Machine learning cybersecurity

**Log file vectorization**

# LAB 1: Writing a script to vectorize log file

**Lab Description:** This lab is to vectorize the log files in order to get the features and labels for supervise learning.

Example of log file:



You are required to write a python script to vectorize the log files

**Lab Environment:**

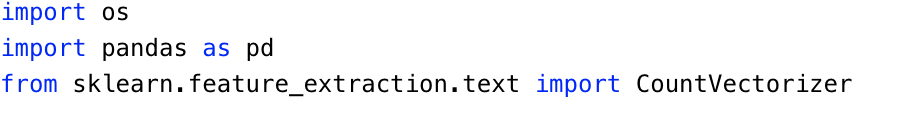
* The students should have access to a machine with Linux system or Windows system
* The environment for python is required as well as some packages such as numpy, tensorflow, pandas and sklearn.

**Lab Files that are Needed:**

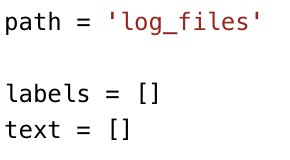
* For this lab you will need several log files, including the log files for normal activity and attack activity.

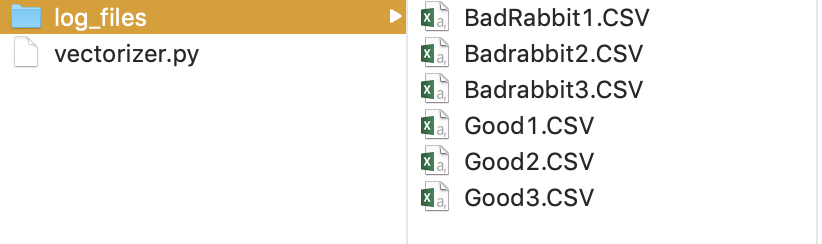
### **Lab exercise 1**

* Import the required libraries.

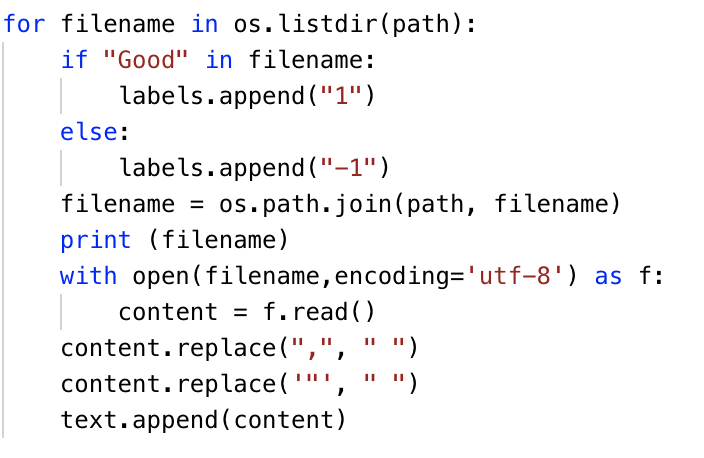


* Define the parameter to store the path for script to read data. And define the parameters to store the labels and text to be vectorized.
  + The name of the files will tell you whether it belongs to a normal activity or not.

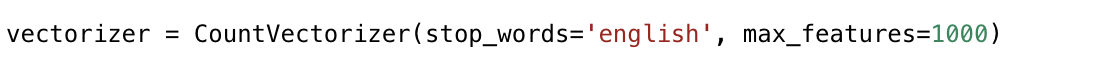




* Read the content from each file and create labels for them



* + If the log file belongs to a normal activity, "1" will be assigned to it as a label. Otherwise, "-1" will be assigned to indicate it belongs to an attack activity.
  + In order to perform vectorization, some characters such as comma and quotation mark need to be removed.
* Use CountVectorizer() to create a vectorizer for the text of log files. After the vectorization, you will be able to get the features and feature names of the content.



* stop\_words='english' indicates that all the stop words in the content will be removed.
* max\_features=1000 indicates that 1000 features will be generated based on the frequency order of the terms in text.
* Save the features to a csv file by using pandas DataFrame.to\_csv() function.
  + You may first need to covert the results to a dataframe
  + The index of dataframe should be the labels
  + Vectorizer can give you the names of the features, and you can use them as the column names for the dataframe.

## What to Submit

You should submit a lab report file which includes:

* + The steps you processed data
  + The necessary code snippet of your vectorizer.
  + The screenshot of the results
  + You can name your report "Lab\_logfile\_vectorization\_yourname.doc".