

**TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ MÔN NHẬP MÔN BẢO MẬT
THÔNG TIN**

DETECT PHISHING WEBSITES USING MACHINE LEARNING

Người hướng dẫn: **TS HUỖNH NGỌC TÚ**

Người thực hiện: **NGUYỄN ĐÔNG HUY – 51800682**

TRẦN QUỐC TÂM – 51800721

NGUYỄN VĂN HUY - 51800783

Lớp : 18050402 - 18050203

Khoá : 22

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021

**TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KÌ MÔN NHẬP MÔN BẢO MẬT
THÔNG TIN**

DETECT PHISHING WEBSITES USING MACHINE LEARNING

Người hướng dẫn: **TS HUỖNH NGỌC TÚ**

Người thực hiện: **NGUYỄN ĐỒNG HUY – 51800682**

TRẦN QUỐC TÂM – 51800721

NGUYỄN VĂN HUY - 51800783

Lớp : 18050402 – 18050203

Khoá : 22

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021

LỜI CẢM ƠN

Trong suốt học kỳ I năm học 2021-2022 nhóm chúng em đã nhận được sự chỉ dẫn, quan tâm và giúp đỡ tận tình của các quý Thầy/Cô và bạn bè. Với lòng biết ơn sâu sắc và chân thành nhất, em xin gửi lời cảm ơn đến Thầy/Cô ở khoa Công Nghệ Thông Tin – Trường Đại học Tôn Đức Thắng đã giúp đỡ chúng em rất nhiều trong việc tiếp thu đầy đủ các kiến thức chuyên ngành.

Nhóm chúng em gửi lời cảm ơn đặc biệt đến cô Huỳnh Ngọc Tú – phụ trách giảng dạy chúng em môn Nhập môn Bảo mật thông tin. Cô đã tận tâm hướng dẫn chúng em qua từng buổi học trên lớp cũng như trao đổi thực tế với sinh viên, đưa ra các ví dụ minh họa rất gần gũi, dễ hình dung và giúp ghi nhớ lâu. Chúng em cũng xin bày tỏ lòng biết ơn đến các ban lãnh đạo của Trường Đại Học Tôn Đức Thắng và các khoa phòng ban chức năng, các bạn học đã trực tiếp và gián tiếp giúp đỡ chúng em trong suốt quá trình học tập và hoàn thành bài báo cáo này.

Với điều kiện về thời gian cũng như kinh nghiệm còn hạn chế của sinh viên, bài báo cáo này không thể tránh được những thiếu sót. Chúng em rất mong nhận được sự đóng góp ý kiến của các quý thầy cô để có cơ hội bổ sung thêm kiến thức và phục vụ tốt hơn cho các công tác thực tế sau này.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là sản phẩm đồ án của chúng tôi và được sự hướng dẫn của cô Huỳnh Ngọc Tú; Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất cứ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày tháng năm

Tác giả

(ký tên và ghi rõ họ tên)

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của GV hướng dẫn

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

Phần đánh giá của GV chấm bài

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

TÓM TẮT

Tấn công giả mạo là một cuộc tấn công phổ biến nhằm vào những người dùng Internet khiến họ tiết lộ thông tin của mình bằng cách sử dụng các trang web giả mạo. Mục tiêu của trang web giả mạo là đánh cắp thông tin cá nhân như username, mật khẩu và các giao dịch ngân hàng trực tuyến. Những kẻ lừa đảo sử dụng các trang web tương tự về mặt hình ảnh và ngữ nghĩa so với các trang web thực đó.

Khi công nghệ tiếp tục phát triển, các kỹ thuật lừa đảo bắt đầu tiến bộ nhanh chóng và điều này cần được ngăn chặn bằng cách sử dụng các cơ chế chống lừa đảo như phát hiện URL giả mạo. Học máy (Machine Learning) là một công cụ mạnh mẽ được sử dụng để chống lại các cuộc tấn công giả mạo. Bài báo cáo này đề cập đến công nghệ học máy để phát hiện các URL giả mạo bằng cách trích xuất và phân tích các đặc điểm khác nhau của các URL hợp pháp và giả mạo. Các thuật toán Random Forest, Logistic Regression và Support Vector Machine được sử dụng để phát hiện các trang web giả mạo.

MỤC LỤC

LỜI CẢM ƠN	i
PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN	iii
TÓM TẮT	iv
MỤC LỤC.....	1
DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT	3
DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ	4
CHƯƠNG 1 – CƠ SỞ LÝ THUYẾT	5
1.1 Tổng quan về Phishing.....	5
1.2 Các phương thức tấn công Phishing.....	6
1.2.1 Gửi email, tin nhắn	6
1.2.2 Giả mạo website.....	7
CHƯƠNG 2 – MÔ TẢ ĐỒ ÁN	8
2.1 Giới thiệu đồ án.....	8
2.2 Các đặc điểm nhận dạng trang web giả mạo của đồ án	8
2.2.1 Các đặc điểm dựa trên thanh địa chỉ.....	8
2.2.2 Các đặc điểm dựa trên bất thường.....	13
2.2.3 Các đặc điểm dựa trên HTML và JavaScript	16
2.2.4 Các đặc điểm dựa trên tên miền	18
2.3 Thuật toán sử dụng.....	20
2.3.1 Random Forest	20
2.3.2 Logistic Regression.....	22
2.3.3 Support Vector Machine	22
CHƯƠNG 3 – DEMO.....	24
3.1 Chức năng có trong demo	24

3.2 Chạy chương trình demo.....	24
3.2.1 Cài đặt thư viện Python.....	24
3.2.2 Chạy các thuật toán.....	31
3.2.3 Chạy chương trình demo trên trình duyệt web	33
3.2.3 Chạy chương trình demo trực tiếp trên Command Prompt	37
3.2.4 Hướng xử lý khi chạy chương trình bị lỗi hoặc không chạy được .	41
3.3 Luồng hoạt động của chương trình	42
3.4 Đánh giá kết quả đạt được	43
CHƯƠNG 4: KẾT LUẬN	46
TÀI LIỆU THAM KHẢO.....	47

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC KÝ HIỆU

CÁC CHỮ VIẾT TẮT

URL	Uniform Resource Locator
SVM	Support Vector Machine
HTTPS	Hyper Text Transfer Protocol with Secure Sockets Layer

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

DANH MỤC HÌNH

Hình 1. 1: Giả mạo email, tin nhắn	6
Hình 1. 2: Giả mạo website.....	7
Hình 2. 1: Thuật toán Random Forest sử dụng trong đồ án.....	21
Hình 2. 2: Thuật toán Logistic Regression sử dụng trong đồ án	22
Hình 2. 3: Thuật toán SVM sử dụng trong đồ án.....	23
Hình 3. 1: Cài đặt thư viện Requests.....	25
Hình 3. 2: Cài đặt thư viện Extract	25
Hình 3. 3: Cài đặt thư viện Tldextract.....	26
Hình 3. 4: Cài đặt thư viện BS4 và Whois	26
Hình 3. 5: Cài đặt thư viện Google	27
Hình 3. 6: Cài đặt thư viện Datetime và Parse	27
Hình 3. 7: Cài đặt thư viện Joblib và Pandas	28
Hình 3. 8: Cài đặt thư viện Print_dict và Flask.....	28
Hình 3. 9: Cài đặt thư viện Jsontify	29
Hình 3. 10: Cài đặt thư viện LXML.....	29
Hình 3. 11: Kiểm tra lại thư viện Python - 1.....	30
Hình 3. 11: Kiểm tra lại thư viện Python - 2.....	31

DANH MỤC BẢNG

Bảng 2. 1: Các công thông thường cần kiểm tra.....	13
--	----

CHƯƠNG 1 – CƠ SỞ LÝ THUYẾT

Ngày nay, Internet đóng vai trò quan trọng trong giao tiếp, nơi mọi người tạo ra môi trường trực tuyến để quản lý chức năng kinh doanh, các hoạt động trực tuyến của ngân hàng, mạng xã hội... Tuy nhiên, Internet cũng ẩn chứa rất nhiều rủi ro bởi vì khi người dùng hoạt động trong môi trường trực tuyến họ có thể dễ bị tấn công bởi các attacker. Và đặc điểm nhận diện của chúng thường là một URL giả mạo. Và các URL giả mạo thường được đặt trên các trang web phổ biến hoặc được gửi đến email người dùng.

1.1 Tổng quan về Phishing

Hiện nay, Phishing (tấn công giả mạo) trở thành một lĩnh vực quan tâm chính của các nhà nghiên cứu bảo mật vì không khó để tạo ra một trang web giả mạo trông giống với trang web hợp pháp. Khi người dùng nhấp vào liên kết web, nó sẽ hướng người dùng đến máy chủ của kẻ tấn công thay vì máy chủ web thực. Các chuyên gia có thể xác định các trang web giả mạo nhưng không phải tất cả người dùng đều có thể xác định được trang web giả mạo và những người dùng đó trở thành nạn nhân của cuộc tấn công giả mạo. Mục đích chính của kẻ tấn công là đánh cắp thông tin đăng nhập tài khoản ngân hàng. Các cuộc tấn công giả mạo đang trở nên phổ biến vì sự thiếu nhận thức của người dùng. Vì tấn công giả mạo khai thác những điểm yếu được tìm thấy ở người dùng nên rất khó để giảm thiểu chúng. Nhưng việc tăng cường các kỹ thuật phát hiện giả mạo là rất quan trọng.

Phương pháp chung để phát hiện các trang web lừa đảo bằng cách cập nhật các URL trong danh sách đen, Giao thức Internet (IP) vào cơ sở dữ liệu chống vi-rút còn được gọi là phương pháp "danh sách đen". Để tránh danh sách đen, kẻ tấn công sử dụng các kỹ thuật sáng tạo để đánh lừa người dùng bằng cách sửa đổi URL để có vẻ hợp pháp thông qua sự xáo trộn và nhiều kỹ thuật đơn giản khác bao gồm: fast-flux, trong đó proxy được tạo tự động để lưu trữ trang web; tạo URL mới theo thuật toán...

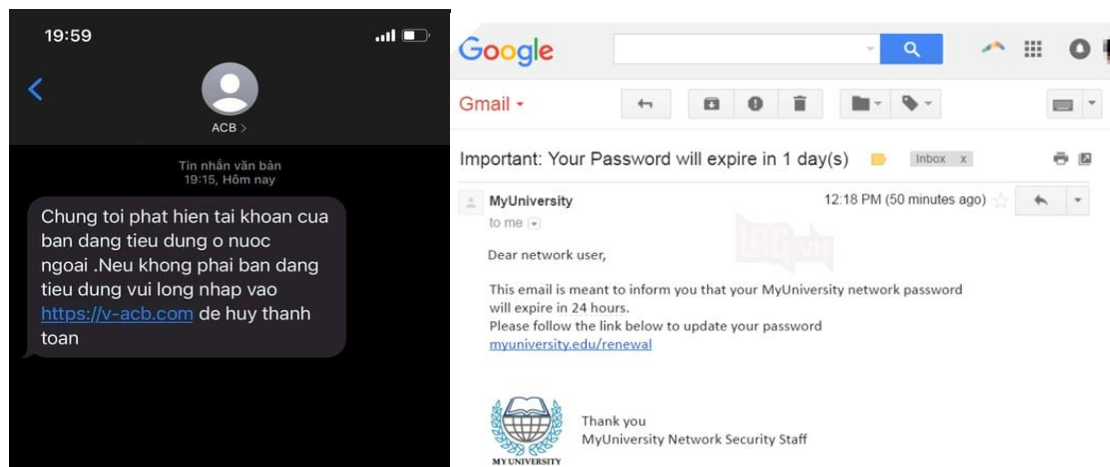
Để khắc phục những hạn chế của phương pháp dựa trên danh sách đen, nhiều nhà nghiên cứu bảo mật hiện tập trung vào các kỹ thuật học máy (machine learning). Công nghệ học máy bao gồm nhiều thuật toán yêu cầu dữ liệu trong quá khứ để đưa ra quyết định hoặc dự đoán về dữ liệu trong tương lai. Sử dụng kỹ thuật này, thuật toán sẽ phân tích các URL hợp pháp và nằm trong danh sách đen khác nhau dựa trên các đặc điểm của chúng để phát hiện chính xác các trang web giả mạo.

1.2 Các phương thức tấn công Phishing

1.2.1 Gửi email, tin nhắn

Đặc điểm nhận dạng:

- Địa chỉ người gửi không đáng tin cậy.
- Thường sở hữu các cụm từ liên quan đến xác minh tài khoản người dùng.
- Tồn tại link hướng người dùng đến trang web giả mạo.
- Yêu cầu người dùng cung cấp các thông tin nhạy cảm.
- Không đề cập đến người nhận cụ thể.
- Xuất hiện lỗi chính tả.
- Đính kèm tệp độc hại.

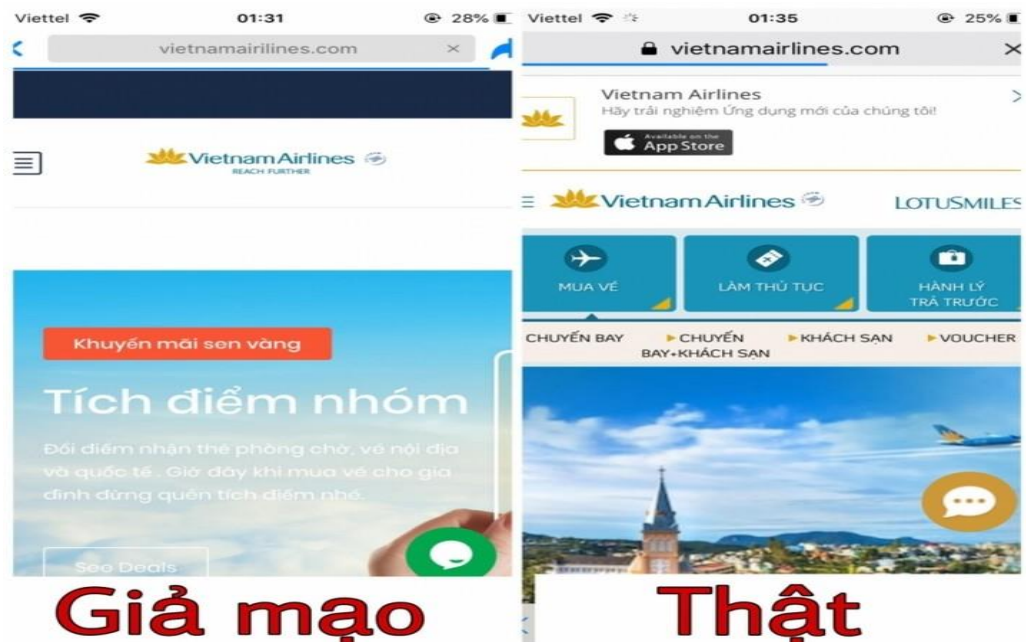


Hình 1. 1: Giả mạo email, tin nhắn

1.2.2 Giả mạo website

Đặc điểm nhận dạng:

- Thường sử dụng giao thức không an toàn (không sử dụng https) và có domain không đáng tin cậy.
- Đường link khác 1 hoặc một vài kí tự.
- Không có đầy đủ các chức năng (Thường chỉ là một trang landing page).
- Luôn có những thông điệp khuyến khích người dùng nhập thông tin cá nhân vào Website.
- Sử dụng nền tảng tạo web miễn phí.



Hình 1. 2: Giả mạo website

CHƯƠNG 2 – MÔ TẢ ĐỒ ÁN

2.1 Giới thiệu đồ án

Nhóm chúng em đã phát triển đồ án của mình bằng cách sử dụng một trang web làm nền tảng cho tất cả người dùng. Đây là một trang web tương tác và đáp ứng được sử dụng để phát hiện xem một trang web là hợp pháp hay giả mạo. Trang web này được tạo bằng các ngôn ngữ thiết kế web khác nhau bao gồm HTML, CSS, Javascript và Python.

