Bộ dữ liệu UNSW-NB15 thuộc lĩnh vực an ninh mạng và dùng để giải bài toán "Nhận biết tấn công mạng". Mỗi sample (mẫu) trong dữ liệu có thể được hiểu như sau:

Giải thích về dữ liệu

1. **Mỗi sample:**
   - **Một mẫu trong dữ liệu này đại diện cho một phiên giao tiếp mạng (network session) hoặc một kết nối giữa các hệ thống mạng.**
   - **Mỗi kết nối chứa thông tin về thời gian, loại giao thức, dữ liệu trao đổi và kết quả hoạt động.**
2. **Các nhóm đặc trưng chính:**
   - **Thông tin cơ bản:**
     - **`id`: ID duy nhất của phiên giao tiếp.**
     - **`dur`: Thời gian phiên giao tiếp kéo dài tính bằng giây (duration).**
     - **`proto`: Giao thức mạng sử dụng, ví dụ: TCP, UDP, …**
     - **`service`: Dịch vụ được sử dụng (ví dụ: HTTP, FTP).**
     - **`state`: Trạng thái của kết nối, ví dụ: FIN (Finished), INT (Interrupted)**
   - **Thông tin về gói dữ liệu:**
     - **`spkts`, `dpkts`: Số lượng gói tin được gửi (`spkts`) và nhận (`dpkts`).**
     - **`sbytes`, `dbytes`: Số lượng byte được gửi (`sbytes`) và nhận (`dbytes`).**
   - **Thông tin tốc độ và tải:**
     - **`rate`: Tốc độ truyền tải của kết nối. (byte/s)**
     - **`sload`, `dload`: Tải dữ liệu từ nguồn (`sload`) và đích (`dload`). (byte/s)**
   - **Thông tin thời gian và tổn thất:**
     - **`sinpkt`, `dinpkt`: Thời gian giữa các gói tin gửi (`sinpkt`) và nhận (`dinpkt`). (s)**
     - **`sloss`, `dloss`: Số lượng gói tin bị mất khi gửi (`sloss`) và nhận (`dloss`).**
   - **Thông tin giao thức TCP:**
     - **`sttl`, `dttl`: Time-to-live (TTL) từ nguồn (`sttl`) và đích (`dttl`). (s)**
     - **`tcprtt`, `synack`, `ackdat`: Các đặc trưng liên quan đến độ trễ và xác nhận trong giao thức TCP.**
   - **Đặc trưng thống kê:**
     - **`smean`, `dmean`: Trung bình số byte được gửi (`smean`) và nhận (`dmean`).**
     - **`trans_depth`: Độ sâu truyền tải.**
     - **`response_body_len`: Độ dài nội dung phản hồi.**
   - **Thông tin về nguồn và đích:**

        - **`ct_srv_src`, `ct_srv_dst`: Số lượng dịch vụ từ nguồn và đích.**
        - **`ct_dst_ltm`, `ct_src_ltm`: Lưu lượng đến từ nguồn và đích theo thời gian.**
    - **Các thông tin khác:**
        - **`is_ftp_login`: Có phải là một phiên đăng nhập FTP không (True/False).**
        - **`ct_ftp_cmd`: Số lượng lệnh FTP trong phiên.**
        - **`ct_flw_http_mthd`: Số phương thức HTTP trong luồng.**
    - **Nhãn và loại tấn công:**
        - **`attack_cat`: Danh mục tấn công (ví dụ: DoS, Probe, Fuzzers, Worms...).**
        - **`label`: Nhãn cho biết đây là tấn công (1) hay kết nối hợp lệ (0).**

Giải thích chi tiết các cột quan trọng

Proto

Các giá trị trong cột `proto` đại diện cho các giao thức mạng (protocols) được sử dụng trong các kết nối mạng. Dưới đây là giải thích về một số giao thức phổ biến trong danh sách:

