DATA SCIENCE PRACTICUM - PROJECT 1

MALWARE CLASSIFICATION

By Team Zentyal

APPROACHES CONSIDERED

- Linear Regression
- Bytes N-gram with Random Forest
- Opcode N-gram with Random Forest
- Image visualization of Malware

LINEAR REGRESSION

Features considered:

- Byte Entropy
- .dll calls
- Mem reads & writes
- Byte histogram

Analysis : Two much time needed to extract features to make sense.

Just for fun : Ran a model with just hex byte count. Accuracy : 55.9

BYTE N-GRAM WITH RANDOM FOREST

- Calculated the byte count for each Hex byte.
- Using CountVectorizer to tokenize the bag of words and extract features.
- Passed the features through Random Forest with number of trees 50 and depth 25

Accuracy : 98.2

OPCODE N-GRAM WITH RANDOM FOREST

- Extract opcodes from the ASM file
- Calculated the byte count for each Hex byte.
- Using CountVectorizer to tokenize the bag of words and extract features.
- Passed the features through Random Forest with number of trees 50 and depth 25

Accuracy 98.5

LEARNINGS\GAINS

- GCP sdk, ssh jupyter notebook web interface setup
- Random Forests
- Pyspark experience
- ISSUE : wholeTextfiles doesn't take files in an ordered manner