Trang web hiển thị thông tin liên quan đến các dịch vụ do nhóm chúng em cung cấp. Nó cũng chứa thông tin liên quan đến những sai lầm xảy ra trong thế giới công nghệ ngày nay. Trang web được tạo ra với một ý tưởng để mọi người không chỉ có thể phân biệt giữa trang web hợp pháp và trang web giả mạo, mà còn nhận thức được các hành vi xấu đang xảy ra trong thế giới hiện tại. Từ đó có thể tránh xa những kẻ tấn công đang cố gắng khai thác thông tin cá nhân như: địa chỉ email, mật khẩu, chi tiết thẻ tín dụng, số tài khoản ngân hàng...

Tập dữ liệu của đồ án nhóm em bao gồm các đặc điểm khác nhau cần được xem xét khi xác định URL là hợp pháp hay giả mạo.

2.2 Các đặc điểm nhận dạng trang web giả mạo của đồ án

Phần này mô tả mô hình phát hiện tấn công giả mạo được đề xuất. Mô hình tập trung vào việc xác định cuộc tấn công giả mạo dựa trên việc kiểm tra các đặc điểm của các trang web giả mạo và cơ sở dữ liệu danh sách đen. Theo công cụ đề xuất của nhóm chúng em, một số đặc điểm được chọn có thể được sử dụng để phân biệt giữa các trang web hợp pháp và giả mạo. Những đặc điểm này bao gồm: các đặc điểm dựa trên thanh địa chỉ, các đặc điểm dựa trên bất thường, các đặc điểm dựa trên HTML và JavaScript và các đặc điểm dựa trên tên miền.

2.2.1 Các đặc điểm dựa trên thanh địa chỉ

2.2.1.1 Sử dụng địa chỉ IP

Nếu địa chỉ IP được sử dụng thay thế cho tên miền trong URL, chẳng hạn như “http://125.98.3.123/fake.html”, người dùng có thể chắc chắn rằng ai đó đang cố lấy cắp thông tin cá nhân của họ. Đôi khi, địa chỉ IP thậm chí còn được chuyển đổi thành mã thập lục phân như được hiển thị trong liên kết sau:

“http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html”

Quy tắc:

- Nếu tên miền có địa chỉ IP => Phishing (giả mạo)
- Nếu không => Legitimate (hợp pháp)

2.2.1.2 URL dài để ẩn phần đáng ngờ

Những kẻ lừa đảo có thể sử dụng URL dài để ẩn phần đáng ngờ trong thanh địa chỉ. Ví dụ:

http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html

Để đảm bảo tính chính xác của nghiên cứu, nhóm chúng em đã tính toán độ dài của các URL trong tập dữ liệu và đưa ra độ dài URL trung bình. Kết quả cho thấy rằng nếu độ dài của URL lớn hơn hoặc bằng 54 ký tự thì URL đó được phân loại là giả mạo. Bằng cách xem xét tập dữ liệu của mình, nhóm chúng em có thể tìm thấy 1220 độ dài URL bằng 54 trở lên, chiếm 48,8% tổng kích thước tập dữ liệu.

Quy tắc:

- Độ dài URL < 54 => Legitimate
- URL $\geq 54 \leq 75$ => Phishing

2.2.1.3 Sử dụng dịch vụ rút ngắn URL “TinyURL”

Rút ngắn URL là một phương pháp trên “World Wide Web” trong đó URL có thể được tạo với độ dài nhỏ hơn đáng kể và vẫn dẫn đến trang web được yêu cầu. Điều này được thực hiện nhờ “Chuyển hướng HTTP” trên một tên miền ngắn, liên kết đến trang

web có URL dài. Ví dụ: URL “http://portal.hud.ac.uk/” có thể được rút ngắn thành “bit.ly/19DXSk4”.

Quy tắc:

- TinyURL => Phishing
- Nếu không => Legitimate

2.2.1.4 URL có biểu tượng “@”

Việc sử dụng biểu tượng “@” trong URL khiến trình duyệt bỏ qua mọi thứ đứng trước biểu tượng “@” và địa chỉ thực thường đứng sau biểu tượng “@”.

Quy tắc:

- URL có ký hiệu @ => Phishing
- Nếu không => Legitimate

2.2.1.5 Chuyển hướng bằng “//”

Sự tồn tại của “//” trong đường dẫn URL có nghĩa là người dùng sẽ được chuyển hướng đến một trang web khác. Ví dụ về URL như vậy là:

“http://www.legitimate.com//http://www.phishing.com”

Chúng em kiểm tra vị trí mà “//” xuất hiện. Chúng em nhận thấy rằng nếu URL bắt đầu bằng “HTTP”, điều đó có nghĩa là “//” sẽ xuất hiện ở vị trí thứ sáu. Tuy nhiên, nếu URL sử dụng “HTTPS” thì “//” sẽ xuất hiện ở vị trí thứ bảy.

Quy tắc:

- Vị trí xuất hiện lần cuối của “//” trong URL > 7 => Phishing
- Nếu không => Legitimate

2.2.1.6 Thêm tiền tố hoặc hậu tố được phân tách bằng (-) vào miền

Biểu tượng gạch ngang hiểm khi được sử dụng trong các URL hợp pháp. Những kẻ lừa đảo có xu hướng thêm tiền tố hoặc hậu tố được phân tách bằng (-) vào tên miền

để người dùng cảm thấy rằng họ đang xử lý một trang web hợp pháp. Ví dụ: <http://www.Confirme-paypal.com/>.

Quy tắc:

- Phần tên miền bao gồm ký hiệu (-) => Phishing
- Nếu không => Legitimate

2.2.1.7 Miền phụ và nhiều miền phụ

Giả sử chúng em có liên kết sau: <http://www.hud.ac.uk/students/>. Tên miền có thể bao gồm các tên miền cấp cao nhất mã quốc gia (ccTLD), trong ví dụ của chúng em là “uk”. Phần “ac” là viết tắt của “học thuật”, “ac.uk” kết hợp được gọi là miền cấp hai (SLD) và “hud” là tên thực của miền. Để đưa ra quy tắc trích xuất đặc điểm này, trước tiên phải bỏ qua (www.) khỏi URL mà trên thực tế, bản thân nó là một miền phụ. Sau đó, chúng ta phải xóa (ccTLD) nếu nó tồn tại. Cuối cùng, chúng ta đếm số chấm còn lại. Nếu số lượng dấu chấm lớn hơn một, thì URL được phân loại là "Suspicious" (đáng ngờ) vì nó có một tên miền phụ. Tuy nhiên, nếu các dấu chấm lớn hơn hai, nó được phân loại là "Phishing" vì nó sẽ có nhiều miền phụ. Nếu URL không có miền phụ, chúng em sẽ chỉ định "Legitimate" cho đối tượng.

Quy tắc:

- Dots In Domain Part = 1 => Legitimate
- Dots In Domain Part = 2 => Suspicious
- Nếu không => Phishing

2.2.1.8 HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)

Sự tồn tại của HTTPS rất quan trọng trong việc tạo ấn tượng về tính hợp pháp của trang web, nhưng điều này rõ ràng là chưa đủ. Các tác giả trong (Mohammad, Thabtah và McCluskey 2012) (Mohammad, Thabtah và McCluskey 2013) đề xuất kiểm tra chứng chỉ được chỉ định với HTTPS bao gồm phạm vi của tổ chức phát hành chứng chỉ tin cậy và tuổi chứng chỉ. Tổ chức phát hành chứng chỉ luôn được liệt kê trong số những cái tên

đáng tin cậy hàng đầu bao gồm: “GeoTrust, GoDaddy, Network Solutions, Thawte, Comodo, Doster và VeriSign”. Hơn nữa, bằng cách kiểm tra bộ dữ liệu của chúng em, chúng em thấy rằng độ tuổi tối thiểu của một chứng chỉ có uy tín là hai năm.

Quy tắc:

- Sử dụng https và Nhà phát hành được tin cậy và Tuổi chứng chỉ ≥ 1 năm => Legitimate
- Sử dụng https và Nhà phát hành không đáng tin cậy => Suspicious
- Nếu không => Phishing

2.2.1.9 Thời hạn đăng ký tên miền

Dựa trên thực tế là một trang web giả mạo tồn tại trong một khoảng thời gian ngắn, chúng em tin rằng các miền đáng tin cậy thường được thanh toán trước vài năm. Trong tập dữ liệu của mình, chúng em thấy rằng các miền gian lận lâu nhất chỉ được sử dụng trong một năm.

Quy tắc:

- Tên miền hết hạn sau ≤ 1 năm => Phishing
- Nếu không => Legitimate

2.2.1.10 Favicon

Favicon là một hình ảnh (biểu tượng) đồ họa được liên kết với một trang web cụ thể. Nhiều tác nhân người dùng hiện tại như trình duyệt đồ họa và trình đọc tin tức hiển thị favicon như một lời nhắc nhở trực quan về danh tính trang web trong thanh địa chỉ. Nếu favicon được tải từ một miền khác với miền được hiển thị trên thanh địa chỉ, thì trang web có thể bị coi là Phishing.

Quy tắc:

- Favicon được tải từ tên miền bên ngoài => Phishing
- Nếu không => Legitimate

2.2.1.11 Sử dụng cổng không chuẩn

Đặc điểm này hữu ích trong việc xác thực nếu một dịch vụ cụ thể (ví dụ: HTTP) lên hoặc xuống trên một máy chủ cụ thể. Với mục đích kiểm soát các cuộc xâm nhập, sẽ tốt hơn nhiều nếu chỉ mở các cổng mà bạn cần. Theo mặc định, một số tường lửa, proxy và máy chủ NAT (Network Address Translation) sẽ chặn tất cả hoặc hầu hết các cổng và chỉ mở những cổng đã chọn. Nếu tất cả các cổng đều mở, những kẻ lừa đảo có thể chạy hầu hết mọi dịch vụ mà chúng muốn và kết quả là thông tin người dùng bị đe dọa. Các cổng quan trọng nhất và trạng thái ưu tiên của chúng được thể hiện trong Bảng 2.1.

Quy tắc:

- Cổng có Trạng thái Ưu tiên (Preferred Status) => Phishing
- Nếu không => Legitimate

Bảng 2. 1: Các cổng thông thường cần kiểm tra

PORT	Service	Meaning	Preferred Status
21	FTP	Transfer files from one host to another	Close
22	SSH	Secure File Transfer Protocol	Close
23	Telnet	provide a bidirectional interactive text-oriented communication	Close
80	HTTP	Hyper text transfer protocol	Open
443	HTTPS	Hypertext transfer protocol secured	Open
445	SMB	Providing shared access to files, printers, serial ports	Close
1433	MSSQL	Store and retrieve data as requested by other software applications	Close
1521	ORACLE	Access oracle database from web.	Close
3306	MySQL	Access MySQL database from web.	Close
3389	Remote Desktop	allow remote access and remote collaboration	Close

2.2.1.12 Sự tồn tại của mã thông báo “HTTPS” trong phần miền của URL

Những kẻ lừa đảo có thể thêm mã thông báo “HTTPS” vào phần miền của URL để lừa người dùng. Ví dụ:

<http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/>

Quy tắc:

- Sử dụng mã thông báo HTTPS trong phần tên miền của URL => Phishing
- Nếu không => Legitimate

2.2.2 Các đặc điểm dựa trên bất thường

2.2.2.1 URL yêu cầu (Request URL)

URL yêu cầu kiểm tra xem các đối tượng bên ngoài có trong trang web như: hình ảnh, video và âm thanh có được tải từ một miền khác hay không. Trong các trang web hợp pháp, địa chỉ trang web và hầu hết các đối tượng được nhúng trong trang web đang chia sẻ cùng một miền.

Quy tắc:

- % of Request URL < 22% → Legitimate
- % of Request URL \geq 22% and 61% → Suspicious
- Nếu không => Phishing

2.2.2.2 URL of Anchor

Anchor là một phần tử được xác định bởi thẻ <a>. Đặc điểm này được coi chính xác là "Request URL".

Tuy nhiên, đối với đặc điểm này, chúng em kiểm tra:

1. Nếu thẻ <a> và trang web có tên miền khác nhau. Điều này tương tự như đặc điểm Request URL.
2. Nếu Anchor không liên kết đến bất kỳ trang web nào, ví dụ:
 -
 -
 -
 -

Quy tắc:

- % of URL of Anchor < 31% => Legitimate
- % of URL Of Anchor \geq 31% And \leq 67% => Suspicious
- Nếu không => Phishing

2.2.2.3 Các liên kết trong thẻ <Meta>, <Script> và <Link>

Cuộc điều tra của chúng em bao gồm tất cả các góc độ có khả năng được sử dụng = source code trang web, chúng em thấy rằng các trang web hợp pháp thường sử dụng <Meta>tags để cung cấp siêu dữ liệu về tài liệu HTML; <Script>tags để tạo một tập lệnh phía khách hàng; và <Link>tags để lấy các tài nguyên web khác. Dự kiến các thẻ này được liên kết với cùng một tên miền của trang web.

Quy tắc:

- % Liên kết trong web " < Meta > "," < Script > " và " < Link>" < 17% => Legitimate
- % Liên kết trong web < Meta > "," < Script > " và " < Link>" $\geq 17\%$ và $\leq 81\%$ => Suspicious
- Nếu không => Phishing

2.2.2.4 Server Form Handler (SFH)

Các SFH có chứa chuỗi trống hoặc “about: blank” được coi là đáng nghi ngờ vì hành động đối với thông tin đã gửi. Ngoài ra, nếu tên miền trong SFH khác với tên miền của trang web, điều này cho thấy rằng trang web đáng ngờ vì thông tin đã gửi hiếm khi được xử lý bởi các miền bên ngoài.

Quy tắc:

- SFH là "about: blank" hoặc là Empty => Phishing
- SFH đề cập đến một miền khác => Suspicious
- Nếu không => Legitimate

2.2.2.5 Gửi thông tin đến email

Biểu mẫu web cho phép người dùng gửi thông tin cá nhân của mình được chuyển hướng đến máy chủ để xử lý. Kẻ lừa đảo có thể chuyển hướng thông tin của người dùng đến email cá nhân của hắn. Vì vậy, ngôn ngữ script phía máy chủ có thể được sử dụng, chẳng hạn như hàm “mail ()” trong PHP. Một chức năng phía máy khách khác có thể được sử dụng cho mục đích này là chức năng “mailto:”.

Quy tắc:

- Sử dụng "mail ()" hoặc "mailto:" Chức năng Gửi Thông tin Người dùng => Phishing
- Nếu không => Legitimate

2.2.2.6 URL bất thường

Đặc điểm này có thể được trích xuất từ cơ sở dữ liệu WHOIS. Đối với một trang web hợp pháp, danh tính thường là một phần của URL của nó.

Quy tắc:

- Tên máy chủ không được bao gồm trong URL => Phishing
- Nếu không => Legitimate

2.2.3 Các đặc điểm dựa trên HTML và JavaScript

2.2.3.1 Chuyển tiếp trang web

Điểm mấu chốt để phân biệt các trang web giả mạo với các trang web hợp pháp là số lần một trang web đã được chuyển hướng. Trong tập dữ liệu của chúng em, chúng em thấy rằng các trang web hợp pháp đã được chuyển hướng tối đa một lần. Mặt khác, các trang web giả mạo chứa đặc điểm này đã bị chuyển hướng ít nhất 4 lần.

Quy tắc:

- Chuyển hướng trang web ≤ 1 => Legitimate
- Chuyển hướng trang web ≥ 2 và < 4 => Suspicious
- Nếu không => Phishing

2.2.3.2 Tùy chỉnh thanh trạng thái

Những kẻ lừa đảo có thể sử dụng JavaScript để hiển thị URL giả trên thanh trạng thái cho người dùng. Để trích xuất đặc điểm này, chúng ta phải tìm hiểu mã nguồn của trang web, đặc biệt là sự kiện “onMouseOver” và kiểm tra xem nó có thực hiện bất kỳ thay đổi nào trên thanh trạng thái hay không.

