C3i Hub, Indian Institute of Technology Kanpur

Problem (Malware Detection)

Description: Static and dynamic analysis of malware using machine learning. Train a model that takes static and dynamic analysis data, extracts features and classifies the input as Malware or Benign.

Dataset description:

Set 1: A directory containing files for static analysis– each file labeled by its hash value in a separate directory which contains 2 different text files – structure and strings. Even though malwares are further classified into various malware classes, you may put them together as just a single class – malware. The same analysis information has been made available for the benign files. Once you download the zipped-up files and extract the directories – you must programmatically extract features for each malware as well as for each benign file.

Set 2: A directory containing files for dynamic analysis– into JSON files each file labeled by its hash value. Even though malwares are further classified into various malware classes, you may put them together as just a single class – malware. The same analysis information has been made available for the benign files. Once you download the zipped-up files and extract the directories – you must programmatically extract features for each malware as well as for each benign file.

Set 3: A directory containing files for dynamic analysis– into JSON files each file labeled by its hash value. Even though malwares are further classified into various malware classes, you may put them together as just a single class – malware. The same analysis information has been made available for the benign files. Once you download the zipped-up files and extract the directories – you must programmatically extract features for each malware as well as for each benign file.

Steps to follow:

- **Data collection**: Collect Static and Dynamic Analysis Data for Malware and Benign samples provided.
- **Feature extraction:** Extract features from the collected dataset using a script.
- **Feature selection:** Select only important features so that prediction time will be reduced.
- **Classification:** Use machine learning classifiers to train the classifiers using extracted features.

Project must fulfill these requirements as mentioned below:

Project must have good accuracy, precision, recall, and F-score for both machine learning models (Static and Dynamic analysis) with low false positive and low false negative rate.

NOTE:

To train the model do not use all files as you will need to test the various figures of efficacy of your models. So, keep 25% of malware and 25% of benign file data for testing purposes.

Deliverable:

Create a program named MalwareDetection.py. The program should take as input the full path to a directory containing static and dynamic analysis information for 1000 or so files (mix of malware and benign ware). Then programs should extract feature vectors from these files – do the feature reduction – and run your model on the feature vector (follow all steps mentioned above) – for each file. At the end programs will output a .CSV file with two columns – one hash of the file you test, and in the second column Prediction result Malware/Benign. **All the source codes (feature extraction, selection, and machine learning model testing), trained model (for testing with random files), Observations during analysis in a document file, software required and readme file (how to use and libraries used) must be submitted in a single folder in zip format.**

