**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

**CHUYÊN NGÀNH AN TOÀN THÔNG TIN**

---------------o0o---------------

****

**Báo cáo đồ án môn học**

**Ứng dụng học máy vào bài toán nhận diện mã độc JavaScript**

**GVHD: Nguyễn Đức Toàn**

**SV: Trần Đức Sơn 20163571**

**HÀ NỘI, 4/2020**

# Đặt vấn đề

Công nghệ web hiện nay đang chiếm vị trí quan trọng trong tổng thể ngành công nghệ thông tin. Ở bất cứ đâu con người đều có thể truy cập vào web thông qua các ứng dụng web, làm mọi thứ từ truy cập thông tin đến giải trí, học tập và làm việc. Công cụ đóng vai trò quan trọng, giúp chúng ta tương tác với Web và Internet là Trình duyệt (browser). Ở bất cứ hệ điều hành nào, ta đều bắt gặp ít nhất một browser được cài sẵn giúp người dùng thuận tiện khi muốn truy cập vào website.

Tuy nhiên, song song với sự phát triển, đi lên của trình duyệt internet thì những hình thức tấn công vào máy tính người dùng thông qua nền tảng này cũng được phát triển, mở rộng và đang là mối nguy lớn. Chủ yếu, nó được lợi dụng bằng cách cài đặt một số những mã nguồn độc hại bằng JavaScript nhằm thực thi chuỗi lệnh từ đó thực hiện hành vi độc hại mà kẻ tấn công đã lên kịch bản trước.

# Phân tích hành vi JavaScript malware

1. **Chèn thuộc tính ẩn vào document**

Những mẫu JavaScript loại này được cài đặt trong những trang web không được xác thực, không an toàn hay sử dụng giao thức HTTP để giao tiếp. Nó có cấu trúc như một đoạn mã JavaScript thông thường nhưng thay vì thực hiện theo mục đích của lập trình viên website thì nó lại mang những hành động xấu. Cụ thể là chèn trực tiếp vào document HTML một thẻ <iframe> với attribute display:none hoặc cài đặt attribute.width, attribute.height tương đối nhỏ, làm cho người dùng không nhận ra.

Mục đích của hành vi này là đánh lừa người dùng click vào đối tượng, hoặc tương tác với đối tượng mà không chủ ý. Dẫn đến sinh lợi cho attacker (thông quan chạy quảng cáo) hoặc thực hiện những hành vi nguy hiểm hơn (chuyển hướng đến trang web độc hại hay tải mã độc về máy tính người dùng), …

1. **Gọi hàm ẩn đối tượng trong document**

Những hàm ẩn này cũng có tác dụng tương tự như set lại thuộc tính ẩn. Tuy nhiên, do khả năng tương tác với mọi đối tượng của document nên có thể sử dụng để tương tác với tất cả các đối tượng trong document. Từ đó dễ dàng đặt lại thuộc tính cho đối tượng hơn.

1. **Thực hiện các câu lệnh jQuery**

Những câu lệnh JQuery có tính tương tác với các đối tượng trong document HTML, tác động đến nội dung file HTML, cấu trúc DOM, có thể lựa chọn những đối tượng trong DOM, trong CSS, hỗ trợ hiệu ứng, hoạt cảnh trong trang web, các phương thức HTML, chuyển đổi JSON, gọi hàm AJAX, các plug-ins, …

Những đoạn mã độc giả mạo những câu lệnh jQuery có tác động trực tiếp đến trang web mà người dùng truy cập đến

1. **Chuyển hướng trang web hiện tại**

Một mã độc JavaScript chuyển hướng được chèn vào một website với mục đích chuyển hướng người dùng tới một website khác. Nó thường được tạo ra bởi các attacker nhằm đáp ứng một nội dung quảng cáo. Tuy nhiên, một vài mã độc chuyển hướng kiểu này có hiệu ứng cao hơn. Nó exploit (khai thác) một lỗ hổng trong máy tính của người dùng đang truy cập vào website, thông qua đoạn mã để cài những loại mã độc mà không được bảo vệ ở phía người dùng, hoặc lợi dụng để cài đặt mã độc cùng loại chuyển hướng mà rất khó có thể loại bỏ trực tiếp khỏi máy người dùng.