Quy tắc:

- Thanh trạng thái onMouseOver Changes => Phishing
- Nó không thay đổi thanh trạng thái => Legitimate

2.2.3.3 Vô hiệu hóa Nhấp chuột phải

Những kẻ lừa đảo sử dụng JavaScript để tắt chức năng nhấp chuột phải, để người dùng không thể xem và lưu source code trang web. Đặc điểm này được coi chính xác là “Sử dụng onMouseOver để ẩn Liên kết”. Tuy nhiên, đối với đặc điểm này, chúng em sẽ tìm kiếm sự kiện “event.button == 2” trong mã nguồn của trang web và kiểm tra xem nhấp chuột phải có bị tắt hay không.

Quy tắc:

- Nhấp chuột phải bị vô hiệu hóa => Phishing
- Nếu không => Legitimate

2.2.3.4 Sử dụng cửa sổ pop-up

Thật bất thường khi tìm thấy một trang web hợp pháp yêu cầu người dùng gửi thông tin cá nhân của họ thông qua một cửa sổ pop-up. Mặt khác, đặc điểm này đã được sử dụng trong một số trang web hợp pháp và mục tiêu chính của nó là cảnh báo người dùng về các hoạt động gian lận hoặc phát đi thông báo chào mừng, mặc dù không có thông tin cá nhân nào được yêu cầu điền thông qua các cửa sổ bật lên này.

Quy tắc:

- Cửa sổ pop-up Chứa các Trường Văn bản => Phishing
- Nếu không => Legitimate

2.2.3.5 Chuyển hướng IFrame

IFrame là một thẻ HTML được sử dụng để hiển thị một trang web bổ sung thành một trang web hiện đang được hiển thị. Những kẻ lừa đảo có thể sử dụng thẻ “iframe”

và làm cho nó ẩn đi, tức là không có Frame Border. Về vấn đề này, những kẻ lừa đảo sử dụng thuộc tính “frameBorder” khiến trình duyệt hiển thị mô tả trực quan.

Quy tắc:

- Sử dụng Iframe => Phishing
- Nếu không => Legitimate

2.2.4 Các đặc điểm dựa trên tên miền

2.2.4.1 Tuổi miền (Age of Domain)

Đặc điểm này có thể được trích xuất từ cơ sở dữ liệu WHOIS (Whois 2005). Hầu hết các trang web giả mạo đều tồn tại trong một khoảng thời gian ngắn. Bằng cách xem xét tập dữ liệu của mình, chúng em thấy rằng tuổi tối thiểu của miền hợp pháp là 6 tháng.

Quy tắc:

- Tuổi miền ≥ 6 tháng => Legitimate
- Nếu không => Phishing

2.2.4.2 Bản ghi DNS (DNS Record)

Đối với các trang web giả mạo, danh tính xác nhận quyền sở hữu không được cơ sở dữ liệu WHOIS công nhận (Whois 2005) hoặc không có hồ sơ nào được thiết lập cho tên máy chủ (Pan và Ding 2006). Nếu bản ghi DNS trống hoặc không được tìm thấy thì trang web được phân loại là "Phishing", nếu không nó được phân loại là "Legitimate".

Quy tắc:

- Không có Bản ghi DNS cho Tên miền => Phishing
- Nếu không => Legitimate

2.2.4.3 Lưu lượng truy cập trang web (Website Traffic)

Đặc điểm này đo lường mức độ phổ biến của trang web bằng cách xác định số lượng người truy cập và số trang họ truy cập. Tuy nhiên, vì các trang web giả mạo tồn tại trong một khoảng thời gian ngắn, chúng có thể không được cơ sở dữ liệu Alexa nhận

dạng (Alexa the Web Information Company, 1996). Bằng cách xem xét tập dữ liệu của mình, chúng em thấy rằng trong các tình huống tội tộ nhất, các trang web hợp pháp được xếp hạng trong số 100.000 trang hàng đầu. Hơn nữa, nếu miền không có lưu lượng truy cập hoặc không được cơ sở dữ liệu Alexa nhận dạng, nó được phân loại là "Phishing". Nếu không, nó được phân loại là "Suspicious".

Quy tắc:

- Website Rank < 100,000 => Legitimate
- Website Rank > 100,000 => Suspicious
- Nếu không => Phishing

2.2.4.4 Xếp hạng trang (PageRank)

PageRank là một giá trị nằm trong khoảng từ “0” đến “1”. PageRank nhằm mục đích đo lường mức độ quan trọng của một trang web trên Internet. Giá trị PageRank càng lớn thì trang web càng quan trọng. Trong bộ dữ liệu của mình, chúng em thấy rằng khoảng 95% trang web giả mạo không có PageRank. Hơn nữa, chúng em nhận thấy rằng 5% trang web giả mạo còn lại có thể đạt giá trị PageRank lên đến “0,2”.

Quy tắc:

- PageRank < 0.2 => Phishing
- Nếu không => Legitimate

2.2.4.5 Chỉ mục của Google (Google Index)

Đặc điểm này kiểm tra xem một trang web có trong chỉ mục của Google hay không. Khi một trang web được Google lập chỉ mục, nó sẽ được hiển thị trên kết quả tìm kiếm (Webmaster resources, 2014). Thông thường, các trang web giả mạo chỉ có thể truy cập được trong một thời gian ngắn và do đó, nhiều trang lừa đảo có thể không được tìm thấy trong chỉ mục của Google.

Quy tắc:

- Trang web được Google lập chỉ mục => Legitimate

- Nếu không => Phishing

2.2.4.6 Số lượng liên kết trở đến trang

Số lượng liên kết trở đến trang web cho biết mức độ hợp pháp của nó, ngay cả khi một số liên kết có cùng tên miền (Dean, 2014). Trong tập dữ liệu của chúng em, chúng em thấy rằng 98% các mục tập dữ liệu giả mạo không có liên kết trở đến chúng. Mặt khác, các trang web hợp pháp có ít nhất 2 liên kết bên ngoài trở đến chúng.

Quy tắc:

- Liên kết Trở đến Trang web = 0 => Phishing
- Liên kết Trở đến Trang web > 0 và ≤ 2 => Suspicious
- Nếu không => Legitimate

2.2.4.7 Đặc điểm dựa trên báo cáo thống kê

Một số bên như PhishTank (PhishTank Stats, 2010-2012) và StopBadware (StopBadware, 2010-2012) lập nhiều báo cáo thống kê về các trang web giả mạo tại mọi khoảng thời gian nhất định; một số là hàng tháng và một số khác là hàng quý. Trong nghiên cứu của chúng em, chúng em đã sử dụng 2 dạng của mười thống kê hàng đầu từ PhishTank: “Top 10 Domains” và “Top 10 IPs” theo báo cáo thống kê được công bố trong ba năm, bắt đầu từ tháng 1 năm 2010 đến tháng 11 năm 2012. Trong khi đối với “StopBadware”, chúng em đã sử dụng “Top 50”.

Quy tắc:

- Máy chủ lưu trữ thuộc các IP lừa đảo hàng đầu hoặc các miền lừa đảo hàng đầu => Phishing
- Nếu không => Legitimate

2.3 Thuật toán sử dụng

2.3.1 *Random Forest*

Random Forest (Rừng ngẫu nhiên) là một thuật toán học có giám sát. Như tên gọi của nó, Rừng ngẫu nhiên sử dụng các cây (tree) để làm nền tảng. Rừng ngẫu nhiên là một tập hợp của các Decision Tree, mà mỗi cây được chọn theo một thuật toán dựa vào ngẫu nhiên.

Decision Tree là tên đại diện cho một nhóm thuật toán phát triển dựa trên Cây quyết định. Ở đó, mỗi Node của cây sẽ là các thuộc tính, và các nhánh là giá trị lựa chọn của thuộc tính đó. Bằng cách đi theo các giá trị thuộc tính trên cây, Cây quyết định sẽ cho ta biết giá trị dự đoán. Nhóm thuật toán cây quyết định có một điểm mạnh đó là có thể sử dụng cho cả bài toán Phân loại (Classification) và Hồi quy (Regression).

Random Forest hoạt động bằng cách đánh giá nhiều Cây quyết định ngẫu nhiên, và lấy ra kết quả được đánh giá tốt nhất trong số kết quả trả về.

```
#splitting the dataset into training set and test set
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size = 0.25, random_state =0 )

#-----applying grid search to find best performing parameters
from sklearn.model_selection import GridSearchCV
parameters = [{'n_estimators': [100, 700],
                'max_features': ['sqrt', 'log2'],
                'criterion' :['gini', 'entropy']}]

grid_search = GridSearchCV(RandomForestClassifier(), parameters,cv =5, n_jobs=-1)
grid_search.fit(x_train, y_train)
#printing best parameters
print("Best Accuracy =" +str( grid_search.best_score_))
print("best parameters =" + str(grid_search.best_params_))
#-----

#fitting RandomForest regression with best params
classifier = RandomForestClassifier(n_estimators = 100, criterion = "gini", max_features = 'log2', random_state = 0)
classifier.fit(x_train, y_train)

#predicting the tests set result
y_pred = classifier.predict(x_test)

#confusion matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)

#pickle file joblib
joblib.dump(classifier, '../final_models/rf_final.pkl')
```

Hình 2. 1: Thuật toán Random Forest sử dụng trong đồ án

2.3.2 Logistic Regression

Logistic Regression (Hồi quy logistic) là một thuật toán đơn giản nhưng lại rất hiệu quả trong bài toán phân loại (Classification). Hồi quy logistic là một phương pháp phân tích thống kê được sử dụng để dự đoán giá trị dữ liệu dựa trên các quan sát trước đó của tập dữ liệu. Mục đích của hồi quy logistic là ước tính xác suất của các sự kiện, bao gồm xác định mối quan hệ giữa các tính năng, từ đó dự đoán xác suất của các kết quả.

```
#splitting the dataset into training set and test set
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size = 0.25, random_state = 0 )

#fitting logistic regression
classifier = LogisticRegression(random_state = 0)
classifier.fit(x_train, y_train)

#predicting the tests set result
y_pred = classifier.predict(x_test)

#confusion matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)

#pickle file joblib
joblib.dump(classifier, '../final_models/logisticR_final.pkl')
```

Hình 2. 2: Thuật toán Logistic Regression sử dụng trong đồ án

2.3.3 Support Vector Machine

Support Vector Machine (SVM) là một thuật toán mạnh mẽ trong công nghệ học máy. Trong Support Vector Machine, mỗi mục dữ liệu được vẽ dưới dạng một điểm trong không gian n chiều và SVM xây dựng đường phân cách để phân loại hai lớp, đường phân tách này còn được gọi là siêu phẳng (hyperplane).

Support Vector Machine tìm kiếm các điểm gần nhất được gọi là support vector và khi nó tìm thấy điểm gần nhất, nó sẽ vẽ một đường nối với chúng. Support Vector Machine sau đó xây dựng đường phân cách chia đôi và vuông góc với đường nối. Để phân loại dữ liệu một cách hoàn hảo, biên độ phải tối đa. Ở đây lẽ là khoảng cách giữa siêu phẳng và support vector. Trong kịch bản thực tế, không thể tách dữ liệu phức tạp và

dữ liệu phi tuyến tính, để giải quyết vấn đề này, SVM sử dụng thủ thuật hạt nhân để biến đổi không gian chiều thấp hơn sang không gian chiều cao hơn.

```
#splitting the dataset into training set and test set
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size = 0.25, random_state = 0 )

#applying grid search to find best performing parameters
from sklearn.model_selection import GridSearchCV
parameters = [{'C':[1, 10, 100, 1000], 'gamma': [ 0.1, 0.2,0.3, 0.5]}]
grid_search = GridSearchCV(SVC(kernel='rbf' ), parameters,cv =5, n_jobs= -1)
grid_search.fit(x_train, y_train)

#printing best parameters
print("Best Accuracy =" +str( grid_search.best_score_))
print("best parameters =" + str(grid_search.best_params_))

#fitting kernel SVM with best parameters calculated

classifier = SVC(C=1000, kernel = 'rbf', gamma = 0.2 , random_state = 0)
classifier.fit(x_train, y_train)

#predicting the tests set result
y_pred = classifier.predict(x_test)

#confusion matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(cm)

#pickle file joblib
joblib.dump(classifier, '../final_models/svm_final.pkl')
```

Hình 2. 3: Thuật toán SVM sử dụng trong đồ án

CHƯƠNG 3 – DEMO

3.1 Chức năng có trong demo

Các model dự đoán Phishing website được dùng để dán nhãn phân loại trang web dựa trên các đặc điểm được trích xuất ra từ URL đưa vào.

Các tài nguyên được sử dụng để xây dựng model:

- Phishing Website Feature dùng để làm dữ liệu cho tập training và tập test.
- Tập dữ liệu URL Phishing chứa các url lừa đảo.

Chương trình của chúng em gồm có 3 chức năng như sau:

- Chạy single URL trên trình duyệt web, người dùng có thể lựa chọn 1 trong 3 thuật toán để kiểm tra.
- Chạy Multi URLs trên trình duyệt web, người dùng có thể lựa chọn 1 trong 3 thuật toán để kiểm tra.
- Chạy trực tiếp trên Command Prompt.

=> Kết quả trả về là nhãn nhận dạng trang web đó có phải là trang web phishing hay không.

3.2 Chạy chương trình demo

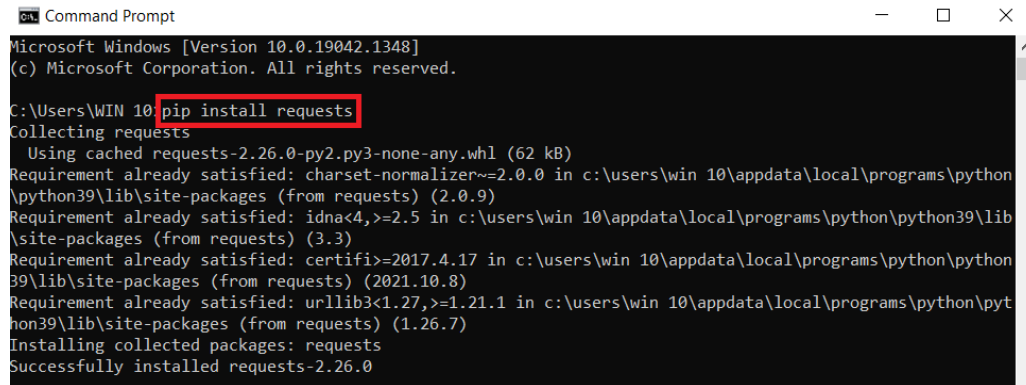
3.2.1 Cài đặt thư viện *Python*

Trước khi chạy chương trình demo, ta cần phải cài đặt đầy đủ các thư viện python sau đây. **Lưu ý** chương trình sẽ chạy **sai** hoặc **bị lỗi** nếu không cài đặt đầy đủ các thư viện yêu cầu.

Mở Command Prompt, cài đặt các thư viện python bằng câu lệnh:

`pip install library_name`

Trong đó library_name là tên thư viện trong python.



