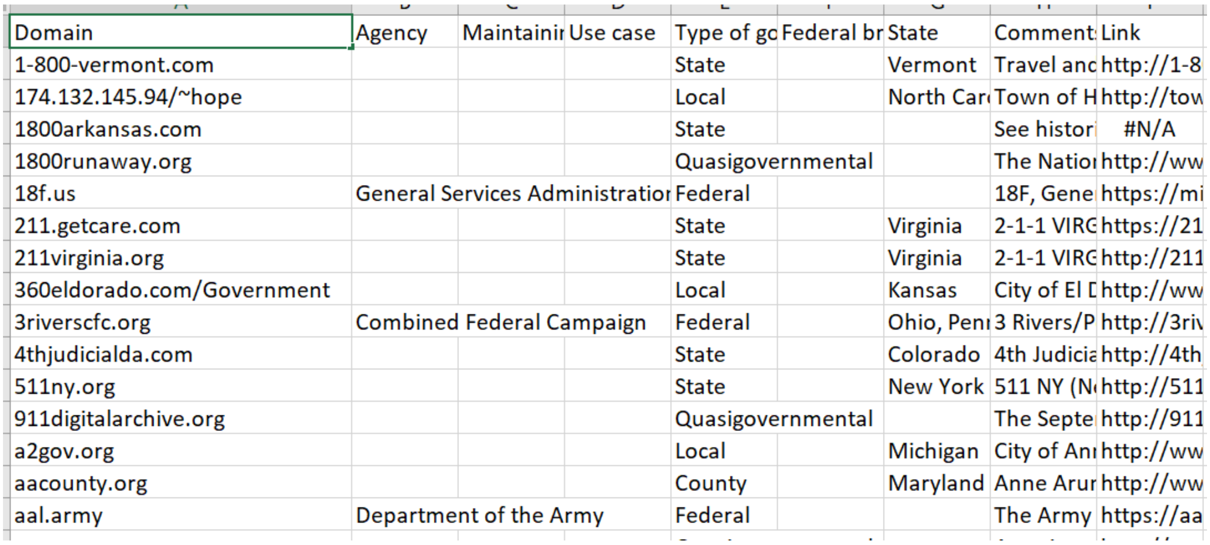
Figure 6: Alexa Top 1 Million

The Moz and the U.S. General Service Administration CSV files present a domain header called “Root Domain” and “Doman name”, respectively. These need to be renamed as “Domain” as per Figures 7 and 8.



Rank	Domain	Linking Root Don	Domain Authority
1	www.blogger.com	30,617,950	100
2	youtube.com	23,184,662	100
3	www.google.com	14,725,970	100
4	linkedin.com	12,376,710	99
5	support.google.com	5,618,215	99
6	play.google.com	3,904,778	99
7	apple.com	6,626,575	99

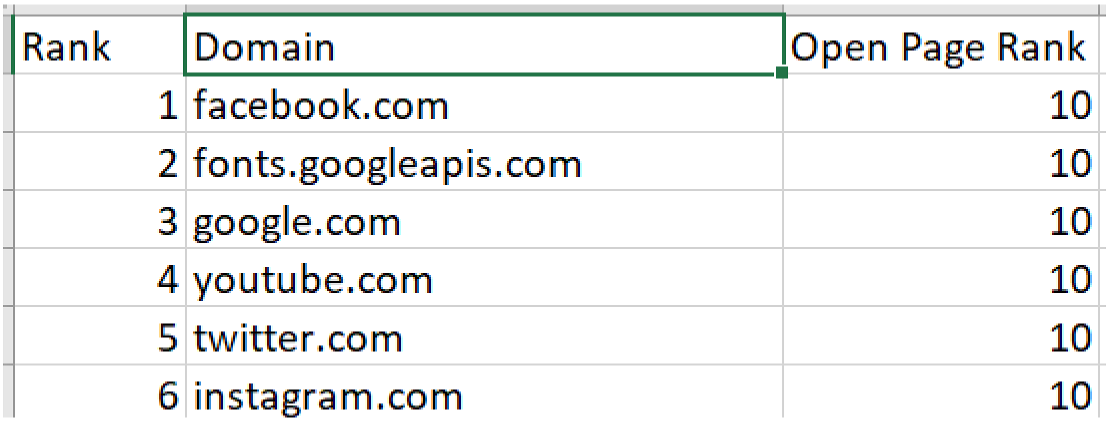
Figure 7: Moz Top 500 domain list



Domain	Agency	Maintainir	Use case	Type of gc	Federal br	State	Comment:	Link
1-800-vermont.com				State		Vermont	Travel anc	http://1-8
174.132.145.94/~hope				Local		North Car	Town of H	http://tow
1800arkansas.com				State			See histor	#N/A
1800runaway.org				Quasigovernmental			The Natior	http://ww
18f.us	General Services Administratior			Federal			18F, Gene	https://mi
211.getcare.com				State		Virginia	2-1-1 VIRG	https://21
211virginia.org				State		Virginia	2-1-1 VIRG	http://211
360eldorado.com/Government				Local		Kansas	City of El D	http://ww
3riverscfc.org	Combined Federal Campaign			Federal		Ohio, Penı	3 Rivers/P	http://3riv
4thjudicialda.com				State		Colorado	4th Judicia	http://4th
511ny.org				State		New York	511 NY (N	http://511
911digitalarchive.org				Quasigovernmental			The Septe	http://911
a2gov.org				Local		Michigan	City of Anı	http://ww
aacounty.org				County		Maryland	Anne Arur	http://ww
aal.army	Department of the Army			Federal			The Army	https://aa

Figure 8: U.S. General Service Administration domain list

On the other hand, the DomCop represented in Figure 9 does not require any editing as the column header is already called Domain.