1. **`udp` (User Datagram Protocol):**
    - **Giao thức kết nối không tin cậy, chủ yếu được sử dụng cho các ứng dụng yêu cầu tốc độ cao và chấp nhận mất gói (như video streaming, gaming).**
2. **`tcp` (Transmission Control Protocol):**
    - **Giao thức kết nối đáng tin cậy, đảm bảo truyền dữ liệu chính xác và theo thứ tự. Phổ biến trong các ứng dụng web, email, và truyền file.**
3. **`arp` (Address Resolution Protocol):**
    - **Dùng để ánh xạ địa chỉ IP thành địa chỉ MAC trong mạng LAN.**
4. **`igmp` (Internet Group Management Protocol):**
    - **Dùng để quản lý thành viên của các nhóm multicast trong mạng.**
5. **`ospf` (Open Shortest Path First):**
    - **Giao thức định tuyến nội miền (Interior Gateway Protocol) sử dụng thuật toán đường đi ngắn nhất.**
6. **`sctp` (Stream Control Transmission Protocol):**
    - **Giao thức truyền tải tin cậy, hỗ trợ truyền tải nhiều luồng trong một kết nối.**
7. **`gre` (Generic Routing Encapsulation):**
    - **Dùng để encapsulate dữ liệu cho các mạng riêng ảo (VPN).**
8. **`ipv6` (Internet Protocol version 6):**

- **Phiên bản mới hơn của IPv4, cung cấp địa chỉ IP lớn hơn và cải thiện các tính năng định tuyến.**

Ngoài ra, danh sách này còn chứa nhiều giao thức khác ít phổ biến hơn, bao gồm:

- **`igp` (Interior Gateway Protocol): Giao thức định tuyến nội miền.**
- **`pim` (Protocol Independent Multicast): Dùng để định tuyến dữ liệu multicast.**
- **`vrrp` (Virtual Router Redundancy Protocol): Cung cấp dự phòng router để đảm bảo kết nối mạng liên tục.**

Ý nghĩa

Mỗi giá trị của cột `proto` thể hiện giao thức được sử dụng trong một kết nối hoặc phiên giao tiếp cụ thể. Phân tích cột này có thể cung cấp thông tin về loại giao thức mạng được sử dụng nhiều nhất trong tập dữ liệu, qua đó hiểu thêm về hành vi của hệ thống mạng.

Service

Cột `service` trong tập dữ liệu biểu thị loại dịch vụ mạng liên quan đến mỗi phiên giao tiếp. Đây là những dịch vụ hoặc giao thức ứng dụng được sử dụng trong mạng. Dưới đây là giải thích từng giá trị:

---

1. `-` (Không xác định hoặc không áp dụng):

- Đại diện cho các kết nối mà không có dịch vụ rõ ràng được nhận diện.
- Có thể là các kết nối mạng không có ứng dụng hoặc dữ liệu dịch vụ cụ thể.

---

2. `http` (HyperText Transfer Protocol):

- Giao thức truyền tải web, được sử dụng để truyền dữ liệu giữa máy khách (browser) và máy chủ web.
- Thường dùng cho các dịch vụ web như trang web tĩnh hoặc động.

---

3. `ftp` (File Transfer Protocol):

- Giao thức truyền tải file, cho phép tải lên và tải xuống file giữa máy khách và máy chủ.
- Được sử dụng phổ biến trong các hệ thống chia sẻ file.

---

4. `ftp-data`:

- Liên quan đến dữ liệu thực tế được truyền trong một phiên FTP.
- Thường chạy trên một kênh riêng biệt với kết nối điều khiển FTP.

---

5. `smtp` (Simple Mail Transfer Protocol):

- Giao thức chính để gửi email qua internet.
- Máy chủ email sử dụng SMTP để gửi thư từ máy khách hoặc chuyển tiếp giữa các máy chủ.

---

6. `pop3` (Post Office Protocol version 3):

- Giao thức để lấy email từ máy chủ về máy khách.
- Thường sử dụng trong các dịch vụ email truyền thống.

---

7. `dns` (Domain Name System):

- Giao thức phân giải tên miền thành địa chỉ IP.
- Một trong những giao thức cơ bản của internet.

---

8. `snmp` (Simple Network Management Protocol):

- Dùng để quản lý và giám sát thiết bị mạng như router, switch, server.
- Thường sử dụng trong quản trị mạng.

---

9. `ssl` (Secure Sockets Layer):

- Giao thức bảo mật, được sử dụng để mã hóa giao tiếp giữa máy khách và máy chủ.

- Phổ biến trong các trang web có HTTPS.

---

10. `dhcp` (Dynamic Host Configuration Protocol):

- Giao thức phân phối địa chỉ IP động và các thông tin cấu hình mạng cho thiết bị trong mạng.

---

11. `irc` (Internet Relay Chat):

- Giao thức nhắn tin thời gian thực, thường dùng cho các phòng chat và nhắn tin nhóm.