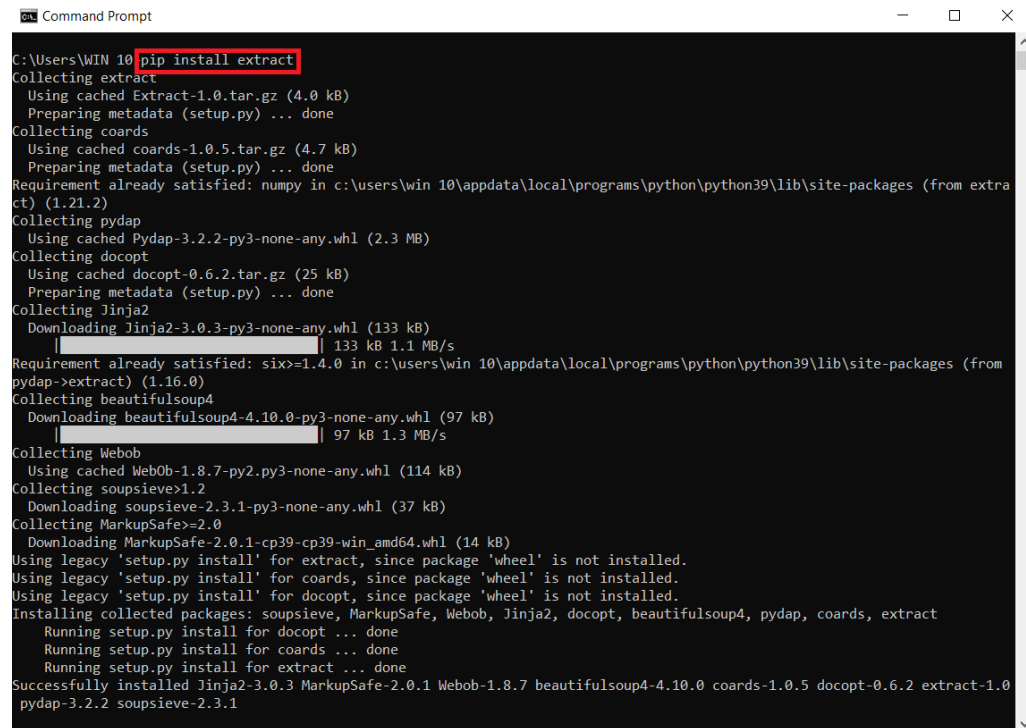
```

Microsoft Windows [Version 10.0.19042.1348]
(c) Microsoft Corporation. All rights reserved.

C:\Users\WIN 10>pip install requests
Collecting requests
  Using cached requests-2.26.0-py2.py3-none-any.whl (62 kB)
Requirement already satisfied: charset-normalizer~=2.0.0 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from requests) (2.0.9)
Requirement already satisfied: idna<4,>=2.5 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from requests) (3.3)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from requests) (2021.10.8)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from requests) (1.26.7)
Installing collected packages: requests
Successfully installed requests-2.26.0

```

Hình 3. 1: Cài đặt thư viện Requests



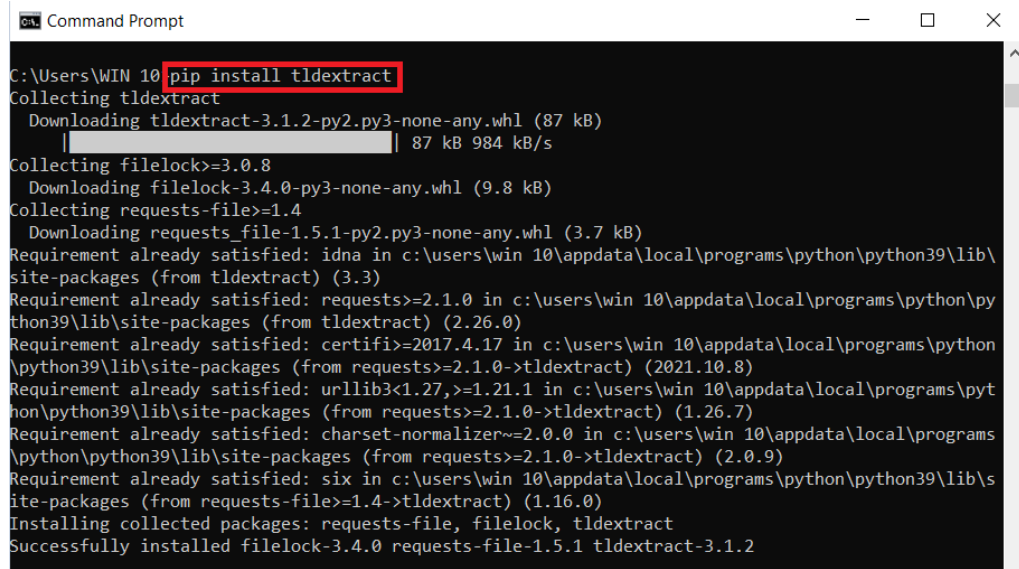
```

Microsoft Windows [Version 10.0.19042.1348]
(c) Microsoft Corporation. All rights reserved.

C:\Users\WIN 10>pip install extract
Collecting extract
  Using cached Extract-1.0.tar.gz (4.0 kB)
  Preparing metadata (setup.py) ... done
Collecting coards
  Using cached coards-1.0.5.tar.gz (4.7 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: numpy in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from extract) (1.21.2)
Collecting pydap
  Using cached Pydap-3.2.2-py3-none-any.whl (2.3 MB)
Collecting docopt
  Using cached docopt-0.6.2.tar.gz (25 kB)
  Preparing metadata (setup.py) ... done
Collecting Jinja2
  Downloading Jinja2-3.0.3-py3-none-any.whl (133 kB)
    |#####| 133 kB 1.1 MB/s
Requirement already satisfied: six>=1.4.0 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from pydap->extract) (1.16.0)
Collecting beautifulsoup4
  Downloading beautifulsoup4-4.10.0-py3-none-any.whl (97 kB)
    |#####| 97 kB 1.3 MB/s
Collecting Webob
  Using cached WebOb-1.8.7-py2.py3-none-any.whl (114 kB)
Collecting soupsieve>1.2
  Downloading soupsieve-2.3.1-py3-none-any.whl (37 kB)
Collecting MarkupSafe>=2.0
  Downloading MarkupSafe-2.0.1-cp39-cp39-win_amd64.whl (14 kB)
Using legacy 'setup.py install' for extract, since package 'wheel' is not installed.
Using legacy 'setup.py install' for coards, since package 'wheel' is not installed.
Using legacy 'setup.py install' for docopt, since package 'wheel' is not installed.
Installing collected packages: soupsieve, MarkupSafe, Webob, Jinja2, docopt, beautifulsoup4, pydap, coards, extract
  Running setup.py install for docopt ... done
  Running setup.py install for coards ... done
  Running setup.py install for extract ... done
Successfully installed Jinja2-3.0.3 MarkupSafe-2.0.1 Webob-1.8.7 beautifulsoup4-4.10.0 coards-1.0.5 docopt-0.6.2 extract-1.0 pydap-3.2.2 soupsieve-2.3.1

```

Hình 3. 2: Cài đặt thư viện Extract

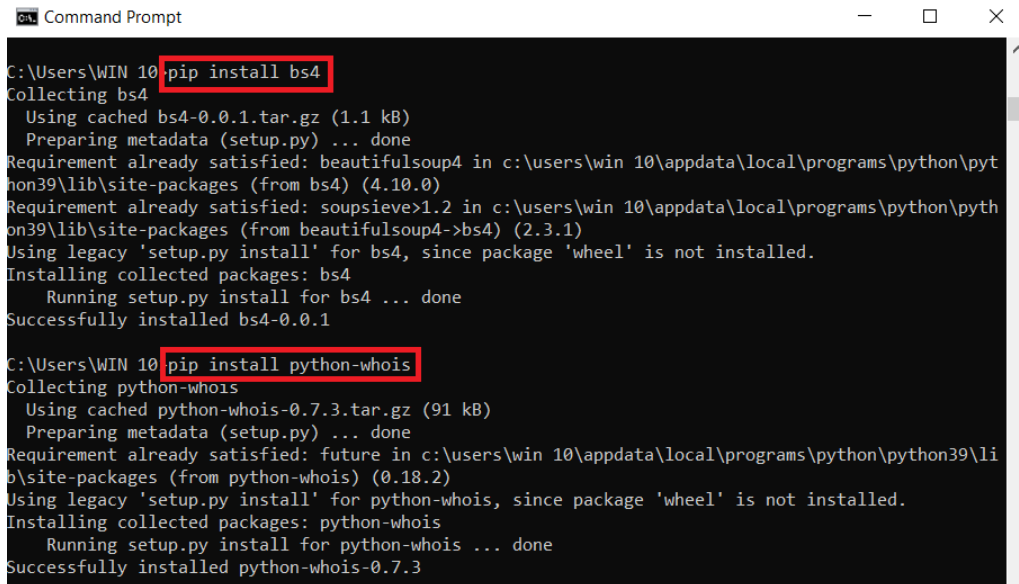


```

C:\Users\WIN 10> pip install tldextract
Collecting tldextract
  Downloading tldextract-3.1.2-py2.py3-none-any.whl (87 kB)
    | 87 kB 984 kB/s
Collecting filelock>=3.0.8
  Downloading filelock-3.4.0-py3-none-any.whl (9.8 kB)
Collecting requests-file>=1.4
  Downloading requests_file-1.5.1-py2.py3-none-any.whl (3.7 kB)
Requirement already satisfied: idna in c:\users\win 10\appdata\local\programs\python\python39\lib\
site-packages (from tldextract) (3.3)
Requirement already satisfied: requests>=2.1.0 in c:\users\win 10\appdata\local\programs\python\py
thon39\lib\site-packages (from tldextract) (2.26.0)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\win 10\appdata\local\programs\python
\python39\lib\site-packages (from requests>=2.1.0->tldextract) (2021.10.8)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\win 10\appdata\local\programs\pyt
hon\python39\lib\site-packages (from requests>=2.1.0->tldextract) (1.26.7)
Requirement already satisfied: charset-normalizer~=2.0.0 in c:\users\win 10\appdata\local\programs
\python\python39\lib\site-packages (from requests>=2.1.0->tldextract) (2.0.9)
Requirement already satisfied: six in c:\users\win 10\appdata\local\programs\python\python39\lib\s
ite-packages (from requests-file>=1.4->tldextract) (1.16.0)
Installing collected packages: requests-file, filelock, tldextract
Successfully installed filelock-3.4.0 requests-file-1.5.1 tldextract-3.1.2

```

Hình 3. 3: Cài đặt thư viện Tldextract



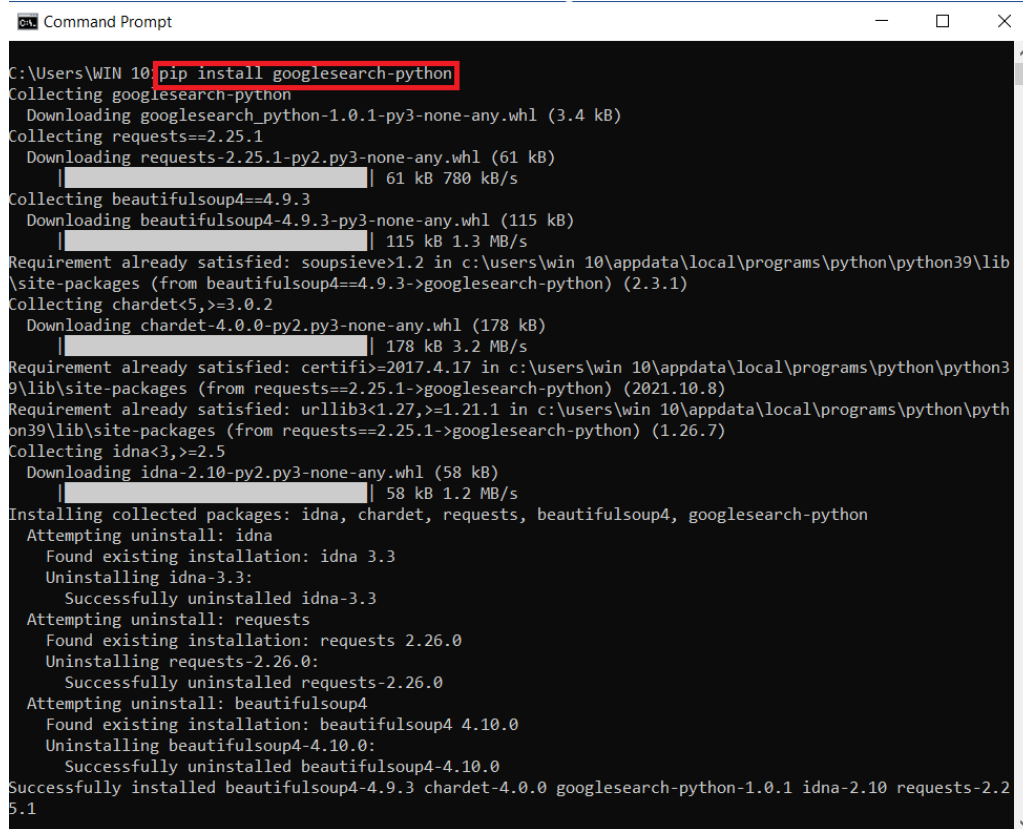
```

C:\Users\WIN 10> pip install bs4
Collecting bs4
  Using cached bs4-0.0.1.tar.gz (1.1 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: beautifulsoup4 in c:\users\win 10\appdata\local\programs\python\pyt
hon39\lib\site-packages (from bs4) (4.10.0)
Requirement already satisfied: soupsieve>1.2 in c:\users\win 10\appdata\local\programs\python\pyt
hon39\lib\site-packages (from beautifulsoup4->bs4) (2.3.1)
Using legacy 'setup.py install' for bs4, since package 'wheel' is not installed.
Installing collected packages: bs4
  Running setup.py install for bs4 ... done
Successfully installed bs4-0.0.1

C:\Users\WIN 10> pip install python-whois
Collecting python-whois
  Using cached python-whois-0.7.3.tar.gz (91 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: future in c:\users\win 10\appdata\local\programs\python\python39\li
b\site-packages (from python-whois) (0.18.2)
Using legacy 'setup.py install' for python-whois, since package 'wheel' is not installed.
Installing collected packages: python-whois
  Running setup.py install for python-whois ... done
Successfully installed python-whois-0.7.3

```

Hình 3. 4: Cài đặt thư viện BS4 và Whois

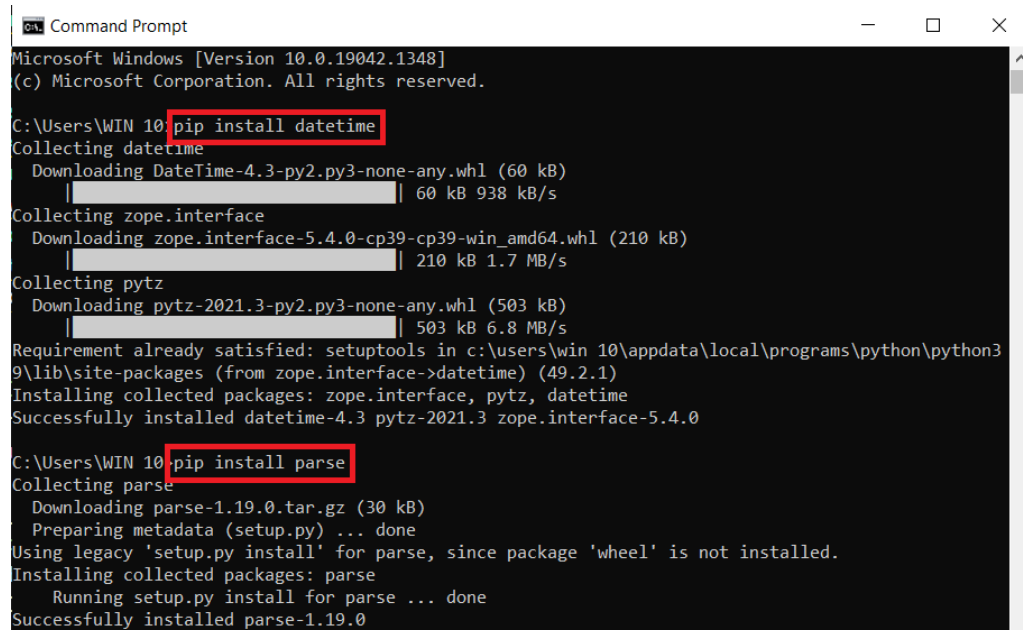


```

C:\Users\WIN 10>pip install googlesearch-python
Collecting googlesearch-python
  Downloading googlesearch-python-1.0.1-py3-none-any.whl (3.4 kB)
Collecting requests==2.25.1
  Downloading requests-2.25.1-py2.py3-none-any.whl (61 kB)
    | 61 kB 780 kB/s
Collecting beautifulsoup4==4.9.3
  Downloading beautifulsoup4-4.9.3-py3-none-any.whl (115 kB)
    | 115 kB 1.3 MB/s
Requirement already satisfied: soupsieve>1.2 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from beautifulsoup4==4.9.3->googlesearch-python) (2.3.1)
Collecting chardet<5,>=3.0.2
  Downloading chardet-4.0.0-py2.py3-none-any.whl (178 kB)
    | 178 kB 3.2 MB/s
Requirement already satisfied: certifi>=2017.4.17 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from requests==2.25.1->googlesearch-python) (2021.10.8)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from requests==2.25.1->googlesearch-python) (1.26.7)
Collecting idna<3,>=2.5
  Downloading idna-2.10-py2.py3-none-any.whl (58 kB)
    | 58 kB 1.2 MB/s
Installing collected packages: idna, chardet, requests, beautifulsoup4, googlesearch-python
  Attempting uninstall: idna
    Found existing installation: idna 3.3
    Uninstalling idna-3.3:
      Successfully uninstalled idna-3.3
  Attempting uninstall: requests
    Found existing installation: requests 2.26.0
    Uninstalling requests-2.26.0:
      Successfully uninstalled requests-2.26.0
  Attempting uninstall: beautifulsoup4
    Found existing installation: beautifulsoup4 4.10.0
    Uninstalling beautifulsoup4-4.10.0:
      Successfully uninstalled beautifulsoup4-4.10.0
Successfully installed beautifulsoup4-4.9.3 chardet-4.0.0 googlesearch-python-1.0.1 idna-2.10 requests-2.25.1

