Sarthak Dhomne
CS 131

CS 131 Mini Project: Phishing Websites

1. Project Objectives
    - Identify common characteristics of phishing sites
    - Extract correlations between attributes
    - Visualize findings to showcase phishing risk factors

2. Dataset Metadata
    - Number of entries: 11055
    - Number of attributes: 31 (last one is result of if phishing site or not)
    - Value ranges: Vary, many are 1: normal, or -1: doubtful/phishing; some on a scale of -1,0,1 lke URL length

3. Summary

The dataset analysis revealed several notable characteristics that are commonly featured for phishing sites. This was done by filtering the data to generate a count of each of the most common negative attributes.

The most common one was the addition of a suffix or prefix to a URL that is separated by a dash. For example, google-info.com may be used. This was a feature in 1341/1550, or about 86.5% of the sites that were phishing sites within the dataset.

Another common one was the SFH, or server form handler, being blank or redirecting to a different domain. This is commonly done to steal user form data. This was the case in 77.8% of the sites. Finally, a common characteristic was low Page Rank, a value used to measure the importance of a site on the internet and on search engines. Around 77% of phishing sites had a PageRank below 0.2.

Overall 14% of the sites in the dataset were phishing sites.

4. Tools/Commands Used
- wc -l and Awk script to count attributes
- Grep to get overall line count
- Sed cmd to remove in headers and clean data
- Awk script to count verified phishing sites in the dataset

Sarthak Dhomne
CS 131

5. 2 plots