Network Security
Project 2: User Behavior Analysis

Instructor: Shiuhpyng Shieh
TA: Jo-Hsuan Huang, Kuan-Hsien Li, Po-Hao Huang, Po-Yu Huang
Email: TA@dsns.css.nctu.edu.tw

Due date: 23:59, June 17, 2020

1 Project Description

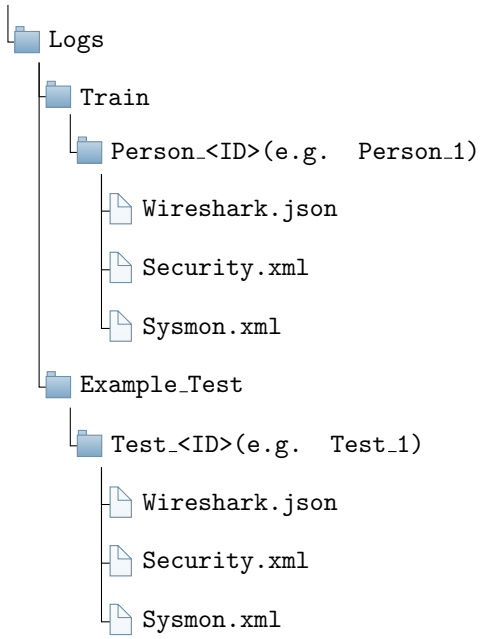
The goal of this project is to practise user behavior analysis. This is a group project, each group will be provided with 6 sets of data collected from different users with different behaviors, you have to write a classifier to classify different users. You can either observe the logs and then write a **rule-based model** or use **machine learning method** to tune a best-result model.

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 practising a simple UBA case to classify users. 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) User Behavior analytics(UBA):
   User behavior analytics (UBA) as defined by Gartner is a cybersecurity process about detection of insider threats, targeted attacks, and financial fraud. UBA solutions look at patterns of human behavior, and then apply algorithms and statistical analysis to detect meaningful anomalies from those patterns—anomalies that indicate potential threats. Instead of tracking devices or security events, UBA tracks a system's users. Big data platforms like Apache Hadoop are increasing UBA functionality by allowing them to analyze petabytes worth of data to detect insider threats and advanced persistent threats.
   https://en.wikipedia.org/wiki/User_behavior_analytics
   (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.
   https://en.wikipedia.org/wiki/Machine_learnings

3 Coding Regulations

1. **Log Structure:**



   (a) **Train:** Contain 6 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(Person_1, Person_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(there might be multiple files).

   (d) **Example:**

```
$ python3 yourcode.py ./Example_Test
testcase 1: person 1
testcase 2: person 2
```

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

4 What to Submit?

A report in PDF format, contains:

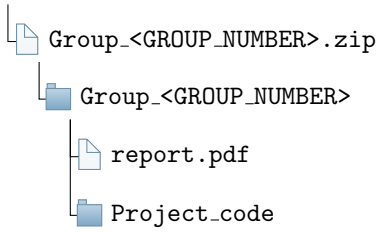
1. What model or algorithm you use? Please describe in as much detail as possible.
2. Anything interesting you find or problems you encounter in the whole process.

A model folder, contains:

1. Source code of your model.
2. A README file in which you explain how your project runs.

5 How to Submit?

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



6 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.

The penalty for late submission is 10% per day, and 10 points will be deducted for handing in wrong file format.