```

Hình 3. 5: Cài đặt thư viện Google



```

C:\Users\WIN 10>pip install datetime
Collecting datetime
  Downloading DateTime-4.3-py2.py3-none-any.whl (60 kB)
    | 60 kB 938 kB/s
Collecting zope.interface
  Downloading zope.interface-5.4.0-cp39-cp39-win_amd64.whl (210 kB)
    | 210 kB 1.7 MB/s
Collecting pytz
  Downloading pytz-2021.3-py2.py3-none-any.whl (503 kB)
    | 503 kB 6.8 MB/s
Requirement already satisfied: setuptools in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from zope.interface->datetime) (49.2.1)
Installing collected packages: zope.interface, pytz, datetime
Successfully installed datetime-4.3 pytz-2021.3 zope.interface-5.4.0

C:\Users\WIN 10>pip install parse
Collecting parse
  Downloading parse-1.19.0.tar.gz (30 kB)
  Preparing metadata (setup.py) ... done
Using legacy 'setup.py install' for parse, since package 'wheel' is not installed.
Installing collected packages: parse
  Running setup.py install for parse ... done
Successfully installed parse-1.19.0

```

Hình 3. 6: Cài đặt thư viện Datetime và Parse

```

C:\Users\WIN 10>pip install joblib
Collecting joblib
  Downloading joblib-1.1.0-py2.py3-none-any.whl (306 kB)
    |-----| 306 kB 1.1 MB/s
Installing collected packages: joblib
Successfully installed joblib-1.1.0

C:\Users\WIN 10>pip install pandas
Collecting pandas
  Downloading pandas-1.3.5-cp39-cp39-win_amd64.whl (10.2 MB)
    |-----| 10.2 MB 3.3 MB/s
Requirement already satisfied: numpy>=1.17.3 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from pandas) (1.21.2)
Requirement already satisfied: python-dateutil>=2.7.3 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2017.3 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from pandas) (2021.3)
Requirement already satisfied: six>=1.5 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from python-dateutil>=2.7.3->pandas) (1.16.0)
Installing collected packages: pandas
Successfully installed pandas-1.3.5

```

Hình 3. 7: Cài đặt thư viện Joblib và Pandas

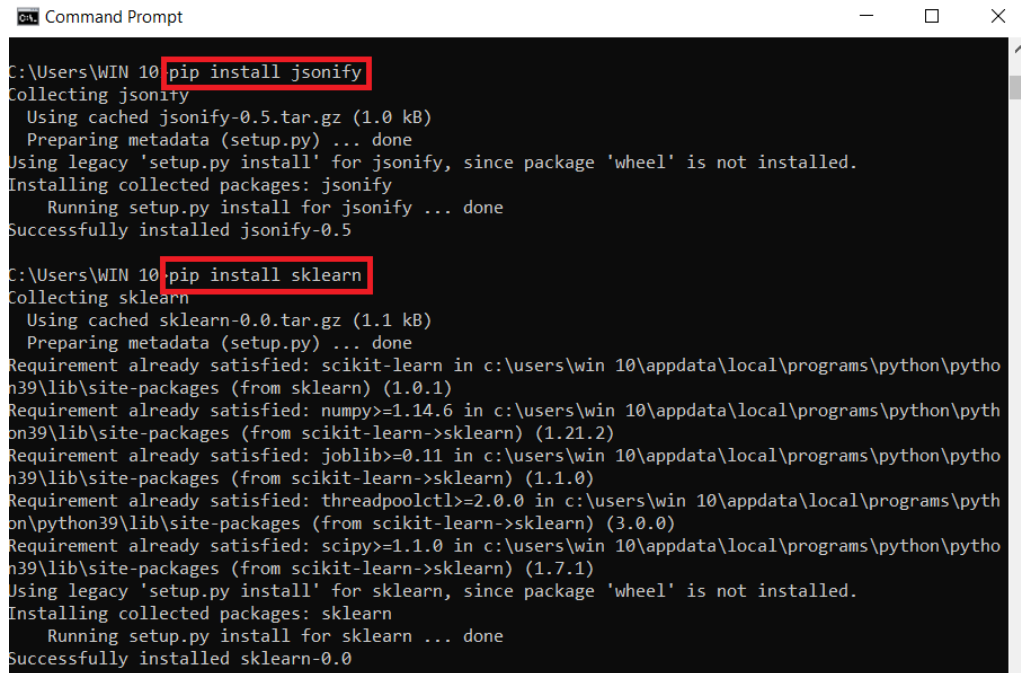
```

C:\Users\WIN 10>pip install print_dict
Collecting print_dict
  Using cached print_dict-0.1.19-py3-none-any.whl (8.3 kB)
Requirement already satisfied: yapf<0.31.0,>=0.30.0 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from print_dict) (0.30.0)
Installing collected packages: print-dict
Successfully installed print-dict-0.1.19

C:\Users\WIN 10>pip install Flask
Collecting Flask
  Downloading Flask-2.0.2-py3-none-any.whl (95 kB)
    |-----| 95 kB 1.0 MB/s
Requirement already satisfied: Jinja2>=3.0 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from Flask) (3.0.3)
Collecting Werkzeug>=2.0
  Downloading Werkzeug-2.0.2-py3-none-any.whl (288 kB)
    |-----| 288 kB 1.6 MB/s
Collecting click>=7.1.2
  Downloading click-8.0.3-py3-none-any.whl (97 kB)
    |-----| 97 kB 2.2 MB/s
Collecting itsdangerous>=2.0
  Downloading itsdangerous-2.0.1-py3-none-any.whl (18 kB)
Collecting colorama
  Downloading colorama-0.4.4-py2.py3-none-any.whl (16 kB)
Requirement already satisfied: MarkupSafe>=2.0 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from Jinja2>=3.0->Flask) (2.0.1)
Installing collected packages: colorama, Werkzeug, itsdangerous, click, Flask
Successfully installed Flask-2.0.2 Werkzeug-2.0.2 click-8.0.3 colorama-0.4.4 itsdangerous-2.0.1

```

Hình 3. 8: Cài đặt thư viện Print_dict và Flask



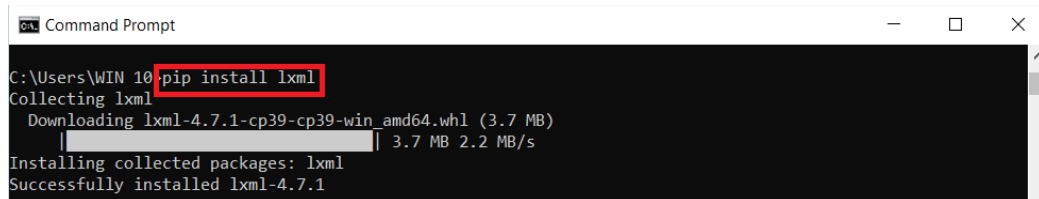
```

C:\Users\WIN 10>pip install jsonify
Collecting jsonify
  Using cached jsonify-0.5.tar.gz (1.0 kB)
  Preparing metadata (setup.py) ... done
Using legacy 'setup.py install' for jsonify, since package 'wheel' is not installed.
Installing collected packages: jsonify
  Running setup.py install for jsonify ... done
Successfully installed jsonify-0.5

C:\Users\WIN 10>pip install sklearn
Collecting sklearn
  Using cached sklearn-0.0.tar.gz (1.1 kB)
  Preparing metadata (setup.py) ... done
Requirement already satisfied: scikit-learn in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from sklearn) (1.0.1)
Requirement already satisfied: numpy>=1.14.6 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from scikit-learn->sklearn) (1.21.2)
Requirement already satisfied: joblib>=0.11 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from scikit-learn->sklearn) (1.1.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from scikit-learn->sklearn) (3.0.0)
Requirement already satisfied: scipy>=1.1.0 in c:\users\win 10\appdata\local\programs\python\python39\lib\site-packages (from scikit-learn->sklearn) (1.7.1)
Using legacy 'setup.py install' for sklearn, since package 'wheel' is not installed.
Installing collected packages: sklearn
  Running setup.py install for sklearn ... done
Successfully installed sklearn-0.0

```

Hình 3. 9: Cài đặt thư viện Jsonify



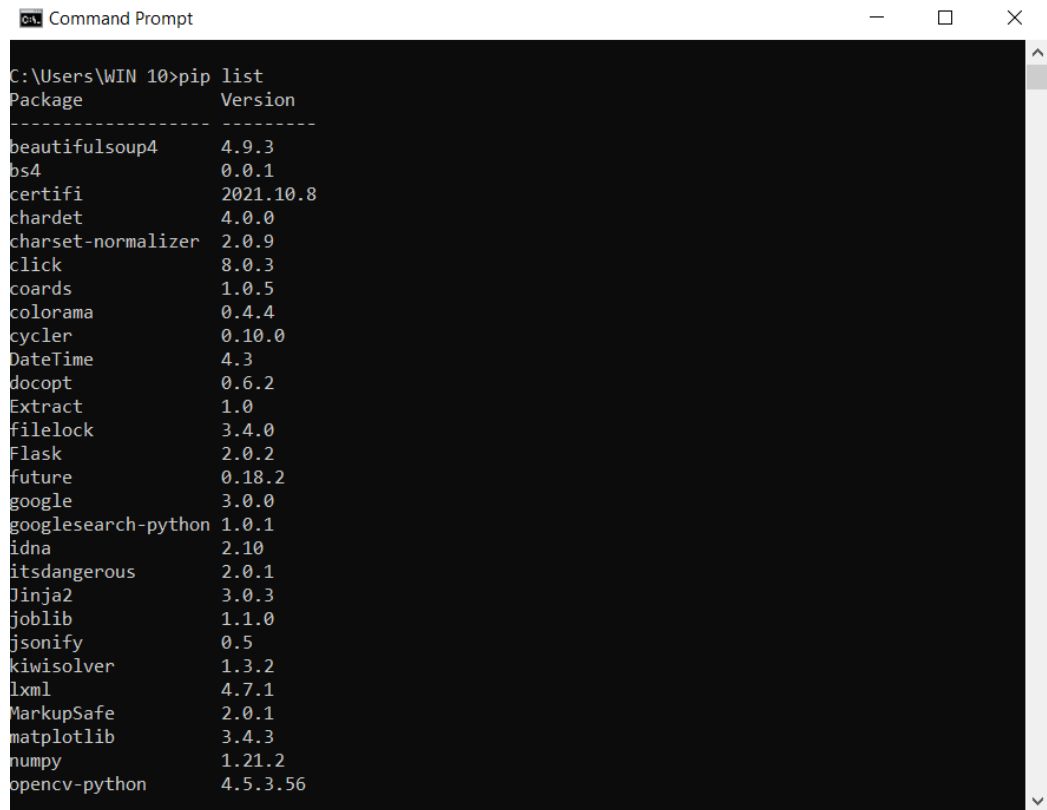
```

C:\Users\WIN 10>pip install lxml
Collecting lxml
  Downloading lxml-4.7.1-cp39-cp39-win_amd64.whl (3.7 MB)
    | 3.7 MB 2.2 MB/s
Installing collected packages: lxml
Successfully installed lxml-4.7.1

```

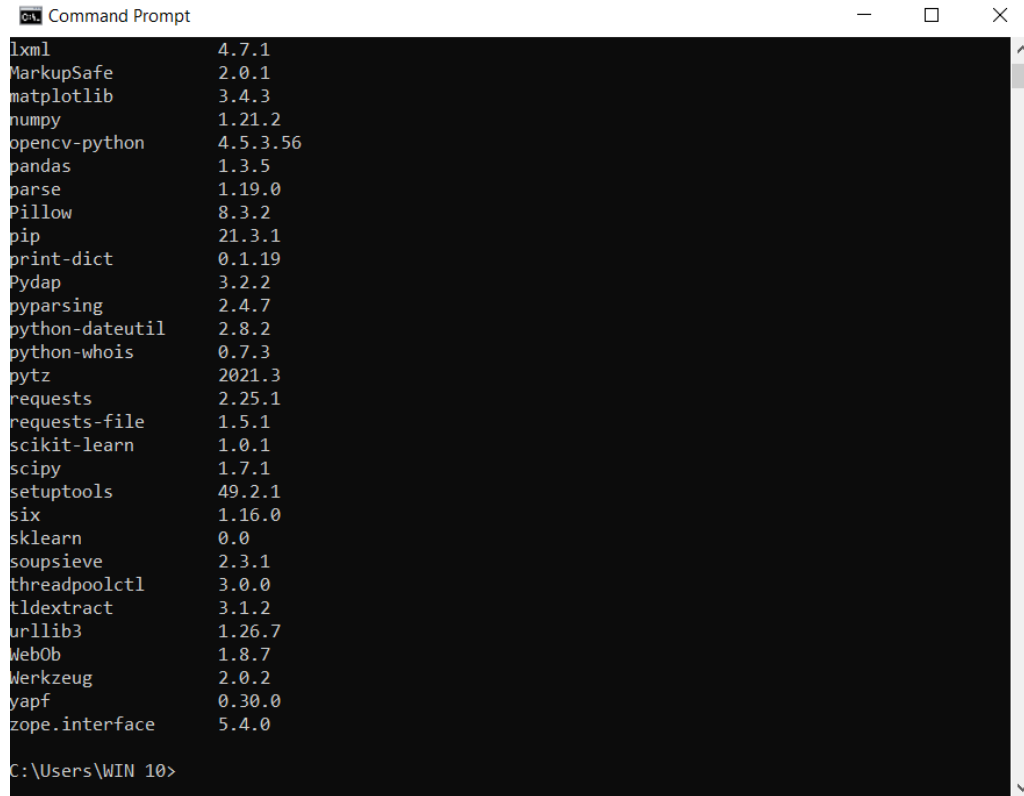
Hình 3. 10: Cài đặt thư viện LXML

Sau khi cài đặt các thư viện xong, ta chạy lệnh **pip list** để kiểm tra lại các thư viện đã cài đặt



```
C:\Users\WIN 10>pip list
Package            Version
-----
beautifulsoup4     4.9.3
bs4                 0.0.1
certifi             2021.10.8
chardet             4.0.0
charset-normalizer  2.0.9
click               8.0.3
coards              1.0.5
colorama            0.4.4
cyclor              0.10.0
DateTime            4.3
docopt              0.6.2
Extract             1.0
filelock            3.4.0
Flask               2.0.2
future              0.18.2
google              3.0.0
googlesearch-python 1.0.1
idna                2.10
itsdangerous        2.0.1
Jinja2              3.0.3
joblib              1.1.0
jsonify             0.5
kiwisolver          1.3.2
lxml                4.7.1
MarkupSafe          2.0.1
matplotlib          3.4.3
numpy               1.21.2
opencv-python       4.5.3.56
```

