Arthur Liou
CS373

Lab 2 Writeup

Lab 2 - The Great Host/Lexical URL Reputation Bake-off

Deliverable (The fun part) - Please provide by due date:

1. Create a 1-page document which summarizes your strategy. What features are you using, what are their weights? How did you figure out a threshold?

   Strategy: As suggested, I used point system where malicious features add to a score, and safe features detract from a that score. I figured out a threshold to use to decide if a URL has exceeded a score which makes it malicious.

   The features/weights that I'm using include:

   - EXE file = +5
   - Young Domain = +2 (all the URLs <= 700 days old or so were malicious)
   - If within Alexa's top 1MM Rank, then reduce score = -5 (URL Prevalence)
   - If no IP Address associated = +5
   - Fragmented URLs = + 2 (lower since there were not as many)
   - Non-traditional ports used (80, 443) = +4

   After looking through the data, I determined weights and determined a threshold of 10 for determining a malicious URL. My reasoning here was that one "big" strike like just having an exe or no IP address associated could be an anomaly, but having two "big" strikes, like exe file and no IP? Now that would be a bigger red flag. Having the Alexa top 500 rank of +5 would help in balancing out the mostly negative attributing factors

2. Implement your strategy by extending the readcorpus script and turning it into your URL microclassifier. Run it against the classification set (the one without malicious_url flag populated) and have it write a results file of in the format : <url string>, <malicious bit> (where 1 =malicious, and 0=safe)

   Done – see code below.

3. Create a final document that combines your code, your results, your thought processes, etc. In other words, write everything up!

Results: Final % of expected malicious URLs (from my script) was 52%.
Thought Processes: I found this lab very informative and fun, especially since in building up our Malicious URL Detector. I also appreciated the more ML/AI side of creating this script.

My Code’s Addition to the given script:

```
# Exe?
x = record["url"].find(".exe")
if x != -1:
    record["my_score"] += 5

# Age of Domain
age = record["domain_age_days"]
if age < 700:
    record["my_score"] += 2

# Alexa Rank. Subtract 5 if in top 1MM
alexa_rank = record["alexa_rank"]
if int(alexa_rank) < 1000000:
    record["my_score"] -= 5

# Is IP Present
ips = record["ips"]
if ips is None or len(ips) == 0:
    record["my_score"] += 5

# Fragmented URL
if record["fragment"] is not None:
    record["my_score"] += 2

# Normal Port?
if record["port"] not in [80,443]:
    record["my_score"] += 4

# Malicious URL Determination
if record["my_score"] >= 10:
    baddies += 1
    guesses.append((record["url"], 1))
else:
    guesses.append((record["url"], 0))
```