Rank	Domain	Open Page Rank
1	facebook.com	10
2	fonts.googleapis.com	10
3	google.com	10
4	youtube.com	10
5	twitter.com	10
6	instagram.com	10

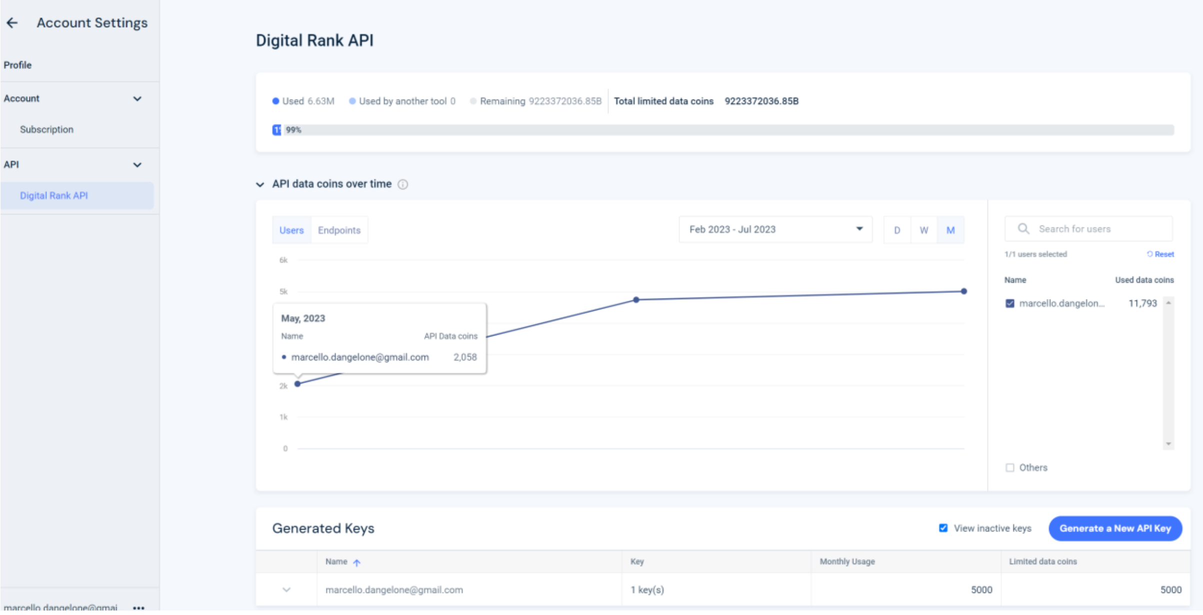
Figure 9: DomCop 10 Million domain list

Of course, any CSV file can be customised and compiled with any number of rows as long as its header is called “Domain”.

4.2 API dataset preparation

The dataset obtained from the Similarweb API requires an account creation through the demo account portal[12]. Once the account is created and the user logged in, the following steps are required to obtain an API key:

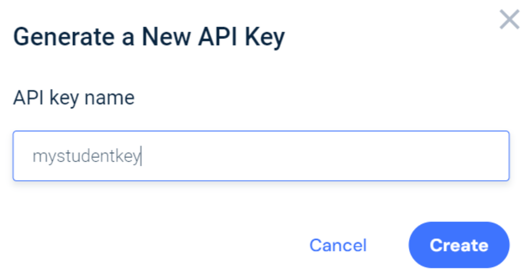
1. Under the Account Settings select “Digital Rank API”
2. From this page click “Generate a New API Key” as depicted in Figure 10



Name	Key	Monthly Usage	Limited data coins
marcello.dangelone@gmail.com	1 key(s)	5000	5000

Figure 10: Digital Rank API key generation

3. Type a name and select Create as described in Figure 11.



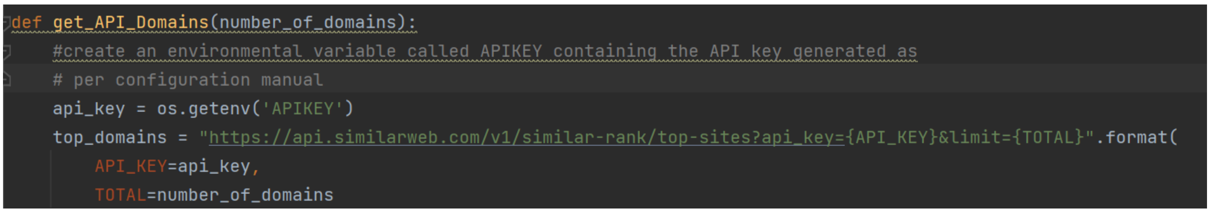
API key name: mystudentkey

Figure 11: Create API key

4. The API key will appear in the “Generated Keys” table.

[12] Similarweb Demo account: https://account.similarweb.com/journey/registration?workspace=marketing&flow=regular