Hình 3. 11: Kiểm tra lại thư viện Python - 1



```

C:\Users\WIN 10>
lxml 4.7.1
MarkupSafe 2.0.1
matplotlib 3.4.3
numpy 1.21.2
opencv-python 4.5.3.56
pandas 1.3.5
parse 1.19.0
Pillow 8.3.2
pip 21.3.1
print-dict 0.1.19
Pydap 3.2.2
pyparsing 2.4.7
python-dateutil 2.8.2
python-whois 0.7.3
pytz 2021.3
requests 2.25.1
requests-file 1.5.1
scikit-learn 1.0.1
scipy 1.7.1
setuptools 49.2.1
six 1.16.0
sklearn 0.0
soupsieve 2.3.1
threadpoolctl 3.0.0
tldextract 3.1.2
urllib3 1.26.7
WebOb 1.8.7
Werkzeug 2.0.2
yapf 0.30.0
zope.interface 5.4.0
C:\Users\WIN 10>

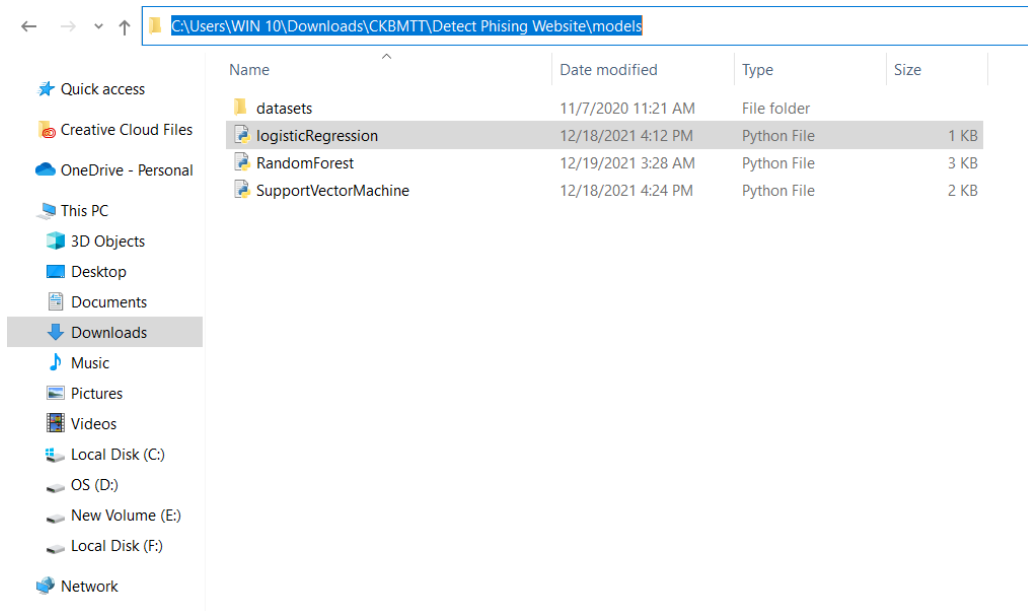
```

Hình 3. 12: Kiểm tra lại thư viện Python - 2

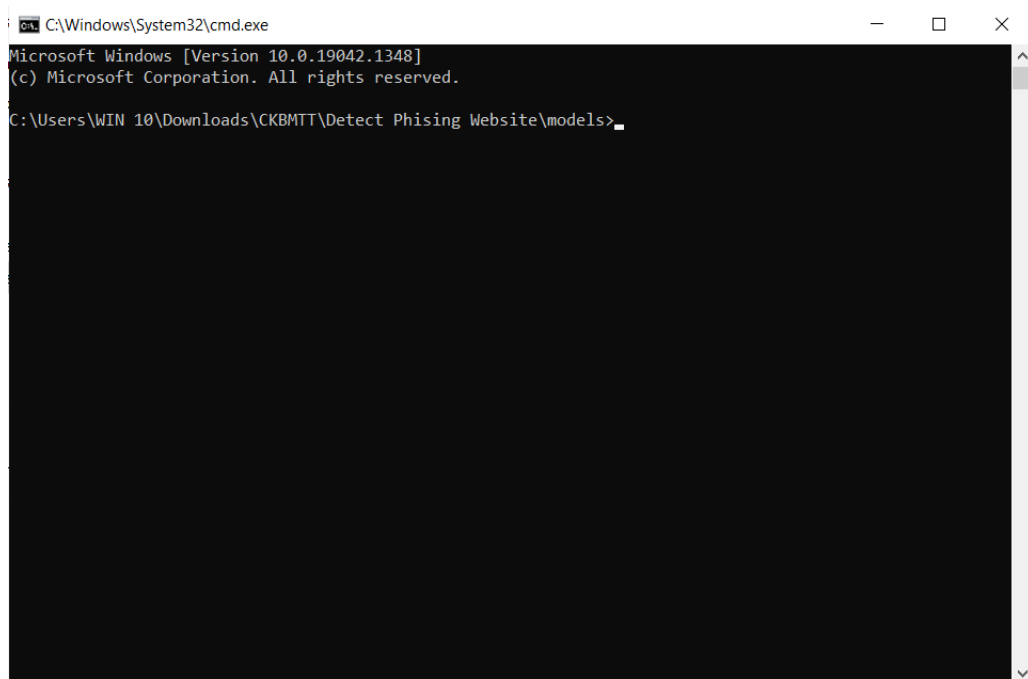
3.2.2 Chạy các thuật toán

Do quá trình nén và giải nén các file chương trình có thể gây nên mất dữ liệu ngoài ý muốn, ta nên chạy trước các model thuật toán ở trong thư mục **models** trước khi chạy chương trình chính. Thư mục models gồm có 3 thuật toán: *logisticRegression.py*, *RandomForest.py*, *SupportVectorMachine.py*. Ta chạy các file này như sau:

- **Bước 1:** Mở Command Prompt chứa đường dẫn tới các file models. Ví dụ:
`C:\Users\WIN 10\Downloads\CKBMTT\Detect Phising Website\models`



- **Bước 2:** Xóa đường dẫn trên và gõ **cmd** để mở Command Prompt



- **Bước 3:** Chạy các thuật toán bằng câu lệnh **python tên_thuật_toán.py**

```

C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.19042.1348]
(c) Microsoft Corporation. All rights reserved.

C:\Users\WIN 10\Downloads\CKBMTT\Detect Phishing Website\models>python logisticRegression.py
C:\Users\WIN 10\Downloads\CKBMTT\Detect Phishing Website\models\logisticRegression.py:12: FutureWarning:
In a future version of pandas all arguments of DataFrame.drop except for the argument 'labels' will be keyword-only
  dataset = dataset.drop('id', 1) #removing unwanted column
[[1121  128]
 [ 84 1431]]

C:\Users\WIN 10\Downloads\CKBMTT\Detect Phishing Website\models>python RandomForest.py
C:\Users\WIN 10\Downloads\CKBMTT\Detect Phishing Website\models\RandomForest.py:14: FutureWarning: In a future version of pandas all arguments of DataFrame.drop except for the argument 'labels' will be keyword-only
  dataset = dataset.drop('id', 1) #removing unwanted column
[-1 -1 -1 ... -1 -1 -1]
Best Accuracy =0.9721377201229394
best parameters ={'criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 100}
[[1181  68]
 [ 19 1496]]

C:\Users\WIN 10\Downloads\CKBMTT\Detect Phishing Website\models>python SupportVectorMachine.py
C:\Users\WIN 10\Downloads\CKBMTT\Detect Phishing Website\models\SupportVectorMachine.py:12: FutureWarning:
In a future version of pandas all arguments of DataFrame.drop except for the argument 'labels' will be keyword-only
  dataset = dataset.drop('id', 1) #removing unwanted column
Best Accuracy =0.964660211399458
best parameters ={'C': 1000, 'gamma': 0.2}
[[1185  64]
 [ 26 1489]]

C:\Users\WIN 10\Downloads\CKBMTT\Detect Phishing Website\models>

```

3.2.3 Chạy chương trình demo trên trình duyệt web

Bước 1: Mở Command Prompt từ đường dẫn chứa file chương trình chính tương tự như bước trên và chạy bằng câu lệnh: **python server.py**

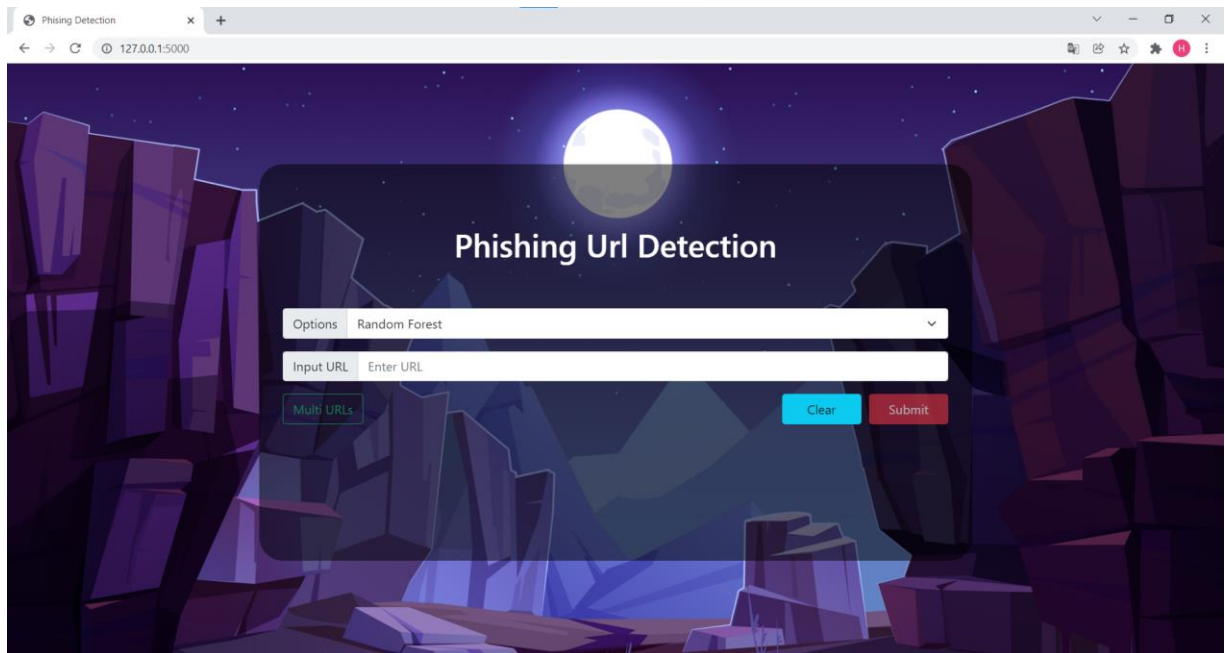
```

C:\Windows\System32\cmd.exe - python server.py
Microsoft Windows [Version 10.0.19042.1348]
(c) Microsoft Corporation. All rights reserved.

C:\Users\WIN 10\Downloads\CKBMTT\Detect Phising Website>python server.py
* Serving Flask app 'server' (lazy loading)
* Environment: production
  WARNING: This is a development server. Do not use it in a production deployment.
  Use a production WSGI server instead.
* Debug mode: on
* Restarting with stat
* Debugger is active!
* Debugger PIN: 435-260-768
* Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)

```

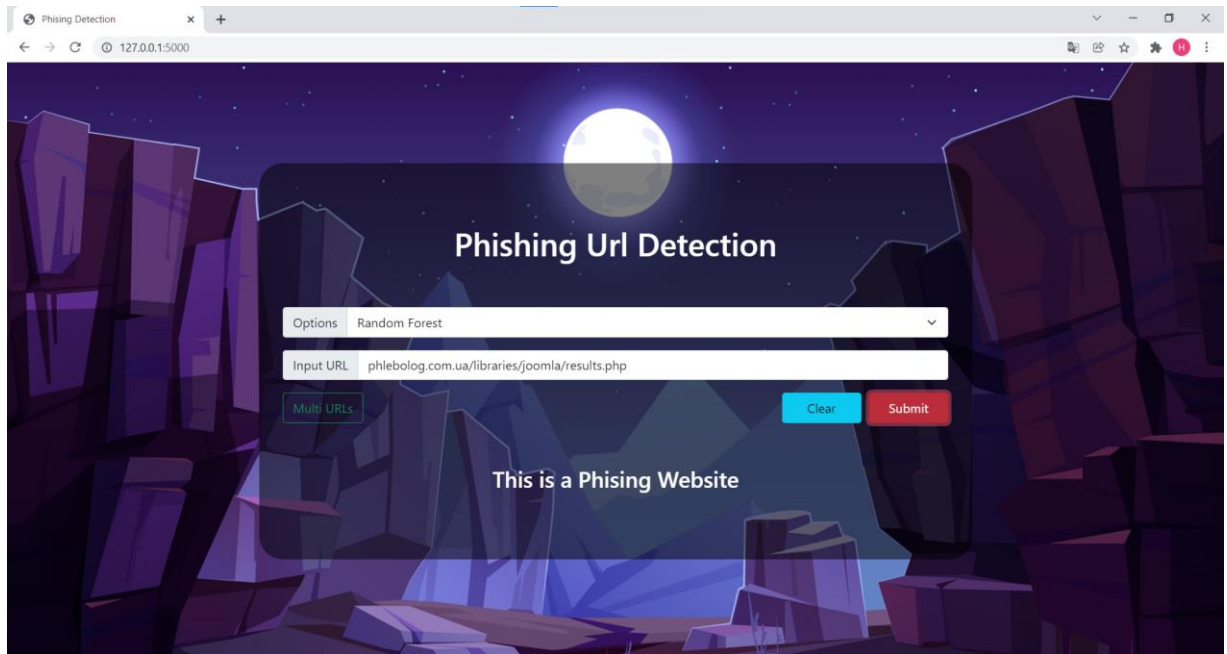
Bước 2: Copy đường dẫn **http://127.0.0.1:5000/** và dán vào thanh địa chỉ của Google Chrome.



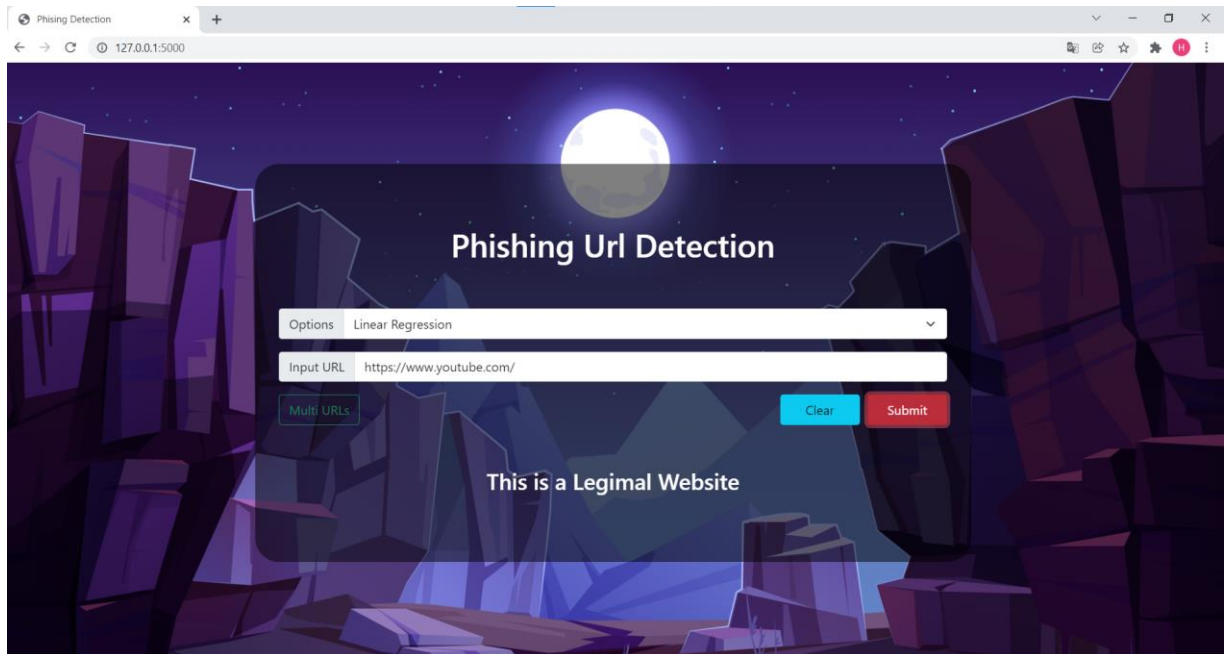
Bước 3: Chọn một thuật toán mà bạn cần chạy, ví dụ ở đây ta chọn thuật toán Random Forest. Tiếp theo chọn một URL tùy ý hoặc URL Phishing (có thể chọn URL

phishing từ 2 file **urls.csv**, **urlset.csv** nằm trong thư mục **files/**). Sau đó chọn Submit để kiểm tra URL là hợp pháp hay giả mạo.

