Classifying Malicious URLs

By Nathan Dullea

What is a malicious URL?

From pandasecurity.com:

“(A) *Malicious URL* is a URL created with malicious purposes, among them, to download any type of malware to the affected computer, which can be contained in spam or phishing messages, or even improve its position in search engines using Blackhat SEO techniques.”

https://www.pandasecurity.com/homeusers/security-info/213894/Malicious

The environment.

- Anaconda for package management.
- Scikit Learn (sklearn) for Machine Learning Framework.
- Sublime Text for editing files.
- Github for version control.
  - https://github.com/ndullea/malicious_URL_Classifier
- Dataset downloaded from Kaggle:
  - https://www.kaggle.com/antonyj453/urldataset/data
  - Dataset is in the form: (url, label)

Main steps in malicious url classification:

1. Read CSV to Dataframe
2. Separate URLs and Labels
3. Transform URLs using tfidfvectorizer
    a. Makes a Count vectorizer with the tokens in url and returns a sparse matrix
    b. Transform sparse matrix by term frequency * inverse document frequency
4. Separate transformed urls and labels into testing and training set
5. Fit a Logistic Regression model on the test set
6. Run model on test set to get accuracy

Quick Code Run Through

Improvements and Thank You.

Improvements:

- Test other models
- Improve feature list
  - Length of url
  - Quality of tokens: are they actual words? Websites?
- Using pickle to export model for external use

Thank you!