DATASETS:

trancolist (Legitimate URLs) and phishtank (Phishing URLs) datasets were used.

LIBRARIES INSTALLED (only till review 2):

beautifulsoup4

matplotlib

pandas

requests

urllib3

CONCEPTS (only till review 2) – WEBSCRAPING, FEATURE ENGINEERING, DATASET CREATION

PROGRESS TILL REVIEW-2:

1. Functions for content-based features were defined for extracting data.

Content-based features – Binary Features and Quantitative Features

|  |  |
| --- | --- |
| BINARY | QUANTITATIVE |
| has\_title | number\_of\_inputs |
| has\_input | number\_of\_buttons |
| has\_button | number\_of\_images |
| has\_image | number\_of\_option |
| has\_submit | number\_of\_list |
| has\_link | number\_of\_TH |
| has\_password | number\_of\_TR |
| has\_email\_input | number\_of\_href |
| has\_hidden\_element | number\_of\_paragraph |
| has\_audio | number\_of\_script |
| has\_video | length\_of\_title |

1. Using BeautifulSoup module, soup object was created and a vector was created by calling all the functions for the soup object. Feature Extraction was defined and the data was saved in a directory.
2. The .csv datasets were converted to dataframe using pandas library. URLs were retrieved from both datasets and the above-mentioned features were extracted. We extracted data from only 1000 URLs from both datasets for now and created structured data.
3. The new dataframes with the values of these features(structured data), were created for both Legitimate and Phishing datasets.
4. These dataframes were converted to .csv files our new datasets upon which we will work using various Machine Learning algorithms.

RESULTS:

Successfully created datasets for Legitimate and Phishing websites by webscraping, feature engineering, dataframes, datasets concepts.