Network Security
Project 2: Log Analysis

Instructor: Shiuhpyng Shieh
TA: Wei-Tung Tai, Li-Ren Zhang, Wei-Gang Sun
Email: TA@dsns.css.nctu.edu.tw

Due date: 23:55, December 13, 2020

1 Project Description

The goal of this project is to practice cyber attack log analysis. This is a personal project, each of you will be provided with 5 sets of data collected when different attacks are carried out. You have to implement a classifier to classify different attack scenarios. You can either observe the logs and then write a **rule-based model** or use **machine learning method** to tune a best-result model. You are encouraged to use ELK stack in project1 to perform the analysis.

2 Project Guide

1. **Project Target:**
   The goal of this project is to practice analyzing logs. Log can be analyzed in many different ways for different purposes, and for this project, you are practicing classifying attack type with simple attack cases. In this project, you need to first observe the logs, then pre-process the logs, and then finally construct a classification model.
2. **Keywords:**
   (a) Log Analysis:
   Log analysis is the process of making sense of computer-generated log messages, also known as log events, audit trail records, or simply logs. Log analysis provides useful metrics that paint a clear picture of what has happened across the infrastructure. You can use this data to improve or solve performance issues within an application or infrastructure. Looking at the bigger picture, companies analyze logs to proactively and reactively mitigate risks, comply with security policies, audits, and regulations, and understand online user behavior.

   (b) Machine Learning:
   Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks.

   (c) Cyber Attack:
   In computers and computer networks an attack is any attempt to expose, alter, disable, destroy, steal or gain unauthorized access to or make unauthorized use of an asset.[1] A cyberattack is any type of offensive maneuver that targets computer information systems, infrastructures, computer networks, or personal computer devices. An attacker is a person or

process that attempts to access data, functions or other restricted areas of the system without authorization, potentially with malicious intent.

3 Attack Scenarios

1. **Attack Scenario Description:**

   - **Port Scan:**
     The logs are collected from a host that was attacked by port scan.
     A port scan is an attack that sends client requests to a range of server port addresses on a host, with the goal of finding an active port and exploiting a known vulnerability of that service. Scanning, as a method for discovering exploitable communication channels, has been around for ages.

   - **SQL Injection:**
     The logs are collected from a web server that has a SQL database vulnerability and was attacked by SQL injection.
     SQL injection is a code injection technique, used to attack data-driven applications, in which malicious SQL statements are inserted into an entry field for execution (e.g. to dump the database contents to the attacker). For example, when user input is either incorrectly filtered for string literal escape characters embedded in SQL statements or user input is not strongly typed and unexpectedly executed.

   - **Brute-Force attack:**
     The logs are collected from a web server that was attacked by brute-force attack. The attacker attempted to guess the username and password.
     A brute-force attack consists of an attacker submitting many passwords or passphrases with the hope of eventually guessing correctly.

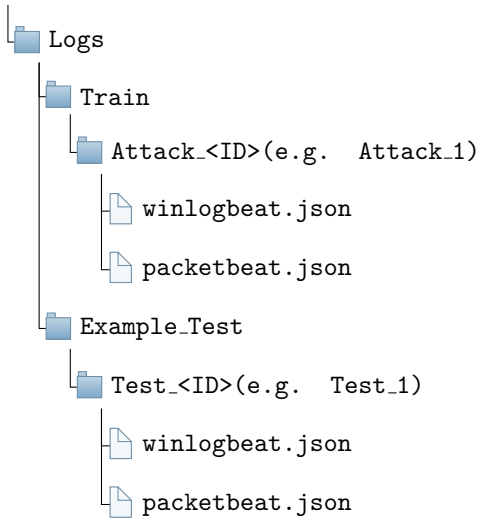
   - **DDoS:**
     The logs are collected from a web server that was attacked by Dos. In other words, the web service was overwhelmed by numerous requests. The web service is running on the default port 80.
     Denial of service(Dos) is typically accomplished by flooding the targeted machine or resource with superfluous requests in an attempt to overload systems and prevent some or all legitimate requests from being fulfilled.

   - **Phishing Email(Malicious Attachment):**
     The log is collected by a host that accidentally downloaded a malicious pdf attachment from an email. The attacker used the pdf file to attack the Adobe acrobat vulnerability and managed to exploit the CVE. After obtaining control of cmd.exe, the attacker searched for the desired data. Finally, compressed the data with tar and transferred to their own server.

2. **Given data:**
   You are given 5 attacks' cases as training data to design your model. **A case is a type of attack** from the list above and there are no repetitive attacks in the cases. Each case contains two log files collected by Winlogbeat and Packetbeat. **The case ID has nothing to do with the order of the above attack list.**

4 Coding Regulations

1. **Log Structure:**



   (a) **Train:** Contain 5 sets of train data separate with ID.

   (b) **Example_Test:** The testing samples we provided(Test_1, Test_2) are the corresponding sets for Training data(Attack_1, Attack_2)

2. **Code Input/Output:**

   (a) **Input:** Example_Test

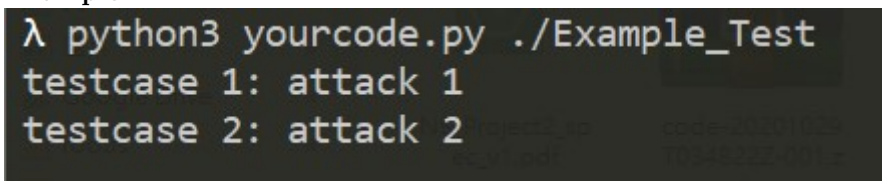
   (b) **Output:** Each test case result should be shown in the end.

   (c) **Notice:**

      i. Your code need to input file_path as an argument. Example: python3 yourcode.py `<FILE_PATH>`

      ii. Testing data will be stored like the above structure, you need to read in all the files in the directory to do the testing part.

   (d) **Example:**



3. **WARNING:** If you don't follow the coding regulations you'll get ZERO point.

5 What to Submit?

A report in PDF format, contains:

1. What model or algorithm you use? Please describe in as much detail as possible.
2. Anything interesting you find or problems you encounter in the whole process.

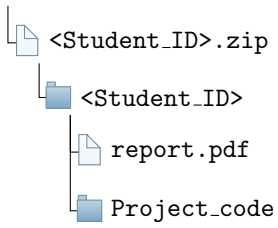
A model folder, contains:

1. Source code of your model.

2. A README file in which you explain how your project runs.

6 How to Submit?

- Compress your report PDF and project code (model folder) into a zip file. Upload the zip file as <Student_ID>.zip " to E3 platform.



7 Demo

- Demo period will be announced on E3 later, please pay attention to our announcement and fill in the demo period table.
- Please bring a computer with you to demo your project.
- We'll test your model with new test cases that we didn't provide.

*** 10 points will be deducted for handing in wrong file format.**