Các phương thức tấn công của loại này:

* Chèn vào file trong trang web người dùng: ví dụ trong WordPress, nó được chèn vào các theme file. Thường được tìm thấy trong các theme header
* Chèn vào trang hoặc post
* Chèn vào các Widgets
* Chèn vào các file .htaccess dùng để cấu hình máy chủ web Apache, hoặc dùng quy định đường dẫn tới server mà tập tin của trang web được kết nối tới
* Ad networks: các mạng quản cáo được chấp nhận trong mạng, kẻ tấn công có thể lợi dụng nó để giả mạo hoặc chèn vào mạng một loại quảng cáo độc hại do hệ thống không có (hoặc ít) cơ chế để kiểm tra và quét.

1. Mã hóa file mã nguồn js hoặc sử dụng kỹ thuật Obfuscated javascript

Với mục đích ẩn giấu mã nguồn, không cho những người không có kỹ năng về phân tích mã nguồn có thể đọc được, kỹ thuật obfuscated ra đời. Tuy nhiên, vì đặc tính che dấu mã nguồn này nên đa số các javascript malware lợi dụng để che dấu mã nguồn thật đằng sau một chuỗi được encode, khiến cho những người làm về phân tích mã nguồn không thể trực tiếp tìm ra được đoạn code mã độc

1. **Theo dõi hành vi và vị trí con trỏ chuột trên trình duyệt**

Hành vi: Đoạn code JS tìm kiếm vị trí hiện tại của con trỏ chuột người dùng, sau đó tạo một vùng hoặc một thẻ tại ví trí đó, thẻ đó sẽ được tạo lại mỗi khi người dùng di chuyển con trỏ chuột

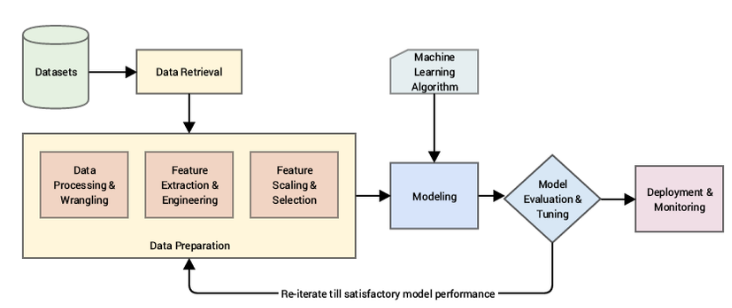
# Lý thuyết cơ bản

1. **Xử lý dữ liệu**

Vai trò của xử lý dữ liệu:

Bất kỳ bài toán học máy nào đều phải trải qua các bước từ dữ liệu thô thu được, sử dụng các kỹ thuật để sắp xếp, xử lý, thiết kế các đặc trưng (feature) và thuộc tính có ý nghĩa từ dữ liệu này. Sau đó mới sử dụng các mô hình thống kê và học máy để xây dựng và giải quyết các yêu cầu đặt ra.

Mô hình xử lý chung cho các bài toán học máy dựa theo mô hình tiêu chuẩn công nghiệp CRISP-DM



Nếu chỉ giữ nguyên dữ liệu thô mà đưa luôn vào mô hình xử lý thì hiệu quả đạt được của mô hình sẽ không cao.

Với mỗi vấn đề cụ thể luôn có những đặc trưng quan trọng quyết định đến hiệu suất hệ thống. Feature Engineering luôn chiếm tới 70% thời gian trong quá trình xây dựng mô hình.

Xử lý dữ liệu trích chọn đặc trưng là quá trình biến đổi dữ liệu thành các đặc trưng đóng vai trò là đầu vào cho các mô hình học máy. Các đặc trưng xử lý tốt sẽ nâng cao hiệu suất mô hình. Tùy vào những dự án khác nhau mà những đặc trưng được lựa chọn cũng sẽ khác nhau.