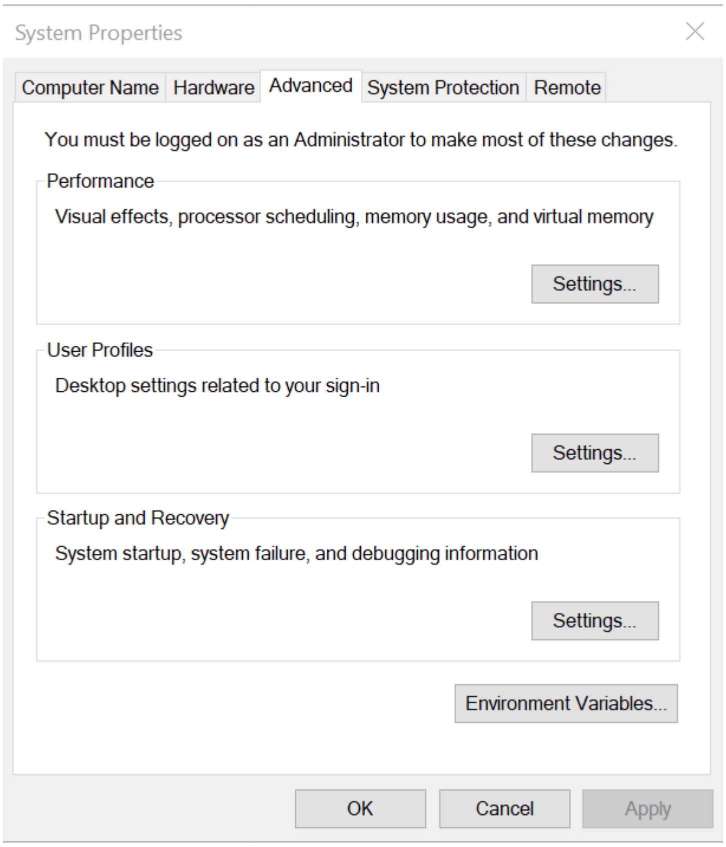
Because it is a poor security practice to hard-code secrets, tokens or other sensitive information directly in the code or configuration files, the following steps are required to store the Similarweb API key in an environmental variable.



```
def get_API_Domains(number_of_domains):
    #create an environmental variable called APIKEY containing the API key generated as
    # per configuration manual
    api_key = os.getenv('APIKEY')
    top_domains = "https://api.similarweb.com/v1/similar-rank/top-sites?api_key={API_KEY}&limit={TOTAL}".format(
        API_KEY=api_key,
        TOTAL=number_of_domains
```

Figure 12: Create API key

Setting the APIKEY environmental variable on Windows, can be done by typing environment in the search bar, this opens the System properties as represented in Figure 13.



You must be logged on as an Administrator to make most of these changes.

Performance

Visual effects, processor scheduling, memory usage, and virtual memory

User Profiles

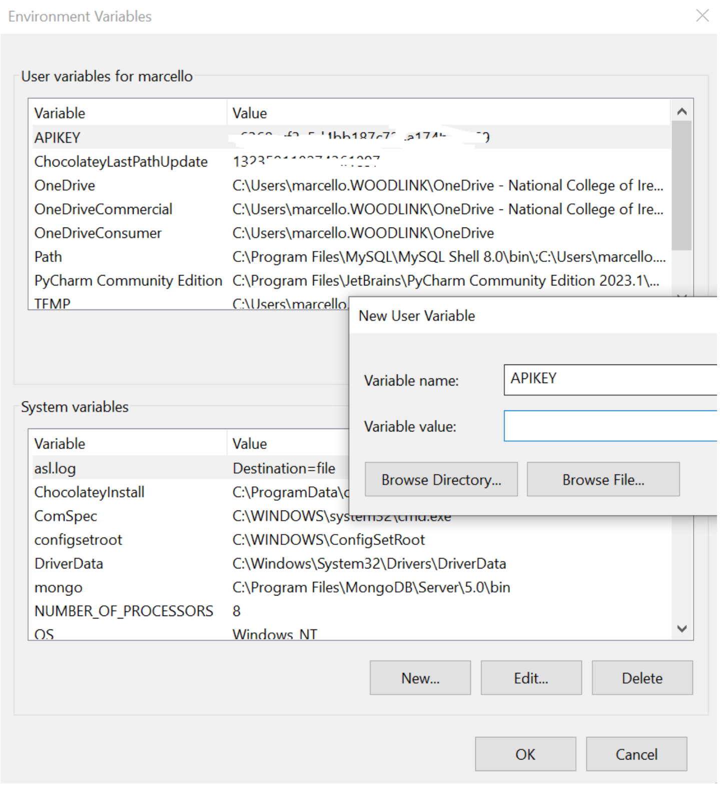
Desktop settings related to your sign-in

Startup and Recovery

System startup, system failure, and debugging information

Figure 13: System properties

Then selecting “Environment Variables…” the Window represented in Figure 14 will open. From here, by selecting “new” under User variable, the user can type APIKEY under Variable Name and paste the key obtained in the steps described to obtain the Similarweb API key.