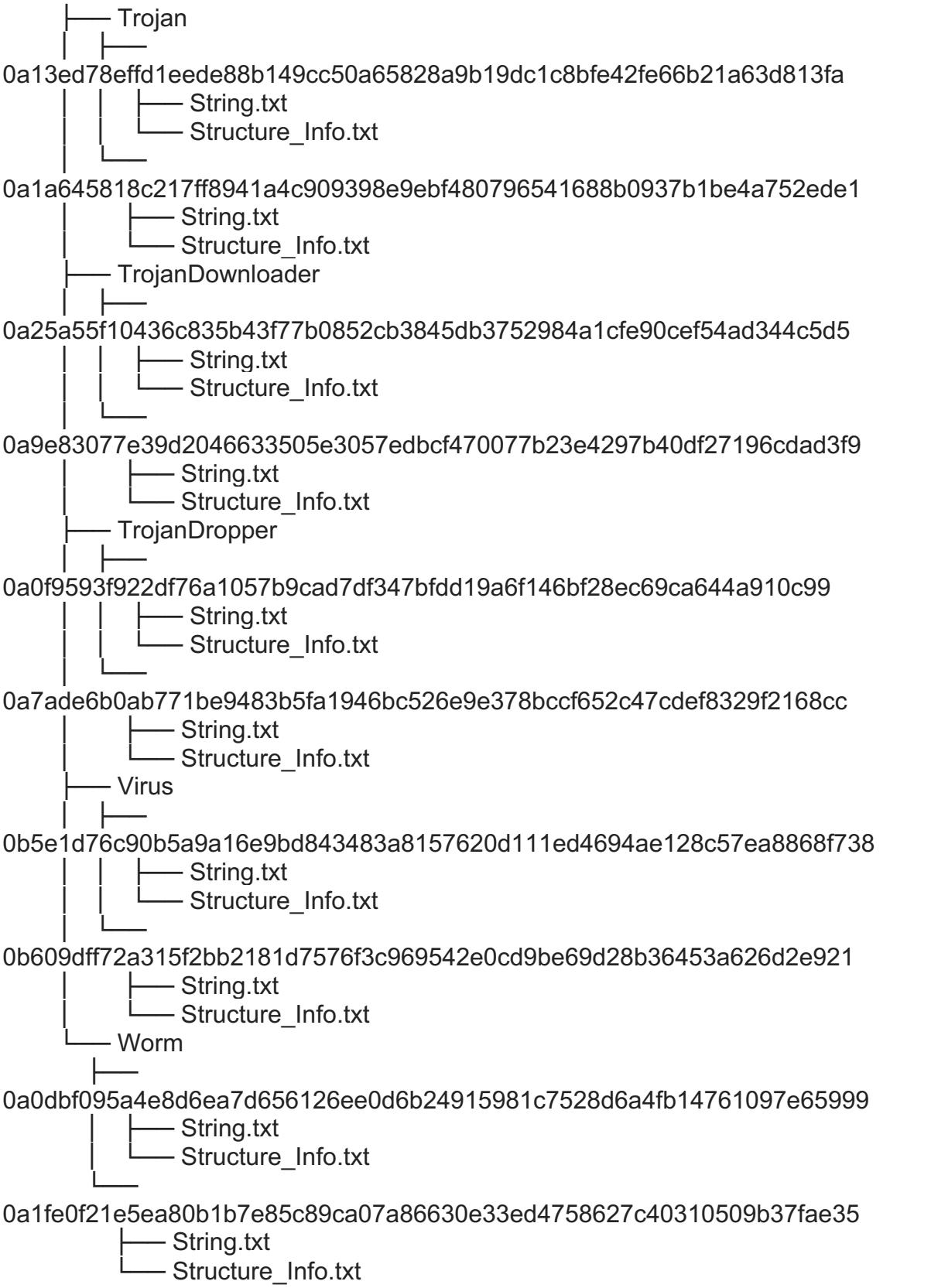
Details of Dataset

1: Static_Analysis_RAWDATA.7z: 1.3GB
Google Drive Link
https://drive.google.com/file/d/1XfnQMagW-yclH-wHZZRvYJHZSBExXzZu/view?usp=sharing
2: Dynamic_Analysis_Data_Part1.7z: 1.4GB
Google Drive Link
https://drive.google.com/file/d/13rmnrPsnoqjRBflDq6e59bW_YoaDeGxx/view?usp=sharing

3: Dynamic_Analysis_Dataset_Part2.7z: 1.5GB
Google Drive Link
https://drive.google.com/file/d/10P5R5WtK5NOV3-KF7yBGqLcidzMZJ8Uv/view?usp=sharing

Tree Structure Static Analysis Data folder for 2 files.

```
Static_Analysis_Data
    ├── Benign
    │   ├──
0a0ee0aa381260d43987e98dd1a6f4bab11164e876f21db6ddb1db7c319c5cf8
    │   │   ├── String.txt
    │   │   └── Structure_Info.txt
    │   └──
0a2adcac2b16b02d475e9d47b4772b77b0b4269132f07557c7ef6081727585da
    │       ├── String.txt
    │       └── Structure_Info.txt
    └── Malware
        ├── Backdoor
        │   ├──
0a21ef18ba03622736a8edd5390afbab6088dcacc3d5877eb0b28206285f569d
        │   │   ├── String.txt
        │   │   └── Structure_Info.txt
        │   └──
0a56a947d9c0be507b6aa0e2b569ca7eed39e5e802c8cf78be71adda9d324eae
        │       ├── String.txt
        │       └── Structure_Info.txt
```


Trojan
0a13ed78effd1eede88b149cc50a65828a9b19dc1c8bfe42fe66b21a63d813fa
String.txt
Structure_Info.txt
0a1a645818c217ff8941a4c909398e9ebf480796541688b0937b1be4a752ede1
String.txt
Structure_Info.txt
TrojanDownloader
0a25a55f10436c835b43f77b0852cb3845db3752984a1cfe90cef54ad344c5d5
String.txt
Structure_Info.txt
0a9e83077e39d2046633505e3057edbcf470077b23e4297b40df27196cdad3f9
String.txt
Structure_Info.txt
TrojanDropper
0a0f9593f922df76a1057b9cad7df347bfdd19a6f146bf28ec69ca644a910c99
String.txt
Structure_Info.txt
0a7ade6b0ab771be9483b5fa1946bc526e9e378bccf652c47cdef8329f2168cc
String.txt
Structure_Info.txt
Virus
0b5e1d76c90b5a9a16e9bd843483a8157620d111ed4694ae128c57ea8868f738
String.txt
Structure_Info.txt
0b609dff72a315f2bb2181d7576f3c969542e0cd9be69d28b36453a626d2e921
String.txt
Structure_Info.txt
Worm
0a0dbf095a4e8d6ea7d656126ee0d6b24915981c7528d6a4fb14761097e65999
String.txt
Structure_Info.txt
0a1fe0f21e5ea80b1b7e85c89ca07a86630e33ed4758627c40310509b37fae35
String.txt
Structure_Info.txt