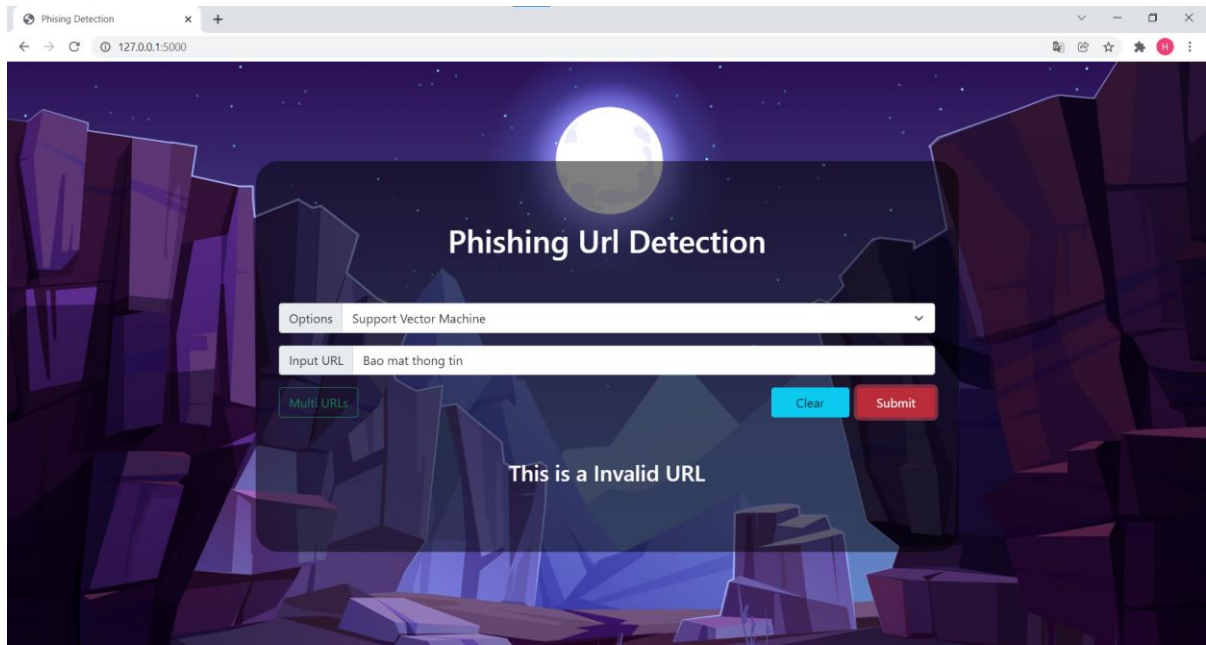
- Trường hợp phát hiện URL giả mạo



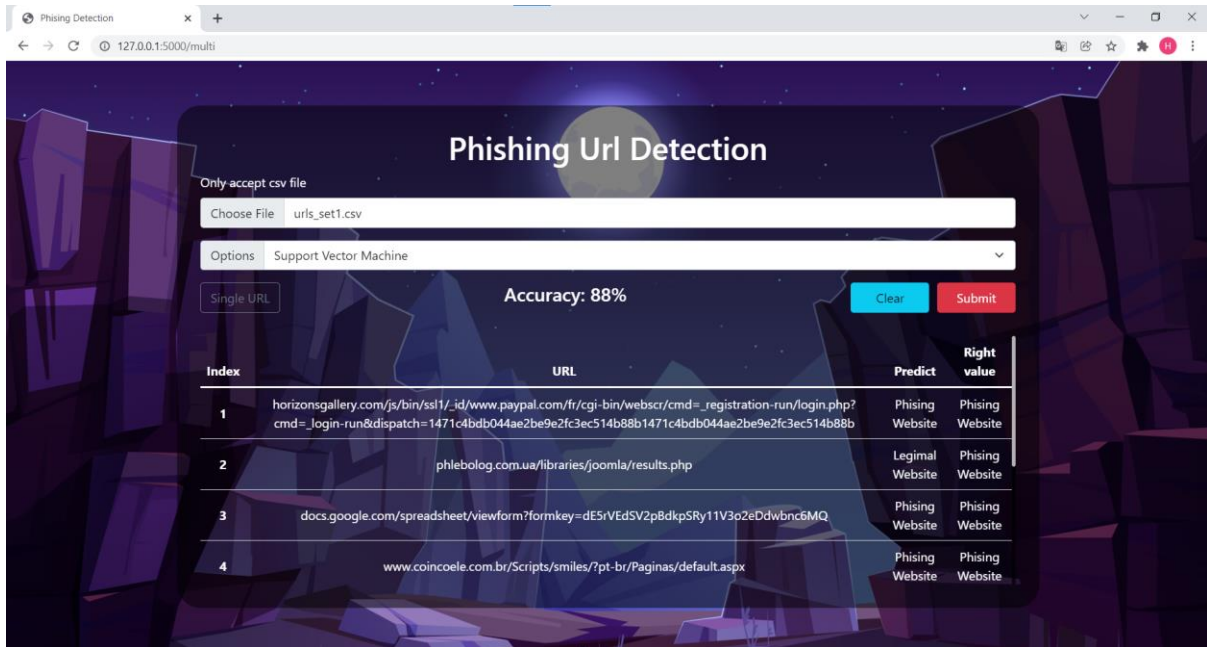
- Trường hợp phát hiện URL hợp pháp



- Trường hợp nhập URL không hợp lệ



Đối với trường hợp chạy nhiều URL cùng một lúc (**Multi URLs**), ta chỉ nên chạy khoảng 10 URL trở lại vì thời gian phản hồi sẽ rất lâu. Và tính năng chạy Multi URLs chỉ chấp nhận file có định dạng csv. Ngoài ra, yêu cầu tối thiểu của file csv là phải bao gồm ít nhất một cột domain (có thể thêm 1 cột label).

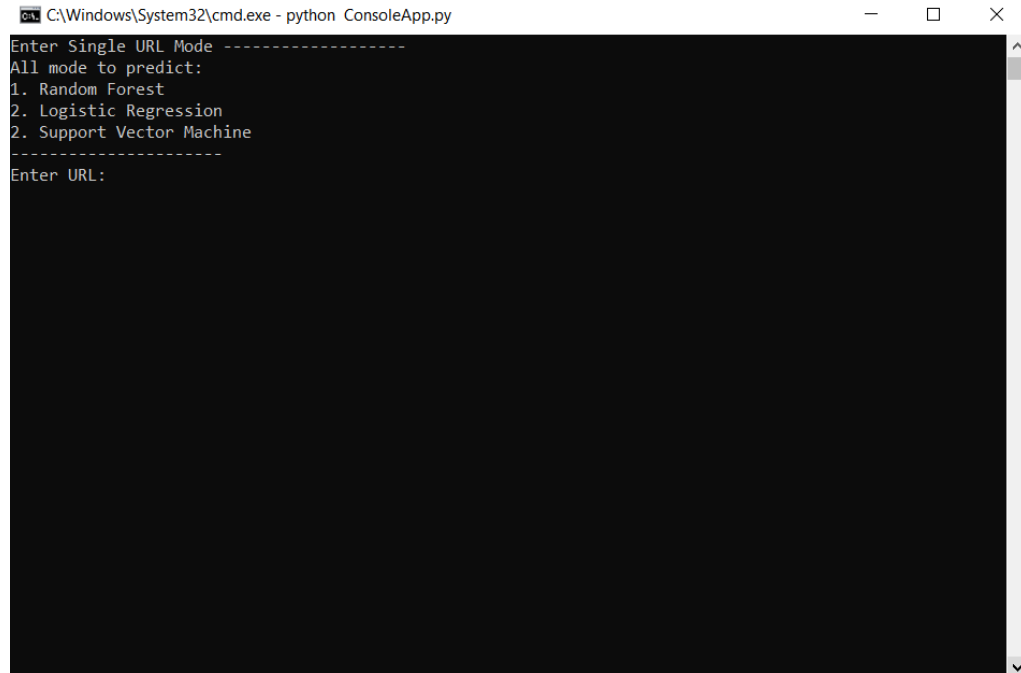


3.2.3 Chạy chương trình demo trực tiếp trên Command Prompt

Bước 1: Mở Command Prompt từ đường dẫn chứa file chương trình chính tương tự như bước trên và chạy bằng câu lệnh: **python ConsoleApp.py**. Sau khi chạy câu lệnh xong, ta sẽ được hiển thị cửa sổ như sau:

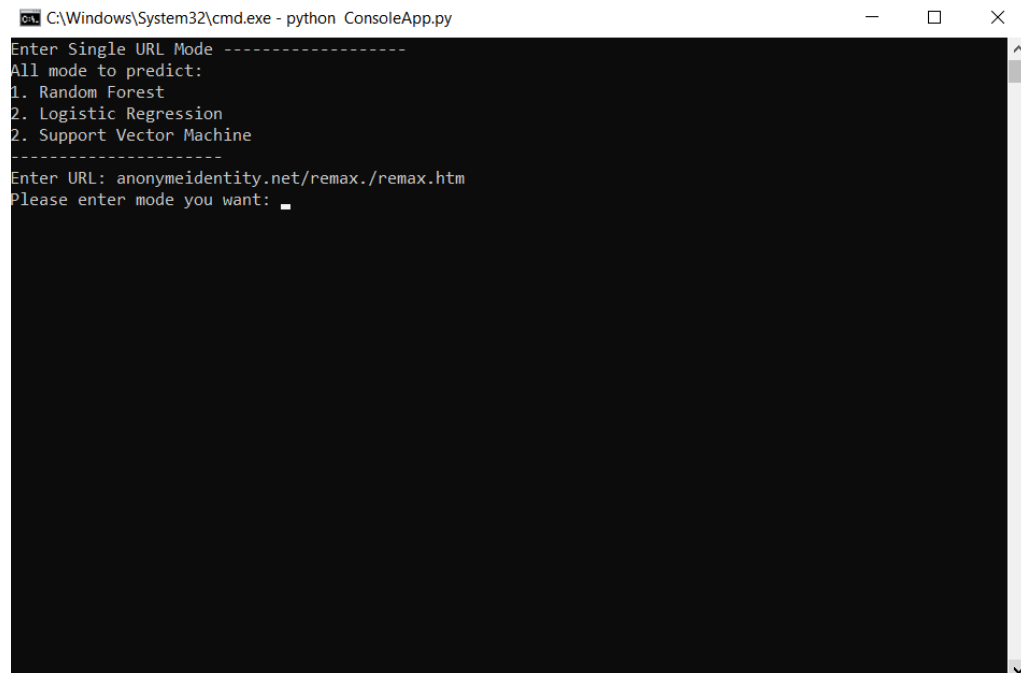
```
C:\Windows\System32\cmd.exe - python ConsoleApp.py
Detect Phishing Website
#-----
1. Single URL
2. Multi URLs
3. Exit
--> Your choice: _
```

Bước 2: Có 2 phương án cho chúng ta lựa chọn: Single URL và Multi URLs. Ví dụ ở đây ta chọn phương án số 1 Single URL.



```
C:\Windows\System32\cmd.exe - python ConsoleApp.py
Enter Single URL Mode -----
All mode to predict:
1. Random Forest
2. Logistic Regression
2. Support Vector Machine
-----
Enter URL:
```

Bước 3: Nhập URL để kiểm tra



```
C:\Windows\System32\cmd.exe - python ConsoleApp.py
Enter Single URL Mode -----
All mode to predict:
1. Random Forest
2. Logistic Regression
2. Support Vector Machine
-----
Enter URL: anonymidentity.net/remax./remax.htm
Please enter mode you want: █
```

Bước 4: Chọn 1 thuật toán trong số 3 thuật toán mà ta muốn kiểm tra URL. Ví dụ chọn thuật toán số 1 – Random Forest. Sau đó kiểm tra kết quả nhận được.

```
C:\Windows\System32\cmd.exe - python ConsoleApp.py
Enter Single URL Mode -----
All mode to predict:
1. Random Forest
2. Logistic Regression
2. Support Vector Machine
-----
Enter URL: anonymidentity.net/remax./remax.htm
Please enter mode you want: 1
--> rf_final
No match for "ANONYMEIDENTITY.NET".
>>> Last update of whois database: 2021-12-24T08:38:23Z <<<

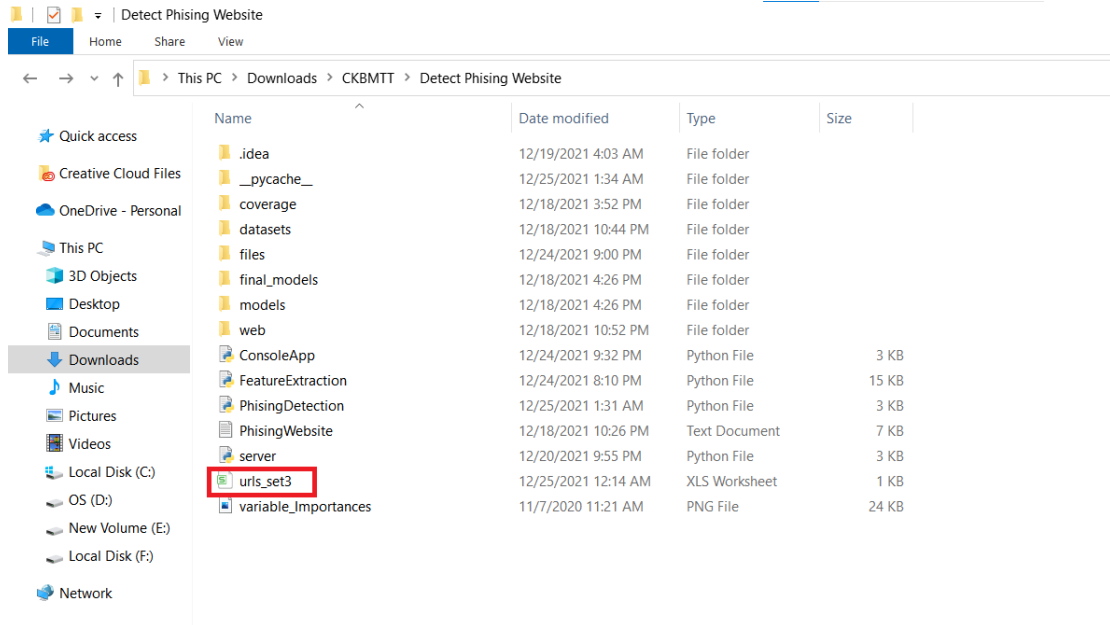
NOTICE: The expiration date displayed in this record is the date the
registrar's sponsorship of the domain name registration in the registry is
currently set to expire. This date does not necessarily reflect the expiration
date of the domain name registrant's agreement with the sponsoring
registrar. Users may consult the sponsoring registrar's Whois database to
view the registrar's reported date of expiration for this registration.

TERMS OF USE: You are not authorized to access or query our Whois
database through the use of electronic processes that are high-volume and
automated except as reasonably necessary to register domain names or
modify existing registrations; the Data in VeriSign Global Registry
Services' ("VeriSign") Whois database is provided by VeriSign for
information purposes only, and to assist persons in obtaining information
about or related to a domain name registration record. VeriSign does not
guarantee its accuracy. By submitting a Whois query, you agree to abide
by the following terms of use: You agree that you may use this Data only
for lawful purposes and that under no circumstances will you use this Data
to: (1) allow, enable, or otherwise support the transmission of mass
unsolicited, commercial advertising or solicitations via e-mail, telephone,
or facsimile; or (2) enable high volume, automated, electronic processes
that apply to VeriSign (or its computer systems). The compilation,
repackaging, dissemination or other use of this Data is expressly
prohibited without the prior written consent of VeriSign. You agree not to
use electronic processes that are automated and high-volume to access or
query the Whois database except as reasonably necessary to register
domain names or modify existing registrations. VeriSign reserves the right
to restrict your access to the Whois database in its sole discretion to ensure
operational stability. VeriSign may restrict or terminate your access to the
Whois database for failure to abide by these terms of use. VeriSign
reserves the right to modify these terms at any time.

The Registry database contains ONLY .COM, .NET, .EDU domains and
Registrars.

[[-1, -1, -1, -1, -1, -1, -1, 1, 0, -1, -1, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, -1, 0, -1]]
this is Phishing Website
```