User variables for marcello

Variable	Value
APIKEY	[redacted]
ChocolateyLastPathUpdate	[redacted]
OneDrive	C:\Users\marcello.WOODLINK\OneDrive - National College of Ire...
OneDriveCommercial	C:\Users\marcello.WOODLINK\OneDrive - National College of Ire...
OneDriveConsumer	C:\Users\marcello.WOODLINK\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\marcello....
PyCharm Community Edition	C:\Program Files\JetBrains\PyCharm Community Edition 2023.1\...
TEMP	C:\Users\marcello

New User Variable

Variable name: APIKEY

Variable value:

System variables

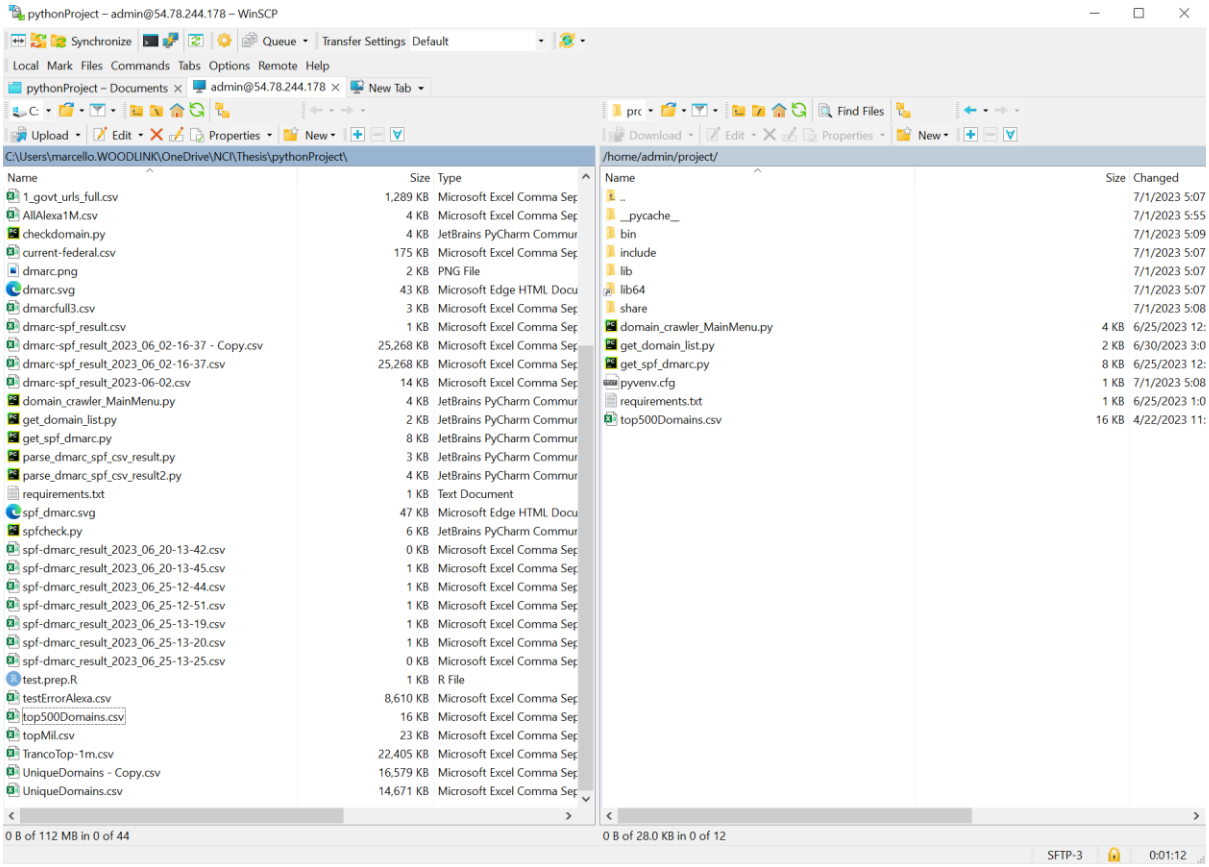
Variable	Value
asl.log	Destination=file
ChocolateyInstall	C:\ProgramData\c
ComSpec	C:\WINDOWS\system32\cmd.exe
configsetroot	C:\WINDOWS\ConfigSetRoot
DriverData	C:\Windows\System32\Drivers\DriverData
mongo	C:\Program Files\MongoDB\Server\5.0\bin
NUMBER_OF_PROCESSORS	8
OS	Windows_NT

Figure 14: Environment Variables

Once this is saved by pressing OK twice the domain crawler tool will be able to access the key via the *os.getenv('APIKEY')* statement in the *get_domain_list.py* function.

