Description of Australian Defence Force Academy-Linux Dataset (ADFA-LD) :

1) The dataset was generated on Linux local server running on Ubuntu 11.04, offering a variety of functions such as file sharing, database, remote access and web server.
2) Six types of attacks occur in ADFA-LD including two brute force password guessing attempts on the open ports enabled by FTP and SSH respectively, an unauthorised attempt to create a new user with root privileges through encoding a malicious payload into a normal executable, the uploads of Java and Linux executable Meterpreter payloads for the remote compromise of a target host, and the compromise and privilege escalation using C100 webshell. These types are termed as Hydra-FTP, Hydra-SSH, Adduser, Java-Meterpreter, Meterpreter and Webshell respectively. You can find these attacks inside the folder "Attack_Data_Master"
3) 833 and 4373 normal traces are generated for training and validation respectively, over a period during which no attacks occur against the host and legitimate application activities ranging from web browsing to document writing are operated as usual. These training and validation can be found in the "Training_Data_Master" and "Validation_Data_Master" folders, respectively.

Assignment Task:

1) Split the **A*ttack*** data of each category (Hydra-FTP, Hydra-SSH, Adduser, Java-Meterpreter, Meterpreter and Webshell) into 70% training data and 30 % test data. For instance there are are 10 folders in "Adduser" attack. Therefore, 7 of these folders are to be used for training and 3 folders are to be used for testing.
2) For the ***Normal*** data, files in "Training_Data_Master" folder are to be used as training data and files in "Validation_Data_Master" folder are to be used as test data.
3) Write a python script to find the frequency of occurences of all unique *3-grams*, *5-grams* and *7-grams* system call sequences in the training data for both **A*ttack*** data (across all categories of attack) and ***Normal*** data. For e.g., consider the following trace file corresponding to the Adduser attack.
   265 168 168 265 168 168 168 265 168 265 168 168 . . .
   Your script to list all 3-grams should produce the following output:
   265 168 168 -->3
   168 168 265 -->2
   168 265 168 -->3
   168 168 168 -->1
   265 168 265 -->1

   NOTE: To save time you can concatenate your entire training file for a particular class of attack and then run your script on the concatenated file instead of running it individually on each file.
4) Perform the same task on files in the "Training_Data_Master" to obtain all the unique *3-grams*, *5-grams* and *7-grams.*
5) Once you have obtained the frequencies of all the unique *n-grams* terms in the training data, use the top 30% *n-grams terms* with the highest frequency to create a data set. For instance consider following results for Adduser data (1st File):
   ('240', '102', '221') 7
   ('204', '203', '5') 2
   ('195', '199', '60') 1
   ('5', '197', '45') 1
   ('5', '195', '5') 12

('6', '220', '4') 1
('191', '5', '133') 9
('13', '45', '5') 2
('60', '5', '197') 4
('3', '142', '7') 2

Hydra-FTP data (2nd File):
('3', '142', '7') 11
('219', '311', '240') 4
('240', '13', '240') 1
('33', '168', '146') 2
('6', '168', '102') 3
('5', '197', '45') 1
('5', '195', '5') 2
('3', '91', '5') 8
('42', '120', '197') 1
('174', '54', '5') 2
('6', '63', '6') 18

Normal training data (3rd File):
('195', '10', '41') 1
('3', '142', '7') 3
('91', '240', '196') 2
('5', '195', '5') 2
('3', '102', '7') 17
('3', '195', '195') 14
('4', '78', '240') 1
('33', '195', '192') 2
('5', '197', '45') 15
('199', '45', '192') 1

The top 30 % 3-grams terms with highest frequencies in Adduser, Hydra-FTP and Normal data are [('5', '195', '5'), ('191', '5', '133'), ('240', '102', '221')], [('6', '63', '6'), ('3', '142', '7'), ('3', '91', '5')] and [('3', '102', '7'), ('5', '197', '45'), ('3', '195', '195')], respectively. Designate ('5', '195', '5') as feature 1(F1), ('191', '5', '133') as feature 2 (F2) ..... and ('3', '195', '195') as F9. Then, the generated dataset should have 9 features and one class label (Adduser, Hydra-FTP, Normal) with each feature corresponding to frequency of occurences of one of these 9 features. For instance for the 1st File, the generated data should be

Freq of F1, Freq of F2, ..., Freq of F9 ----->12, 9, 7, 0, 2, 0, 0,1,0, Adduser
Freq of F1, Freq of F2, ..., Freq of F9 ----->2, 0, 0, 0, 3, 0, 17,15,14, Normal

This will be the final training data which will be used to train various classifiers.

6) Apply the same procedure to generate the test dataset from the test files of the attack data (for all attack types) and the normal files in the "Validation_Data_Master" using the top 30% 3-grams terms with highest frequencies obtained during the training phase. The classifier model developed during the training phase will finally be validated on the Test dataset.

NOTE : You can refer the paper availabe at http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6743952 for further reference on ADFA-LD dataset.