---

12. `radius` (Remote Authentication Dial-In User Service):

- Giao thức xác thực và quản lý quyền truy cập từ xa.
- Thường dùng trong các mạng doanh nghiệp để kiểm soát quyền truy cập.

---

13. `ssh` (Secure Shell):

- Giao thức bảo mật để truy cập từ xa đến máy chủ và thực hiện các lệnh.
- Được sử dụng rộng rãi trong quản trị hệ thống.

Ý nghĩa

- Các giá trị trong cột `service` cung cấp thông tin về dịch vụ hoặc ứng dụng liên quan đến mỗi phiên giao tiếp mạng.
- Phân tích cột này giúp xác định các dịch vụ phổ biến trong tập dữ liệu và có thể giúp phát hiện bất thường trong việc sử dụng dịch vụ (như các cuộc tấn công mạng liên quan đến HTTP, SSH, hoặc FTP).

State

Cột `state` trong tập dữ liệu biểu thị trạng thái của kết nối mạng hoặc phiên giao tiếp trong mạng. Các trạng thái này thường được xác định dựa trên các gói tin và sự tương tác trong kết nối TCP/IP hoặc các giao thức khác. Dưới đây là giải thích từng giá trị:

---

1. `INT` (Interrupted - Gián đoạn):

- Trạng thái cho biết kết nối hoặc phiên giao tiếp bị gián đoạn.
- Có thể xảy ra do lỗi mạng, ngắt kết nối hoặc các vấn đề khác.

---

2. `FIN` (Finished - Kết thúc):

- Trạng thái cho biết kết nối đã được đóng lại một cách đúng quy trình.
- Thường được kích hoạt bởi một thông báo `FIN` trong giao thức TCP để kết thúc một phiên.

---

3. `REQ` (Request - Yêu cầu):

- Thể hiện rằng một yêu cầu đã được gửi từ một thiết bị (máy khách) đến một thiết bị khác (máy chủ).
- Thường xuất hiện trong giao thức như HTTP, FTP hoặc DNS.

---

4. `ACC` (Accepted - Được chấp nhận):

- Trạng thái cho biết một yêu cầu đã được chấp nhận thành công.
- Ví dụ: một kết nối TCP được thiết lập hoặc yêu cầu dữ liệu được xác nhận.

---

5. `CON` (Connected - Đã kết nối):

- Thể hiện rằng một kết nối đã được thiết lập thành công giữa hai thiết bị.
- Thường là trạng thái trung gian trong một phiên TCP đang hoạt động.

---

6. `RST` (Reset - Đặt lại):

- Trạng thái cho biết kết nối đã bị đặt lại hoặc kết thúc bất thường.
- Giao thức TCP sử dụng gói tin `RST` để báo hiệu rằng kết nối không còn hợp lệ.

---

7. `CLO` (Closed - Đóng):

- Trạng thái cho biết kết nối hoặc phiên đã được đóng lại.
- Thường xảy ra sau khi quá trình trao đổi dữ liệu hoàn tất.

Ý nghĩa

- **Phân tích cột `state`** giúp hiểu hành vi của các kết nối mạng trong tập dữ liệu.
- Các trạng thái bất thường (như `RST` hoặc `INT`) có thể chỉ ra các sự cố mạng hoặc dấu hiệu của một cuộc tấn công.
- Các trạng thái phổ biến (như `REQ`, `ACC`, `CON`) cho thấy sự hoạt động bình thường của các dịch vụ mạng.

Phân phối các trạng thái này có thể cung cấp thông tin hữu ích cho việc phát hiện tấn công hoặc xác định sự bất thường trong mạng.

State

Cột `state` trong tập dữ liệu biểu thị trạng thái của kết nối mạng hoặc phiên giao tiếp trong mạng. Các trạng thái này thường được xác định dựa trên các gói tin và sự tương tác trong kết nối TCP/IP hoặc các giao thức khác. Dưới đây là giải thích từng giá trị:

---

1. `INT` (Interrupted - Gián đoạn):

- Trạng thái cho biết kết nối hoặc phiên giao tiếp bị gián đoạn.
- Có thể xảy ra do lỗi mạng, ngắt kết nối hoặc các vấn đề khác.

---

2. `FIN` (Finished - Kết thúc):

- Trạng thái cho biết kết nối đã được đóng lại một cách đúng quy trình.
- Thường được kích hoạt bởi một thông báo `FIN` trong giao thức TCP để kết thúc một phiên.

