**README FILE**

**for**

**A Web Scraping of Attack.mitre website**

**PROJECT**

**By**

**Rohit Dhaipule**

Contents

[1. Getting Started 3](#_Toc44338668)

[2. The Website 4](#_Toc44338669)

[3. Displaying the Website 4](#_Toc44338670)

[4. Crawling a different website 4](#_Toc44338671)

[5. Extracting Mitigations, Detections and its keywords 4](#_Toc44338672)

[6. Creating a super user 5](#_Toc44338673)

[7. Adding data to website 5](#_Toc44338674)

# Getting Started

If you don’t have python3 installed in your pc then, install python3. You can refer to python documentation for installation steps. ( <https://docs.python.org/3/> ). Make sure python works from command prompt

Then install the following python libraries. The commands for installing them are given below. Run the commands in the command prompt to install the corresponding libraries.

* re library

*pip install regex*

* beautifulsoup library

*pip install beautifulsoup4*

* rake-nltk library

*pip install rake-nltk*

* xlsx-writer library

*pip install XlsxWriter*

You can refer to the documentations of those libraries for more information.

Now install Django framework in your system. It can be done by the command:

*pip install Django*

For further queries refer to Django documentation. (<https://www.djangoproject.com/download/> )

Now installation of following Django libraries is to be completed. The commands for installing them are given below. Run the commands in the command prompt to install the corresponding libraries.

* django pagination library

*pip install django-pagination*

* django import-export library

*pip install django-import-export*

# The Website

The website shows all the link, the detections, their keywords and the related mitigations. A search bar is also available which helps to filter the links based on certain keywords.

It also helps us to add and remove the bad key words which we want to filter.

# Displaying the Website

Extract the files from the “Web Scraping.zip” file. Open command prompt and navigate to detections folder in the extracted folder. Now run the following command in the command prompt which runs a local server.

*python manage.py runserver*

Now open <http://127.0.0.1:8000/> in any web browser. The webpage with all the technique links, their mitigations, detections and their keywords are shown.

# Crawling a different website

Open the url.txt file in crawler folder in the extracted folder. Replace the existing url with the url you need to crawl. Clear all the data in crawl.txt and queue.txt as they contain data of crawling of old link. If you need that data then save it at a place.

Now open the crawl.py file and run it. After the program has completed executing you will get all the crawled links in crawl.txt. You can stop the program at any time and you can run the program later. It will start from the same point it has stopped because all the discovered and crawled links have been saved to queue.txt and crawl.txt respectively.

# Extracting Mitigations, Detections and its keywords

First add all the bad words (separate each word by a comma) you want to filter from the detection in the bad\_words.txt file in techniques folder. Then open the techniques.py in the same folder and execute it. You will get all the links (in column C), the detections (in column D), their keywords (in column E) and the related mitigations (in column F) in the excel sheet “data.xlsx”.

# Creating a super user

Super user helps to add necessary data to the website through admin panel. Open the detections folder in command prompt. Enter the command given below.

*python manage.py createsuperuser*

Then it asks for username, email and password. Email is an optional option but the other two are required. After entering a super user is created.

# Adding data to website

First run the local server (command is provided in “Displaying the Website” section). Then open <http://127.0.0.1:8000/admin> in any web browser. Login with your credentials and a window appears as shown below.

Then click on “Add Badwords” in the homepage. A page with all the badwords appears. You can add or delete the badwords. After making suitable changes click on run to update the homepage.