Midterm #2

Dataset

You can download the dataset from here:

https://mega.nz/file/bU1HhCSC#sO7RSM2VkiVv9Rb-vTcWestE3oNBmHiFk0HxnY7FHWc

It contains files from… 3 dangerous Malware families!!!

But don't worry, the malicious files have been already disassembled and are now in innocuous .txt format

- Furthermore, to save your time, they have already been split in directories (one per each family)

- The files contain all the opcodes (instructions) that form the binary file

Name	Type
00a0437aa0555680f83bbb6072e0b79ea95bd25d.asm.txt	TXT File
00b84eda68d82edf2245a7eb5b656f8e888507fc.asm.txt	TXT File
00c84e2f48826dac8306e5d72ae049aa37bb78bb.asm.txt	TXT File
00cc2a1aaf7e5cb1ccbf4791a87964ac1e250eea.asm.txt	TXT File
00e71103640973661c8ed632beba4af89ff0a3de.asm.txt	TXT File
00ea896f9b7a9732c299e47ab320b8e486a41fa2.asm.txt	TXT File
00f3378ae795e2003ebfc542c6839b7436cc7e91.asm.txt	TXT File
0a4da66b67ee14db74aa982fb86d495ecb1ad229.asm.txt	TXT File
0a4dc90779f809c2066162079c33a1b1f54c9e7e.asm.txt	TXT File
0a5fb765fe69f84dd968a3fe40924c41b92ab079.asm.txt	TXT File
0a7bfb6633cedd55d91854ffb3dee1175c85fdbb.asm.txt	TXT File
0a8e52730a6b296884c4bf5a391c41d8f3b5f5b0.asm.txt	TXT File
0a8fc8c792c20c4855ed555207a408630d1bdbab.asm.txt	TXT File
0a9f3a469f882aac5a929333afaff7d8e87f4af5.asm.txt	TXT File
0a22e4623135b517f056150d0e44862935c3051c.asm.txt	TXT File
0a29e05f092debc7f4a94e74181c24e684b21f8e.asm.txt	TXT File
0a675b510d2a996f96409bccafc71d0a14b2f6cb.asm.txt	TXT File

```
push
mov
sub
push
push
push
mov
mov
mov
mov
mov
mov
mov
mov
cmp
jz
mov
mov
add
mov
push
lea
lea
call
mov
imul
mov
mov
imul
mov
add
```

Preprocessing

- Order the opcodes based on number of occurrence PER FAMILY and associate to each of them a specific symbol (ex: A, B, 1, 9, X, C, ...)

➢Note that the least common opcodes should all be converted to the same character/digit

How many unique symbols to use becomes a tuning parameter

- Per each file in the training set, you will convert the opcodes accordingly to the previous preprocessing steps

Algorithms

You can follow one of these two methods to feed the ML algorithm:

- [Method 1] The opcodes read from the files will be directly used to train the ML algorithms
- [Method 2] Instead of converting the opcodes directly to symbols, you will convert the n-grams of the opcodes

The main algorithm to use will be HMM

- You will need to train an HMM per each malware family based on the chosen Method

NOTE: You can apply Ensemble Learning to increase your detection rate

Experiments

The dataset contains three families:

Winwebsec, Zbot, and ZeroAccess

You will need to return the classification accuracy of these SIX tests:

Test 1 and 2:	Winwebsec vs Zbot
Test 3 and 4:	Winwebsec vs ZeroAccess
Test 5 and 6:	Zbot vs ZeroAccess

Report

You will need to submit a report split in these sections:

1. Pre-processing: Explain all the steps and eventual tools used to preprocess the dataset

2. Tuning: Describe all the tuning parameters used (and why) and collect them in a well-defined table

3. Experiments: Describe the output of your experiments in term of accuracy

4. Conclusions: Conclude the report reviewing all the steps, mentioning the best outcome, and proposing a Future Work paragraph where you hypothesize new possible experiments