---

3. `REQ` (Request - Yêu cầu):

- Thể hiện rằng một yêu cầu đã được gửi từ một thiết bị (máy khách) đến một thiết bị khác (máy chủ).
- Thường xuất hiện trong giao thức như HTTP, FTP hoặc DNS.

---

4. `ACC` (Accepted - Được chấp nhận):

- Trạng thái cho biết một yêu cầu đã được chấp nhận thành công.
- Ví dụ: một kết nối TCP được thiết lập hoặc yêu cầu dữ liệu được xác nhận.

---

5. `CON` (Connected - Đã kết nối):

- Thể hiện rằng một kết nối đã được thiết lập thành công giữa hai thiết bị.
- Thường là trạng thái trung gian trong một phiên TCP đang hoạt động.

---

6. `RST` (Reset - Đặt lại):

- Trạng thái cho biết kết nối đã bị đặt lại hoặc kết thúc bất thường.
- Giao thức TCP sử dụng gói tin `RST` để báo hiệu rằng kết nối không còn hợp lệ.

---

7. `CLO` (Closed - Đóng):

- Trạng thái cho biết kết nối hoặc phiên đã được đóng lại.
- Thường xảy ra sau khi quá trình trao đổi dữ liệu hoàn tất.

Ý nghĩa

- **Phân tích cột `state`** giúp hiểu hành vi của các kết nối mạng trong tập dữ liệu.
- Các trạng thái bất thường (như `RST` hoặc `INT`) có thể chỉ ra các sự cố mạng hoặc dấu hiệu của một cuộc tấn công.

- Các trạng thái phổ biến (như `REQ`, `ACC`, `CON`) cho thấy sự hoạt động bình thường của các dịch vụ mạng.

Phân phối các trạng thái này có thể cung cấp thông tin hữu ích cho việc phát hiện tấn công hoặc xác định sự bất thường trong mạng.

Attack_cat

Cột **`attack_cat`** trong dữ liệu của bạn đại diện cho các loại tấn công mạng được phân loại theo từng nhóm. Dưới đây là giải thích chi tiết về mỗi loại tấn công trong cột này:

---

1. `Normal` (Bình thường):

- **Mô tả:** Không phải là tấn công, mà là hành vi mạng hợp pháp và bình thường.
- **Giải thích:** Đây là những kết nối hoặc giao tiếp mạng bình thường giữa các thiết bị, không liên quan đến bất kỳ hành vi xâm nhập hay tấn công nào. Dữ liệu trong nhóm này sẽ không chứa các dấu hiệu của các cuộc tấn công mạng.

---

2. `Reconnaissance` (Do thám):

- **Mô tả:** Các hành vi do thám hoặc quét mạng để thu thập thông tin trước khi thực hiện một cuộc tấn công.
- **Giải thích:** Các cuộc tấn công loại này chủ yếu tập trung vào việc tìm kiếm các mục tiêu tiềm năng hoặc điểm yếu trong hệ thống. Các kỹ thuật phổ biến trong nhóm này bao gồm:
    - **Port scanning** (quét cổng) để xác định các cổng mở trên một hệ thống.
    - **Network sniffing** để thu thập thông tin về các giao tiếp mạng.
    - **Ping sweeps** để kiểm tra sự tồn tại của các thiết bị trên mạng.

---

3. `Backdoor` (Cổng hậu):

- **Mô tả:** Tấn công cài đặt một "cổng hậu" để truy cập vào hệ thống mà không bị phát hiện.
- **Giải thích:** Các tấn công backdoor liên quan đến việc cài đặt phần mềm độc hại (malware) vào một hệ thống, cho phép kẻ tấn công truy cập vào hệ thống đó mà không cần thông qua các phương thức bảo mật thông thường. Các backdoor có thể được sử dụng để thực hiện các hành động sau:

    - Truy cập từ xa không có sự giám sát.
    - Đánh cắp thông tin hoặc thực hiện các hoạt động xấu mà không bị phát hiện.

---

4. `DoS` (Denial of Service - Từ chối Dịch vụ):

- **Mô tả:** Tấn công làm tê liệt hoặc gián đoạn dịch vụ hoặc hệ thống.
- **Giải thích:** Tấn công từ chối dịch vụ (Denial of Service - DoS) là các cuộc tấn công nhằm mục đích làm cho một dịch vụ hoặc hệ thống không thể sử dụng được. Mục tiêu có thể là:
    - **Overloading** (quá tải) máy chủ hoặc mạng bằng cách gửi quá nhiều yêu cầu hoặc dữ liệu.
    - **Flooding** (ngập lụt) hệ thống với gói tin để làm tắc nghẽn và gây ra sự gián đoạn dịch vụ.