5 Execution

The tool can be executed directly from the IDE or from the EC2 instance after copying the required files to the EC2 virtual environment via WinSCP, as depicted in Figure 15. With a simple drag and drop from the required files from the window on the left-hand side to the one on the right-hand side.


pythonProject – admin@54.78.244.178 – WinSCP
Synchronize
Queue
Transfer Settings Default
Local Mark Files Commands Tabs Options Remote Help
pythonProject – Documents
admin@54.78.244.178
New Tab
C:
Upload
Edit
Properties
New
prc
Find Files
Download
SFTP-3
0:01:12

C:\Users\marcello.WOODLINK\OneDrive\NCI\Thesis\pythonProject\

Name	Size	Type
1_govt_urls_full.csv	1,289 KB	Microsoft Excel Comma Sep
AllAlexa1M.csv	4 KB	Microsoft Excel Comma Sep
checkdomain.py	4 KB	JetBrains PyCharm Commur
current-federal.csv	175 KB	Microsoft Excel Comma Sep
dmarc.png	2 KB	PNG File
dmarc.svg	43 KB	Microsoft Edge HTML Docu
dmarcfull3.csv	3 KB	Microsoft Excel Comma Sep
dmarc-spf_result.csv	1 KB	Microsoft Excel Comma Sep
dmarc-spf_result_2023_06_02-16-37 - Copy.csv	25,268 KB	Microsoft Excel Comma Sep
dmarc-spf_result_2023_06_02-16-37.csv	25,268 KB	Microsoft Excel Comma Sep
dmarc-spf_result_2023-06-02.csv	14 KB	Microsoft Excel Comma Sep
domain_crawler_MainMenu.py	4 KB	JetBrains PyCharm Commur
get_domain_list.py	2 KB	JetBrains PyCharm Commur
get_spf_dmarc.py	8 KB	JetBrains PyCharm Commur
parse_dmarc_spf_csv_result.py	3 KB	JetBrains PyCharm Commur
parse_dmarc_spf_csv_result2.py	4 KB	JetBrains PyCharm Commur
requirements.txt	1 KB	Text Document
spf_dmarc.svg	47 KB	Microsoft Edge HTML Docu
spfcheck.py	6 KB	JetBrains PyCharm Commur
spf-dmarc_result_2023_06_20-13-42.csv	0 KB	Microsoft Excel Comma Sep
spf-dmarc_result_2023_06_20-13-45.csv	1 KB	Microsoft Excel Comma Sep
spf-dmarc_result_2023_06_25-12-44.csv	1 KB	Microsoft Excel Comma Sep
spf-dmarc_result_2023_06_25-12-51.csv	1 KB	Microsoft Excel Comma Sep
spf-dmarc_result_2023_06_25-13-19.csv	1 KB	Microsoft Excel Comma Sep
spf-dmarc_result_2023_06_25-13-20.csv	1 KB	Microsoft Excel Comma Sep
spf-dmarc_result_2023_06_25-13-25.csv	0 KB	Microsoft Excel Comma Sep
test.prep.R	1 KB	R File
testErrorAlexa.csv	8,610 KB	Microsoft Excel Comma Sep
top500Domains.csv	16 KB	Microsoft Excel Comma Sep
topMil.csv	23 KB	Microsoft Excel Comma Sep
TrancoTop-1m.csv	22,405 KB	Microsoft Excel Comma Sep
UniqueDomains - Copy.csv	16,579 KB	Microsoft Excel Comma Sep
UniqueDomains.csv	14,671 KB	Microsoft Excel Comma Sep