1. **Thuật toán Machine Learning**

* Thuật toán Decision trees
* Thuật toán random forests

# Giải pháp đề xuất

1. Thử nghiệm với thuật toán Machine Learning
2. Sử dụng mạng neuron

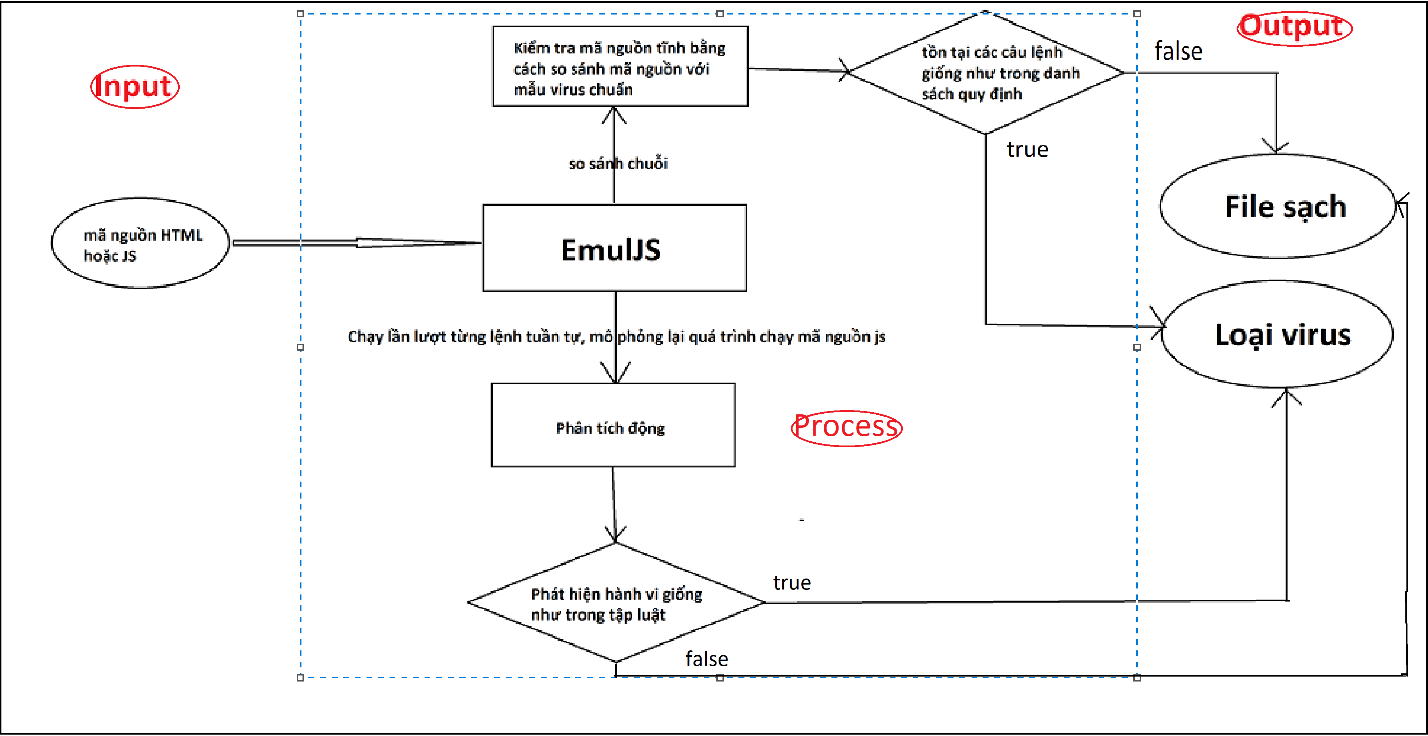
# Công nghệ sử dụng

1. **Giới thiệu về Engine EmulJS**

Engine EmulJS hiện đang được triển khai để nhận diện mẫu virus trên VirusTotal của Bkav. Công cụ này sử dụng kết hợp 2 phương pháp phát hiện code tĩnh và phân tích code động để tìm hành vi độc hại, từ đó sẽ trả về kết quả loại virus tùy theo hành vi phân tích được

