Project Proposal

Spam URL Classification

Mahesh B (1NT19CS105), Nikhil K Rajput (1NT19CS127),

Niraj S Joshi (1NT19CS129), Womika Khan (1NT19CS218)

Data Mining project. We will be classifying URL's as a spam URL or not.

Data Mining Task

The data mining task used in this application is classification. The above dataset can be used to create a binary classification model.

Data Set

This dataset contains about 87.5K URLs in which one-third are flagged as a spam URL and restrict are not spam. The F flagging system identifies if a link is a spam or not, as it parses links from over 100 newsletters every 30 minutes. A link is programmatically F flagged if it appears 3+ times in a single newsletter or contains a likely subscribe/unsubscribe URL.

Methods and Models

Firstly, we read the csv file, and train the model to classify if the URL is spam or not based on the length, number of words, count of digits. We implement this idea using Python Language.
Finally we find the accuracy score of the model using ROC-AUC curve which is a graph which shows the performance of a classification model.

Assessment

We find the accuracy score of the model using ROC-AUC curve which is a graph which to validate that you have found meaningful patterns in the classification model at all classification thresholds.

Presentation and Visualization

The graphs plotted based on attributes such as number of letters, count of digits can be used to analyse if the URL is a spam or not.

Roles

- Data gathering – Womika Khan
- Training the model – Niraj S Joshi
- Graph plotting – Mahesh B
- Accuracy testing – Nikhil K Rajput

Schedule

Date	Tasks to be Completed
03/01/21	Project Proposal
17/01/22	Project Report Submission
20/01/22	Tasks completed by the class presentation

Bibliography

https://www.kaggle.com/shivamb/spam-url-classification-getting-started/data

https://www.kaggle.com/sasakitetsuya/spam-url-classification-by-extra-trees-classifier

https://www.kaggle.com/kkhandekar/spam-url-feature-extraction

https://www.kaggle.com/shivamb/spam-url-classification-getting-started

https://www.real-statistics.com/logistic-regression/classification-table/

https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5