0 B of 112 MB in 0 of 44

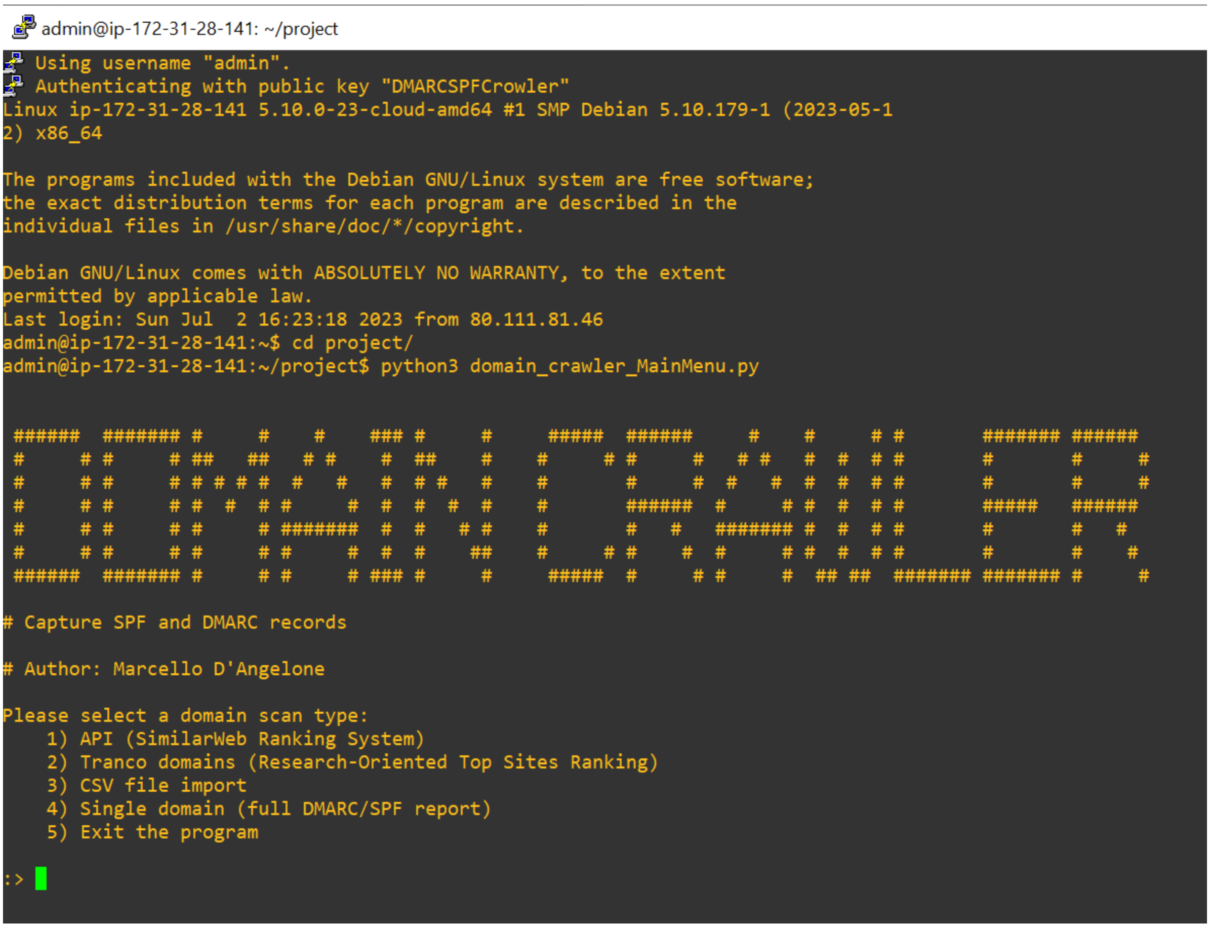
/home/admin/project/

Name	Size	Changed
..		7/1/2023 5:07
__pycache__		7/1/2023 5:55
bin		7/1/2023 5:09
include		7/1/2023 5:07
lib		7/1/2023 5:07
lib64		7/1/2023 5:07
share		7/1/2023 5:08
domain_crawler_MainMenu.py	4 KB	6/25/2023 12:
get_domain_list.py	2 KB	6/30/2023 3:0
get_spf_dmarc.py	8 KB	6/25/2023 12:
pyvenv.cfg	1 KB	7/1/2023 5:08
requirements.txt	1 KB	6/25/2023 1:0
top500Domains.csv	16 KB	4/22/2023 11:

0 B of 28.0 KB in 0 of 12

Figure 15: WinSCP file transfer

When executed from an SSH client connecting to the EC2 instance, the command *python3 domain_crawler_MainMenu.py* must be typed as depicted in Figure 16.

