DSP Project 1

Malware Classification

Team Dragora

Anuj Panchmia, Sumer Singh, Vishakha Atole

Feature Extraction

- Single Byte Counts (from .byte)
- Header Counts (from .asm)
- Opcode Counts (from .asm)
- .Dll Counts (from .asm)

Parallel Approach

1. The raw .asm and .byte content for all files are loaded into a spark dataframe.
2. The .byte and .asm content is processed to only contain bytes and headers, respectively.
3. CountVectorizer.
4. Vector Assembler is used to create a feature vector.
5. Random Forest is used to classify.

Sequential Approach

1. Load .asm or .byte file sequentially into memory.
2. Process and take the counts simultaneously.
3. Random Forest is used to classify.

This additional counting step is done in **constant time** (only requires a dictionary lookup).

Random Forest

- Best score achieved with Header Count & Byte Count features.
- Hyperparameters for best score:
  - Number of Trees – 40
  - Maximum Depth - 23
- Accuracy of 99.0077%

Issues

- Gcloud shell disconnections.
- Overuse of credits.
- Out of memory errors.