Hiện tại em đang chỉnh sửa công cụ này để thay vì trả về kết quả virus như bình thường thì sẽ chỉ có tác dụng chạy động virus để thu thập những hành vi đặc trưng của virus từ đó lấy được dữ liệu output làm đầu vào cho mô hình học máy

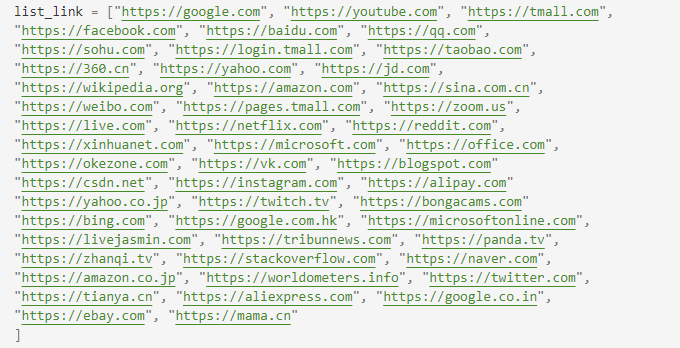
Luồng xử lý của EmulJS:



1. **Giới thiệu về Tool Crawl mẫu HTML sạch**

Công cụ được viết bằng ngôn ngữ Python và sử dụng các thư viện làm việc với trang web và hỗ trợ crawl dữ liệu về như: urllib (download các trang web theo định dạng HTML), BeautifulSoup, time, os

Nguồn link ban đầu bao gồm:



Trong quá trình duyệt trang web, nếu thấy link Website khác với danh sách hoặc không nằm trong những link đã download thì sẽ thêm vào list để download tiếp. Tạo một list lưu danh sách những trang web đã crawl dữ liệu, danh sách sẽ tự động refresh lại về rỗng khi kích thước lớn hơn 2000 link. Có thể điều chỉnh kích thước nếu thực tế không hiệu quả.

Sau khi cho chạy lệnh này bằng python, mã nguồn HTML của các trang web được duyệt sẽ được lưu vào một thư mục làm đầu vào cho Tool Heuristic FS để trích xuất thông tin cần thiết ra file text

1. **Giới thiệu về công cụ Heuristic FS**

Công cụ sử dụng dll được tạo ra từ emulJS để quét offline một tập mẫu JavaScript đầu vào. Có thể tùy chỉnh đầu ra khi chạy công cụ bằng cách sửa mã nguồn của dll EmulJS

1. **Giới thiệu về nguồn lấy mẫu JavaScript và công cụ xử lý liên quan**

Nguồn lấy mẫu: VirusTotal

* VirusTotal là một dịch vụ miễn phí phân tích tập tin và URL nghi ngờ, nó có chứa một kho dữ liệu bao gồm rất nhiều mẫu JavaScript malware. Mỗi engine dùng để phân tích được hỗ trợ bởi cộng đồng các công cụ phân tích và diệt mã độc nổi tiếng trên thế giới. Mỗi một công ty cung cấp dịch vụ như vậy đều được cấp một tài khoản có thể truy cập và tải về kho tập mẫu của VirusTotal chứa nội dung Virus và một file log chứa thông tin các engine phát hiện.

1. **Giới thiệu về Jupyter Notebook**

Jupyter Notebook được cài sẵn trong Anaconda cung cấp một giao diện sử dụng ngôn ngữ python được cài sẵn trong máy để chạy code. Nó có ưu điểm là giao diện trực quan, chạy code theo từng dòng lệnh, có thể lưu lại các checkpoint, …

Jupyter Notebook được sử dụng là bản được cài kèm cùng với bản Anaconda 3 mới nhất.

# Triển khai

# Kết luận

Sử dụng thuật toán học máy Random Forest có thể đem đến một số kết quả nhất định. Tuy nhiên, thực tế cho thấy thuật toán này có nhiều điểm hạn chế. Đầu tiên là nó phụ thuộc quá nhiều vào dữ liệu đầu vào. Trường hợp nếu tập mẫu