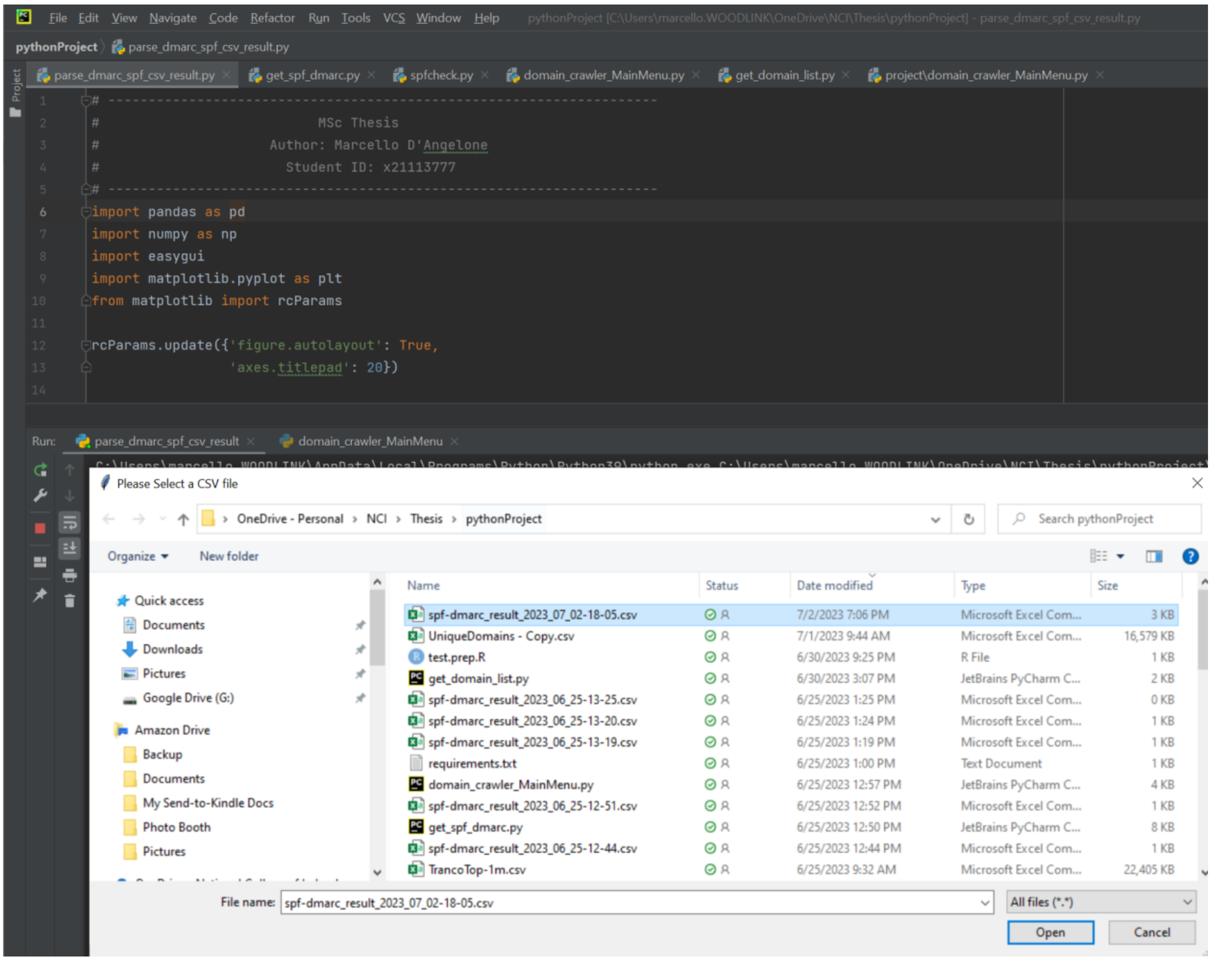


admin@ip-172-31-28-141: ~/project

```
 Using username "admin".
 Authenticating with public key "DMARCSPFCrowler"
Linux ip-172-31-28-141 5.10.0-23-cloud-amd64 #1 SMP Debian 5.10.179-1 (2023-05-1
2) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Sun Jul  2 16:23:18 2023 from 80.111.81.46
admin@ip-172-31-28-141:~$ cd project/
admin@ip-172-31-28-141:~/project$ python3 domain_crawler_MainMenu.py


 ######  ####### #     #    #    ### #     #     #####  ######     #    #     # #       ####### ######
 #     # #     # ##   ##   # #    #  ##    #    #     # #     #   # #   #  #  # #       #       #     #
 #     # #     # # # # #  #   #   #  # #   #    #       #     #  #   #  #  #  # #       #       #     #
 #     # #     # #  #  # #     #  #  #  #  #    #       ######  #     # #  #  # #       #####   ######
 #     # #     # #     # #######  #  #   # #    #       #   #   ####### #  #  # #       #       #   #
 #     # #     # #     # #     #  #  #    ##    #     # #    #  #     # #  #  # #       #       #    #
 ######  ####### #     # #     # ### #     #     #####  #     # #     #  ## ##  ####### ####### #     #

# Capture SPF and DMARC records

# Author: Marcello D'Angelone

Please select a domain scan type:
    1) API (SimilarWeb Ranking System)
    2) Tranco domains (Research-Oriented Top Sites Ranking)
    3) CSV file import
    4) Single domain (full DMARC/SPF report)
    5) Exit the program

:>
```

Figure 16: WinSCP file transfer

The output generated by selecting options one, two, or three consists of a CSV file, which can be imported by Excel, or the stand-alone parsing tool created for this project. Since the parsing tool requires a graphical user interface to display the pie chart, it will prompt to select the output generated by the domain crawler from an Explorer Window, as depicted in Figure 17.


File Edit View Navigate Code Refactor Run Tools VCS Window Help
pythonProject [C:\Users\marcello.WOODLINK\OneDrive\NCI\Thesis\pythonProject] - parse_dmarc_spf_csv_result.py
pythonProject 〉 parse_dmarc_spf_csv_result.py
parse_dmarc_spf_csv_result.py
get_spf_dmarc.py
spfcheck.py
domain_crawler_MainMenu.py
get_domain_list.py
project\domain_crawler_MainMenu.py
Project

```python
# --------------------------------------------------------------------
#                         MSc Thesis
#                   Author: Marcello D'Angelone
#                     Student ID: x21113777
# --------------------------------------------------------------------
import pandas as pd
import numpy as np
import easygui
import matplotlib.pyplot as plt
from matplotlib import rcParams

rcParams.update({'figure.autolayout': True,
                 'axes.titlepad': 20})
```

Run: parse_dmarc_spf_csv_result
domain_crawler_MainMenu

```
C:\Users\marcello.WOODLINK\AppData\Local\Programs\Python\Python39\python.exe C:\Users\marcello.WOODLINK\OneDrive\NCI\Thesis\pythonProject\
```

Name	Status	Date modified	Type	Size
spf-dmarc_result_2023_07_02-18-05.csv	✓	7/2/2023 7:06 PM	Microsoft Excel Com...	3 KB
UniqueDomains - Copy.csv	✓	7/1/2023 9:44 AM	Microsoft Excel Com...	16,579 KB
test.prep.R	✓	6/30/2023 9:25 PM	R File	1 KB
get_domain_list.py	✓	6/30/2023 3:07 PM	JetBrains PyCharm C...	2 KB
spf-dmarc_result_2023_06_25-13-25.csv	✓	6/25/2023 1:25 PM	Microsoft Excel Com...	0 KB
spf-dmarc_result_2023_06_25-13-20.csv	✓	6/25/2023 1:24 PM	Microsoft Excel Com...	1 KB
spf-dmarc_result_2023_06_25-13-19.csv	✓	6/25/2023 1:19 PM	Microsoft Excel Com...	1 KB
requirements.txt	✓	6/25/2023 1:00 PM	Text Document	1 KB
domain_crawler_MainMenu.py	✓	6/25/2023 12:57 PM	JetBrains PyCharm C...	4 KB
spf-dmarc_result_2023_06_25-12-51.csv	✓	6/25/2023 12:52 PM	Microsoft Excel Com...	1 KB
get_spf_dmarc.py	✓	6/25/2023 12:50 PM	JetBrains PyCharm C...	8 KB
spf-dmarc_result_2023_06_25-12-44.csv	✓	6/25/2023 12:44 PM	Microsoft Excel Com...	1 KB
TrancoTop-1m.csv	✓	6/25/2023 9:32 AM	Microsoft Excel Com...	22,405 KB

