Malware Classification

Team Sabayon
(Marcus Hill, Saed Rezayi, Jayant Parashar)

Overview

The Good

- Our team was able to achieve 95% accuracy on the full dataset.
- We are able to examine many features and model structures combinations
  - Naive Bayes
  - Logistic Regression
  - Random Forest
  - SVM

The Bad

- At least a week of our initial development time was spent debugging.
  - Imperative to understand the library implementation, and not make assumptions.
- Time was lost many times trying to find the correct memory configuration for a cluster
  - Memory issues would occur hours into training. and require a restart

Reflections

Lessons Learned

- Simple models can outperform more sophisticated techniques. (Occam's Razor)
- Even the most modern advancements in feature representations, may not work the best with the data.
    - Time must be spent evaluating the actual dataset to discern which features extraction methods can be the most effective
- Understanding the theory behind the models can allow for effective parameter tuning

Results

Model Accuracies

Classifier	Model Pipeline	Dataset	Accuracy
Naive Bayes	Tokenize, Trigram, Stopwords, HashingTF	Small	72%
Logistic Regression	Tokenize, Trigram, Stopwords, HashingTF	Small	84%
Random Forest	Tokenize, Stopwords, Bigram, HashingTF, Max_Depth = 7	Small	92%
Random Forest	Tokenize, Stopwords, Bigram, HashingTF, Max_Depth = 7	Large	95%
Random Forest (asm)	Tokenize, Stopwords, Bigram, HashingTF, Max_Depth = 5	Small	94%