==========

Dynamic_Analysis_Data is divided into two parts, Part1 and Part2. Candidates can download one part and do training and they can download the other part to retrain.

Dynamic_Analysis_Data_Part1: Zip file size is **1.4 GB** and after unzip **23 GB** (Downloading size is equal to ZIP size)

1.4 GB: Dynamic_Analysis_Data_Part1.7z
After extraction of the Dynamic analysis data Part1 folder size will be 23GB.
Total Storage needed for analysis is 23GB.

Dynamic_Analysis_Data_Part1.7z:
Google Drive Link

https://drive.google.com/file/d/13rmnrPsnoqjRBflDq6e59bW_YoaDeGxx/view?usp=sharing

Tree Structure of Dynamic Analysis Data Part1 Folder for 2 sample files.

```
Dynamic_Analysis_Data_Part1
├── Benign
│   ├── 0a0ee0aa381260d43987e98dd1a6f4bab11164e876f21db6ddb1db7c319c5cf8.json
│   └── 0a2adcac2b16b02d475e9d47b4772b77b0b4269132f07557c7ef6081727585da.json
└── Malware
    ├── Backdoor
    │   ├── 0a21ef18ba03622736a8edd5390afbab6088dcacc3d5877eb0b28206285f569d.json
    │   └── 0a56a947d9c0be507b6aa0e2b569ca7eed39e5e802c8cf78be71adda9d324eae.json
    ├── Trojan
    │   ├── 0a13ed78effd1eede88b149cc50a65828a9b19dc1c8bfe42fe66b21a63d813fa.json
    │   └── 0a1a645818c217ff8941a4c909398e9ebf480796541688b0937b1be4a752ede1.json
```

…. Continued on next page

Dynamic_Analysis_Data_Part2: Zip file size is **1.5 GB** and after unzip **21.4 GB** (Downloading size is equal to ZIP size)

1.5GB: Dynamic_Analysis_Dataset_Part2.7z
After extraction of the Dynamic analysis data Part2 folder size will be 21.4 GB.
Total Storage needed for analysis is 21.4GB.

Dynamic_Analysis_Dataset_Part2.7z:
Google Drive Link

https://drive.google.com/file/d/10P5R5WtK5NOV3-KF7yBGqLcidzMZJ8Uv/view?usp=sharing

Tree Structure of Dynamic Analysis Data Part2 Folder for 2 sample files.

```
Dynamic_Analysis_Data_Part2
├── Benign
│   ├── 0a0ee0aa381260d43987e98dd1a6f4bab11164e876f21db6ddb1db7c319c5cf8.json
│   └── 0a2adcac2b16b02d475e9d47b4772b77b0b4269132f07557c7ef6081727585da.json
└── Malware
    ├── TrojanDownloader
    │   ├── 0a25a55f10436c835b43f77b0852cb3845db3752984a1cfe90cef54ad344c5d5.json
    │   └── 0a9e83077e39d2046633505e3057edbcf470077b23e4297b40df27196cdad3f9.json
    ├── TrojanDropper
    │   ├── 0a0f9593f922df76a1057b9cad7df347bfdd19a6f146bf28ec69ca644a910c99.json
    │   └── 0a7ade6b0ab771be9483b5fa1946bc526e9e378bccf652c47cdef8329f2168cc.json
    ├── Virus
    │   ├── 0b5e1d76c90b5a9a16e9bd843483a8157620d111ed4694ae128c57ea8868f738.json
    │   └── 0b609dff72a315f2bb2181d7576f3c969542e0cd9be69d28b36453a626d2e921.json
    └── Worm
        ├── 0a0dbf095a4e8d6ea7d656126ee0d6b24915981c7528d6a4fb14761097e65999.json
        └── 0a1fe0f21e5ea80b1b7e85c89ca07a86630e33ed4758627c40310509b37fae35.json
```