File name: spf-dmarc_result_2023_07_02-18-05.csv

All files (*.*)

Figure 17: DMARC and SPF CSV output parser

Once the CSV output file has been selected, and Opened via the Explorer Window, a pie chart will be displayed with the result as depicted in Figure 18.

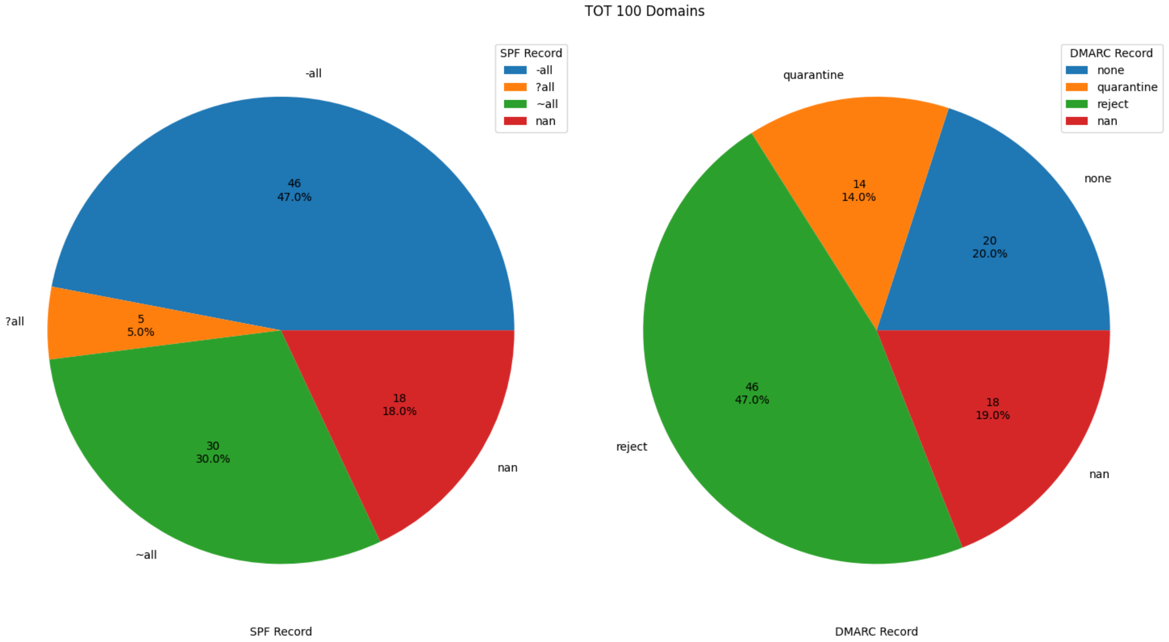


Figure 18: DMARC and SPF CSV output parser

The Title of the chart displays the total number of Domain parsed, while the pie chart on the left displays the total number of and the percentage of the domains with an SPF policy, while the pie chart on the right displays the total number and the percentage of the domains with a DMARC policy.

6 Statistical test

The statistical hypothesis compared one of the last published Alexa Top 1 Million measurements with the one gathered using the domain crawler tool to determine an increase in the anti-spoofing protocols adoption rate.
The Null hypothesis (H_0) stated there is no increase in adopting SPF and DMARC protocols. The alternative hypothesis (H_1) states that there is a significant increase. The confidence level interval was set 95% (alpha value of 0.05).

Because the two measurements have a different sample size, the 2-sample test for equality of proportions with continuity correction has been used with the data displayed in Table 11.

Table 4: Alexa Top 1 Million measurements comparison

	Total valid DMARC		**Total valid SPF**		**Total**
Tatang et al. (2021)	114,706	11.47%	503,310	50.33%	1,000,000
Domain Crawler	221,607	25.63%	515,164	59.59%	864,552

```
#DMARC measurement comparison between 2020 and 2023
> prop.test(c(221607, 114706), c(864552, 1000000))

        2-sample test for equality of proportions with continuity
correction

data:  c(221607, 114706) out of c(864552, 1e+06)
X-squared = 62903, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1405065 0.1427331
sample estimates:
   prop 1    prop 2 
0.2563258 0.1147060 

> 
> #SPF measurement comparison between 2020 and 2023
> prop.test(c(515164, 503310), c(864552, 1000000))

        2-sample test for equality of proportions with continuity
correction

data:  c(515164, 503310) out of c(864552, 1e+06)
X-squared = 16028, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.09113797 0.09398990
sample estimates:
   prop 1    prop 2 
0.5958739 0.5033100 
```