---

5. `Exploits` (Lợi dụng Lỗ hổng):

- **Mô tả:** Tấn công lợi dụng các lỗ hổng bảo mật trong phần mềm hoặc hệ thống.
- **Giải thích:** Tấn công khai thác lỗ hổng (exploits) là hành động lợi dụng các điểm yếu trong hệ thống hoặc phần mềm để truy cập trái phép hoặc làm hỏng hệ thống. Các lỗ hổng này có thể là:
    - Lỗ hổng bảo mật trong phần mềm (ví dụ: lỗi trong mã nguồn).
    - Lỗ hổng trong giao thức mạng (ví dụ: các giao thức không được bảo vệ hoặc mã hóa).
- Các tấn công này có thể bao gồm việc chiếm quyền điều khiển hệ thống, xâm nhập vào dữ liệu hoặc cài đặt mã độc.

---

6. `Analysis` (Phân tích):

- **Mô tả:** Các hành vi nhằm phân tích và kiểm tra hệ thống để tìm hiểu cách hoạt động.
- **Giải thích:** Các cuộc tấn công phân tích thường không gây ra thiệt hại ngay lập tức mà mục tiêu là thu thập thông tin về cấu trúc, phần mềm hoặc các lỗ hổng bảo mật của hệ thống. Điều này có thể bao gồm:
    - **Vulnerability scanning** (quét lỗ hổng) để tìm điểm yếu.
    - **Social engineering** (tấn công kỹ thuật xã hội) để thu thập thông tin từ con người.

---

7. `Fuzzers` (Kiểm tra độ bền):

- **Mô tả:** Sử dụng các kỹ thuật fuzzing để kiểm tra độ ổn định của phần mềm và tìm kiếm các lỗi.
- **Giải thích:** Tấn công bằng kỹ thuật **fuzzing** là phương pháp gửi dữ liệu ngẫu nhiên (hoặc không hợp lệ) vào các ứng dụng phần mềm để phát hiện lỗi hoặc các điểm yếu bảo mật. Mục tiêu của fuzzing là:
    - Tìm các lỗi có thể bị khai thác.
    - Làm cho phần mềm gặp sự cố hoặc bị treo.

---

8. `Worms` (Sâu máy tính):

- **Mô tả:** Phần mềm độc hại tự nhân bản và lây lan qua các mạng.
- **Giải thích: Worms** là các chương trình độc hại có thể tự sao chép và lây lan qua mạng mà không cần sự can thiệp của người dùng. Chúng có thể tấn công và lây lan rất nhanh chóng, gây tắc nghẽn mạng và làm hỏng hệ thống.

---

9. `Shellcode` (Mã Shell):

- **Mô tả:** Mã thực thi được thiết kế để khai thác lỗ hổng và thực hiện các lệnh trên hệ thống mục tiêu.
- **Giải thích: Shellcode** là đoạn mã được sử dụng để thực thi lệnh shell (giao diện dòng lệnh) trên hệ thống mục tiêu. Mã này thường được sử dụng trong các cuộc tấn công khai thác lỗ hổng để chạy các lệnh trái phép trên hệ thống.

---

10. `Generic` (Chung):

- **Mô tả:** Các tấn công không thuộc bất kỳ loại nào cụ thể trong danh sách.
- **Giải thích:** Tấn công **generic** có thể là các hành vi hoặc kỹ thuật không dễ dàng phân loại vào một nhóm cụ thể nào. Đây có thể là các cuộc tấn công phức tạp hoặc những cuộc tấn công mới, chưa được phân loại.

Ý Nghĩa

Các loại tấn công trong cột `attack_cat` phản ánh các chiến thuật khác nhau mà kẻ tấn công có thể sử dụng để xâm nhập vào hệ thống hoặc làm gián đoạn các dịch vụ mạng. Các tấn công

này có thể bao gồm từ các cuộc tấn công cơ bản (như tấn công từ chối dịch vụ - DoS) đến các chiến thuật phức tạp hơn như tấn công qua các backdoor hoặc khai thác các lỗ hổng bảo mật. Việc phân loại các tấn công giúp nhận diện và bảo vệ các hệ thống khỏi các mối đe dọa mạng.