Improving Detection of Malicious URL's using Ensemble Learning Techniques

Configuration Manual

MSc Research Project
Cyber Security

Santosh Raj. Bingi
Student ID: x20123035

School of Computing
National College of Ireland

Supervisor: Michael Pantridge

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Santosh Raj Bingi ……………………………………………………………………………………………………

Student ID: X20123035 ……………………………………………………………………………………………………..……

Programme: M.S in Cyber Security ……………………………………………………… **Year:** 2020-2021 …………………………..

Module: Research Project ……………………………………………………………………………………………….………

Lecturer: Michael Pantridge ……………………………………………………………………………………………….………

Submission Due Date: 16th August 2021 ……………………………………………………………………………………………….………

Project Title: Improving Detection of Malicious URL's using Ensemble Learning Tech ……………………………………………………………………………………………….………

Word Count: …………………………………… **Page Count:** ………………………………….…….………

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: Santosh Raj. Bingi ………………………………………………………………………………………………………

Date: !6th August 2021 ………………………………………………………………………………………………………

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

Santosh Raj. Bingi
Student ID: x20123035

1 Pre Requisites

Hardware:

A Machine with Pentium Processor with multiple cores and hyperthreading enabled with minimum of 8 GB of RAM is required for being able to setup and

Software:

Operating System

Ubuntu 20

Application Software

Anaconda Jupyter Python Notebook – 2.04

Opensource Modules:

Pandas - Pandas is a library which can helps us with csv files and working with large datasets as through it were a spreadsheet. We are using pandas 1.1.5 version of the library. It is a very powerful, flexible and easy to use opensource data analysis and data manipulation tool for processing large data. This is very cumbersome for processing on a conventional

Numpy – Numpy is a library which is used for scientific number crunching. It is mostly used in conjunction with Pandas. However it can be used separately as well.

Requests: we use the python requests module to make http get API to capture the http_headers,

re – Python Regular Expression Module. Is used to pre process the required fields on the input dataset based on string patterns.

Urllib.parse – Used to Parse URL'S from the input dataset.

Plotly – Used to Generate plots of various kinds

Matplotlib – Used to generate graphs/plots of different kind.

Seaborn –

Time – Used to time the execution

Sklearn - Used to implement ML algorithms and Ensemble Machine Learning Algorithms

Lightgbm – used to implement light gradient boosting

Xgboost & LighGBM – Used to implement boosting alogruthms

Custom Modules:

DataEditor – parsing of the input csv into interesting feature columns and is a custom module for the purpose

2 Configuration & Implementation

2a. Create a project directory

$ mkdir url_scanner

2b. Copy the Source code tar ball into the project directory

$ cp malicious_url_analysis.tar.gz url_scanner

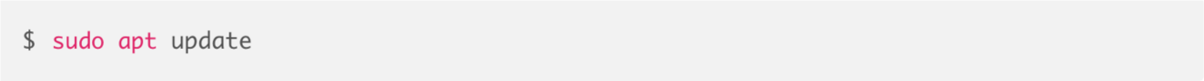
2c. Extract the code tar ball to the project directory

$tar - zavf malicious_url_analysis.tar.gz

2d Install Python Jupyter Notebook

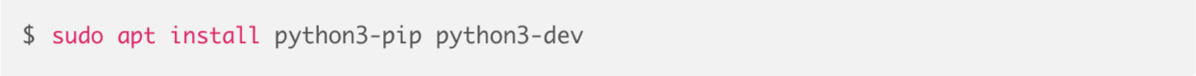
Setting up Python

- Update the package indexes for the apt repositories



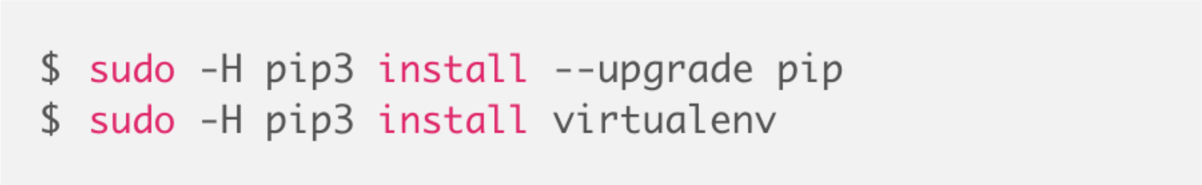
```
$ sudo apt update
```

- Install Jupyter dependencies (pip and PYTHON Dev packages)



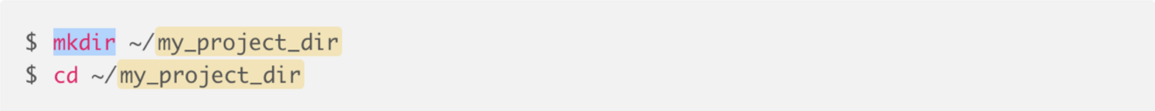
```
$ sudo apt install python3-pip python3-dev
```

- Create a virtualenv for the project



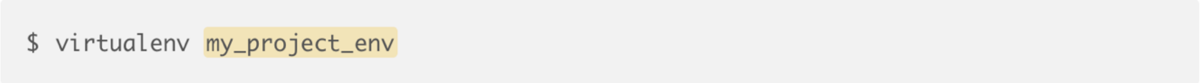
```
$ sudo -H pip3 install --upgrade pip
$ sudo -H pip3 install virtualenv
```

- Create the Project Directory Structure for hosting the local libraries



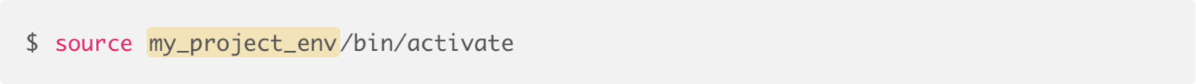
```
$ mkdir ~/my_project_dir
$ cd ~/my_project_dir
```

* Create the virtualenv for the Project



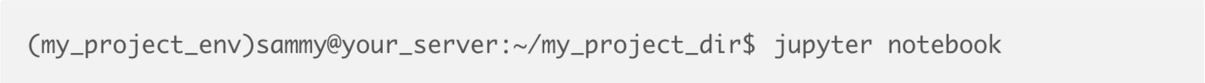
```
$ virtualenv my_project_env
```

* Activate the virtual Environment



```
$ source my_project_env/bin/activate
```

Install Jupyter Notebook



```
(my_project_env)sammy@your_server:~/my_project_dir$ jupyter notebook
```

Install the pre requisite libraries for the project. It is all accumulated in a single file called requirements.txt using the command

```
$cd ~/project_dir
$pip install -r requirements.txt
```

3. Validation

Start the Jupyter Notebook. You can replace the IP and the port to suit your purpose and that there is no conflicting port on another application

```
$jupyter notebook --no-browser --ip
<IP> --port=<PORT>
```

Open a browser and access the URL accordingly with the token from the link output on the screen

References

References should be formatted using APA or Harvard style as detailed in NCI Library Referencing Guide available at https://libguides.ncirl.ie/referencing
You can use a reference management system such as Zotero or Mendeley to cite in MS Word.

Tagliaferri, L. (no date) *How to Install, Run, and Connect to Jupyter Notebook | DigitalOcean.* Available at: https://www.digitalocean.com/community/tutorials/how-to-set-up-jupyter-notebook-with-python-3-on-ubuntu-20-04-and-connect-via-ssh-tunneling (Accessed: 16 August 2021).

Youtube – Presentation and recording Link: https://youtu.be/TqrSF3ONlcw
https://drive.google.com/drive/folders/13RrMTcBoixHD2OdKnXMLXczBNca0sUHw?usp=sharing