Đối với trường hợp **Multi URLs**. Thay vì Bước 2 ta nhập URL thì ở bước này ta nhập tên file csv. Yêu cầu về file cvs cũng tương tự như yêu cầu đối với trình duyệt web. Lưu ý file csv dùng để chạy trên Command Prompt phải nằm cùng thư mục với file chương trình. Ví dụ:



- Nhập tên file csv muốn kiểm tra và lựa chọn 1 thuật toán

```

C:\Windows\System32\cmd.exe - python ConsoleApp.py
Enter Multi URLs Mode -----
All mode to predict:
1. Random Forest
2. Logistic Regression
2. Support Vector Machine
-----
Enter your file location: urls_set3.csv
Please enter mode you want: 2

```

- Kết quả thu được sau khi chạy Multi URLs

```

C:\Windows\System32\cmd.exe - python ConsoleApp.py
[[-1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, -1, 0, 0, -1, -1, 1, 1, 0, 1, 1, -1, -1, -1]]
[[-1, -1, -1, -1, -1, -1, -1, 0, 0, 1, -1, -1, -1, -1, 0, 1, -1, -1, -1, -1, 1, -1, -1, -1, -1]]
{
  'accuracy score': 0.9,
  'result': [{
    'domain': 'www.henkdeinumboomkwekerij.nl/language/pdf_fonts/smiles.php',
    'predict': 'Phising Website',
    'right value': 'Phising Website'
  }, {
    'domain': 'perfectsolutionofall.net/wp-content/themes/twentyten/wiresource/',
    'predict': 'Phising Website',
    'right value': 'Phising Website'
  }, {
    'domain': 'lingshc.com/old_aol.1.3/?Login=&Lis=10&LigertID=1993745&us=1',
    'predict': 'Phising Website',
    'right value': 'Phising Website'
  }, {
    'domain': 'anonymidentity.net/remax/remax.htm',
    'predict': 'Phising Website',
    'right value': 'Phising Website'
  }, {
    'domain': 'dutchweb.gtphost.com/zimbra/exch/owa/uleth/index.html',
    'predict': 'Legimal Website',
    'right value': 'Phising Website'
  }, {
    'domain': 'www.avedeoiro.com/site/plugins/chase/',
    'predict': 'Phising Website',
    'right value': 'Phising Website'
  }, {
    'domain': 'https://facebook.com/',
    'predict': 'Legimal Website',
    'right value': 'Legimal Website'
  }, {
    'domain': 'https://youtube.com/',
    'predict': 'Legimal Website',
    'right value': 'Legimal Website'
  }, {
    'domain': 'https://stdportal.tdtu.edu.vn/',
    'predict': 'Legimal Website',
    'right value': 'Legimal Website'
  }, {
    'domain': 'https://google.com',
    'predict': 'Legimal Website',
    'right value': 'Legimal Website'
  }
]}

```

3.2.4 Hướng xử lý khi chạy chương trình bị lỗi hoặc không chạy được

Khi chương trình không chạy được hoặc chạy chưa chính xác, ta nên lưu ý 2 tình huống sau đây:

- Chưa cài đặt đủ thư viện: Đối với tình huống này, ta nên kiểm tra lại xem mình đã cài đặt đầy đủ các thư viện mà chương trình yêu cầu chưa. Kiểm tra lại thư viện python bằng câu lệnh **`pip list`**.
- Chưa chạy các thuật toán mà đã chạy luôn chương trình chính. Do quá trình nén và giải nén dữ liệu nhiều lần, chương trình có thể bị mất đi một số dữ liệu nên có thể dẫn đến việc cho ra sai kết quả.

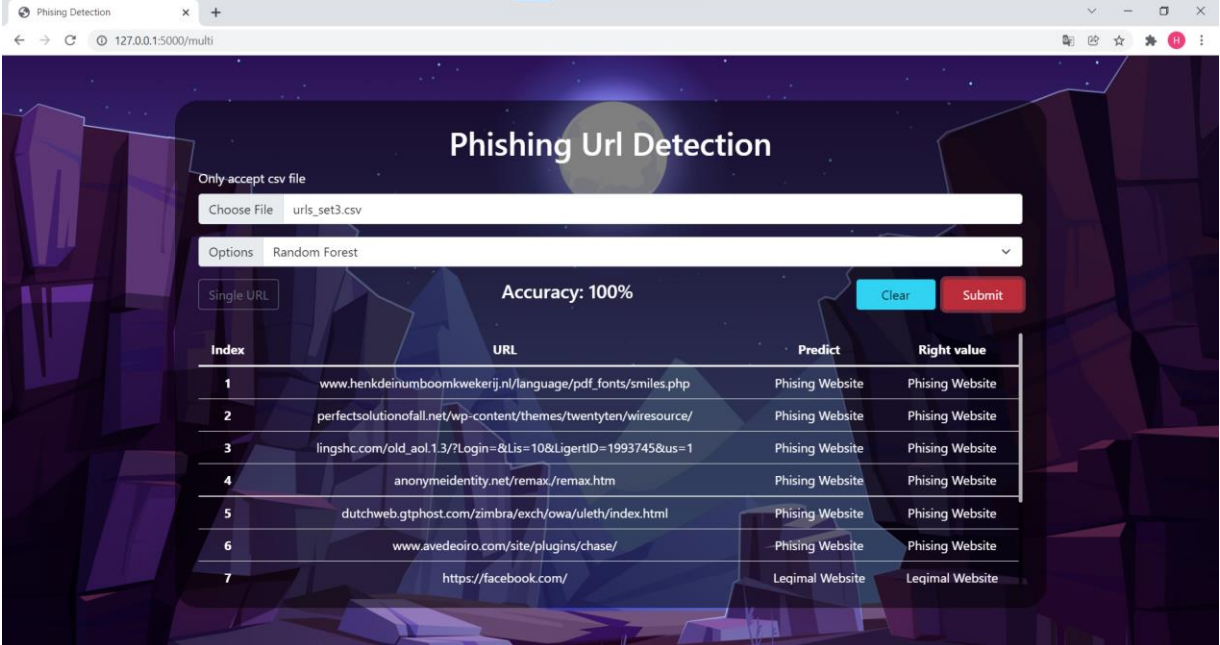
3.3 Luồng hoạt động của chương trình

- Đối với single URL:
 - Người dùng nhập input, client sẽ gửi cái URL từ input về server.
 - Server nhận được và gọi hàm từ phishing detection.
 - Trong phishing detection, đầu tiên là tiền xử lý bằng cách trích xuất ra các đặc điểm (feature extraction). Sau đó sử dụng model mà người dùng lựa chọn để predict URL đó.
 - Tiếp theo, một tập dữ liệu được tạo ra, trong đó mỗi chi tiết đặc điểm kết hợp với cặp (1, 0, -1), sau đó được chuyển đến các bộ phân loại khác nhau.
 - Tiếp đến, bộ phân loại phân tích tập dữ liệu (1, 0, -1) vừa được tạo ra dựa trên độ chính xác của từng thuật toán.
 - Cuối cùng, trả về kết quả cho người dùng.
- Đối với Multi URLs:
 - Người dùng nhập vào một file csv với cấu trúc gồm 2 thuộc tính là domain và label hoặc ít nhất phải có trường domain và chọn model cần dùng để dự đoán, client sẽ gửi file về server thông qua phương thức POST.
 - Server nhận được và gọi hàm từ phishing detection.
 - Trong phishing detection, đầu tiên là tiền xử lý bằng cách trích xuất ra các đặc điểm (feature extraction). Sau đó sử dụng model mà người dùng lựa chọn để predict tập URL đó.
 - Tiếp theo, một tập dữ liệu được tạo ra, trong đó mỗi chi tiết đặc điểm kết hợp với cặp (1, 0, -1), sau đó được chuyển đến các bộ phân loại khác nhau.
 - Tiếp đến, bộ phân loại phân tích tập dữ liệu (1, 0, -1) vừa được tạo ra dựa trên độ chính xác của từng thuật toán.
 - Model sẽ tiến hành dự đoán cho từng loại URL và đưa kết quả vào một dictionary chứa accuracy rate và dữ liệu dự đoán cho trang web.

- Cuối cùng, trả về kết quả cho người dùng và hiển thị kết quả lên trình duyệt web
- Đối với ứng dụng Console: Cách thức hoạt động tương tự như 2 chức năng đã đề cập ở trước. Tuy nhiên ở việc dự đoán cho file csv chứa nhiều URL, dữ liệu đầu vào chính là đường dẫn đến file đó.

3.4 Đánh giá kết quả đạt được

Độ chính xác của thuật toán Random Forest



The screenshot shows a web browser window with the address bar displaying '127.0.0.1:5000/multi'. The page title is 'Phishing Detection'. The main content area is titled 'Phishing Url Detection' and features a dark, stylized background with a full moon and jagged rock formations. The interface includes a file upload section with the text 'Only accept csv file', a 'Choose File' button, and a file named 'urls_set3.csv' selected. Below this is an 'Options' dropdown menu set to 'Random Forest'. A 'Single URL' button is also present. The 'Accuracy: 100%' is displayed. There are 'Clear' and 'Submit' buttons. A table with 4 columns: 'Index', 'URL', 'Predict', and 'Right value' is shown. The table contains 7 rows of data, all correctly classified as 'Phishing Website' except for the last row which is 'Legimal Website'.

Index	URL	Predict	Right value
1	www.henkdeinumboomkwekerij.nl/language/pdf_fonts/smiles.php	Phising Website	Phising Website
2	perfectsolutionofall.net/wp-content/themes/twentyten/wiresource/	Phising Website	Phising Website
3	lingshc.com/old_aol.1.3/?Login=8&Is=10&LigertID=1993745&us=1	Phising Website	Phising Website
4	anonymidentity.net/remax/remax.htm	Phising Website	Phising Website
5	dutchweb.gtphost.com/zimbra/exch/owa/uleth/index.html	Phising Website	Phising Website
6	www.avedeoiro.com/site/plugins/chase/	Phising Website	Phising Website
7	https://facebook.com/	Legimal Website	Legimal Website

Độ chính xác của thuật toán Logistic Regression

Phishing Detection

Only accept csv file

Choose File

Options

Single URL

Accuracy: 90%

Clear Submit

Index	URL	Predict	Right value
1	www.henkdeinumboomkwekerij.nl/language/pdf_fonts/smiles.php	Phising Website	Phising Website
2	perfectsolutionofall.net/wp-content/themes/twentyten/wiresource/	Phising Website	Phising Website
3	lingshc.com/old_aol.1.3/?Login=&Lis=10&LigertID=1993745&us=1	Phising Website	Phising Website
4	anonymidentity.net/remax/remax.htm	Phising Website	Phising Website
5	dutchweb.gtpghost.com/zimbra/exch/owa/uleth/index.html	Legimal Website	Phising Website
6	www.avedeoiro.com/site/plugins/chase/	Phising Website	Phising Website
7	https://facebook.com/	Legimal Website	Legimal Website

Độ chính xác của thuật toán Support Vector Machine

Phishing Detection

Only accept csv file

Choose File

Options

Single URL

Accuracy: 70%

Clear Submit

Index	URL	Predict	Right value
1	www.henkdeinumboomkwekerij.nl/language/pdf_fonts/smiles.php	Phising Website	Phising Website
2	perfectsolutionofall.net/wp-content/themes/twentyten/wiresource/	Phising Website	Phising Website
3	lingshc.com/old_aol.1.3/?Login=&Lis=10&LigertID=1993745&us=1	Phising Website	Phising Website
4	anonymidentity.net/remax/remax.htm	Phising Website	Phising Website
5	dutchweb.gtpghost.com/zimbra/exch/owa/uleth/index.html	Phising Website	Phising Website
6	www.avedeoiro.com/site/plugins/chase/	Phising Website	Phising Website
7	https://facebook.com/	Phising Website	Legimal Website

Kết quả cho thấy thuật toán Random Forest cho độ chính xác phát hiện tốt hơn so với Logistic Regression và Support Vector Machine. Kết quả cũng cho thấy độ chính

xác của việc phát hiện các trang web giả mạo tăng lên khi có nhiều tập dữ liệu được sử dụng làm tập dữ liệu training. Tất cả các bộ phân loại hoạt động tốt khi 90% dữ liệu được sử dụng làm tập dữ liệu training, tuy nhiên trong trường hợp sử dụng quá nhiều dữ liệu cho tập training sẽ dẫn đến số lượng dữ liệu cho tập test ít đi dẫn đến đánh giá độ chính xác trong quá trình chạy thuật toán.

Đối với việc model dự đoán kết quả, thời gian xử lý khá nhanh nhưng với việc trích xuất các đặc điểm để tiến hành đưa vào model dự đoán thì thời gian xử lý cho nhiều URL là vô cùng lâu vì số đặc điểm cần trích xuất để dự đoán trang web là 30.

CHƯƠNG 4: KẾT LUẬN

Bài báo cáo này nhằm mục đích nâng cao phương pháp phát hiện để phát hiện các trang web giả mạo bằng cách sử dụng công nghệ học máy. Từ đó chúng ta đã thấy rằng lừa đảo là một mối đe dọa to lớn như thế nào đối với an ninh và an toàn của web và cách phát hiện lừa đảo là một vấn đề quan trọng.

Đối với các cải tiến trong tương lai, chúng em dự định xây dựng hệ thống phát hiện giả mạo dưới dạng một dịch vụ web có thể mở rộng, thêm các tính năng bằng cách phân tích hình ảnh, video và phân loại nội dung của các trang. Bên cạnh đó, cải thiện độ chính xác của mô hình chúng em bằng cách khai thác tính năng tốt hơn.

TÀI LIỆU THAM KHẢO

1. Rami M. Mohammad, Fadi Thabtah, Lee McCluskey, “Phishing Websites Features”
2. Atharva Deshpande, Omkar Pdamkar, Nachiket Chaudhary, Dr. Swapna Borde, “Detection of Phishing Websites using Machine Learning”
3. Abdul Razaque, Mohamed Ben Haj Frej, Dauren Sabyrov, Aidana Shaikhyn, Fathi Amsaad, Ahmed Oun, “Detection of Phishing Websites using Machine Learning”
4. Rishikesh Mahajan, Irfan Siddavatam, “Phishing Website Detection using Machine Learning Algorithms”
5. Tập dữ liệu URLs Phishing <https://www.kaggle.com/shashwatwork/web-page-phishing-detection-dataset>
6. Tập dữ liệu Phishing Feature <https://www.kaggle.com/eswarchandt/phishing-website-detector?select=phishing.csv>