**Classification of malicious URL**

END SEM PROJECT REPORT

***Submitted by***

BL.EN.U4AIE20006 ------ B Srivathsan

BL.EN.U4AIE20058 ------ Samirit Saha

BL.EN.U4AIE20025 ------ Kanisettypalli Karthik

**19AIE214 – Big Data Analytics**

**B.Tech. in Computer Science and Engineering (Artificial Intelligence)**



AMRITA SCHOOL OF ENGINEERING, BANGALORE

AMRITA VISHWA VIDYAPEETHAM

BANGALORE 560 035

June – 2022

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.no** | **Topic** | **Page.no** |
| **1.** | **Problem Statement** | **3** |
| **2.** | **Dataset Description** | **4** |
| **3.** | **System Workflow** | **4-5** |
| **4.** | **Execution** | **5** |
| **4.** | **Conclusion** | **10** |
| **5.** | **References** | **11** |

**Problem Statement**

With the rise of dominance in the usage of the internet in today’s world, there is also an increase in the opportunity of exploitation of information-related resources by hackers seeking to misuse the World Wide Web for their personal gain. In our project, a malicious URL detection method is implemented based on machine learning methodologies, wherein the nature of a URL being malicious or not is assessed based on its ‘attributes and behavior.

The technology of Bigdata is also implemented for the improving the model’s capability to detect malicious URL based on its’ behavior and attributes. Therefore, the prediction model consists of a set of features and behaviors for the URL that are new, a algorithm that implements machine learning, as well as technology that operates on bigdata. The results from the model show that the behavior and attributes of a URL can improve the ability of the model to detect malicious URLs. This is suggested that the proposed system may be considered as an optimizable solution for malicious URL detection.

In our project, a malicious website being malicious or not is assessed based on its ‘attributes and behavior.  The technology of Bigdata is also implemented for the improving the model’s capability to detect malicious URL based on its’ behavior and attributes. Therefore, the prediction model consists of a set of features and behaviors for the URL that are new, a algorithm that implements machine learning, as well as technology that operates on bigdata. We use features such as url JavaScript content obfuscated JavaScript code and converting them to one feature to classify it. Py spark will be used for this project.

**Dataset Description**

We obtained the dataset which was used for malicious website classification using lexical tokens content tokens of the given dataset.

URL gives the url domain of the website.

TLD gives the domain name such as .org .com etc.

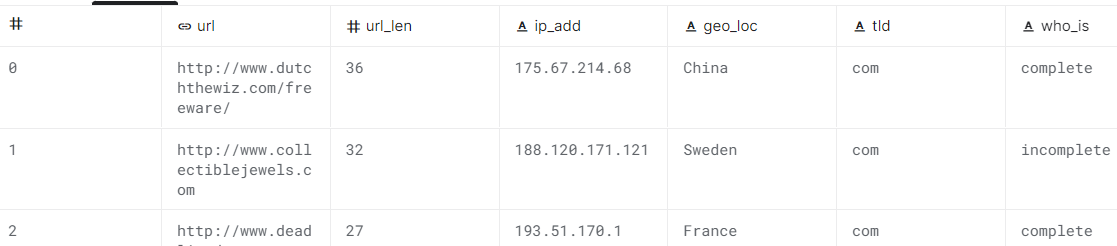
Who is tells us whether the domain information Is complete or anonymous

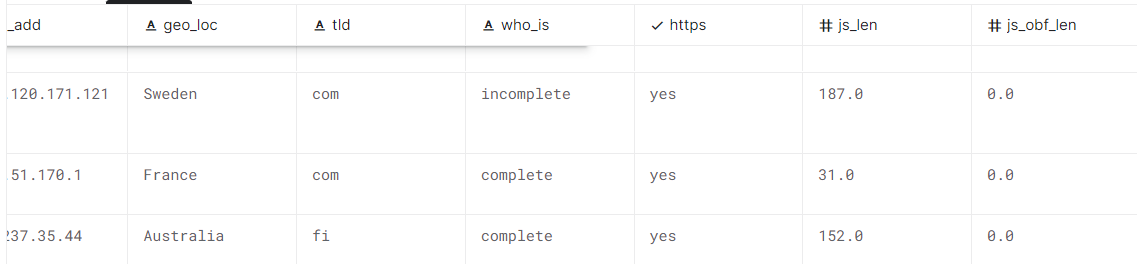
Javascipt length and javascript obf length display the code length .

Content is the actual content of the website body to be displayed.

Label is the classification label to be used.







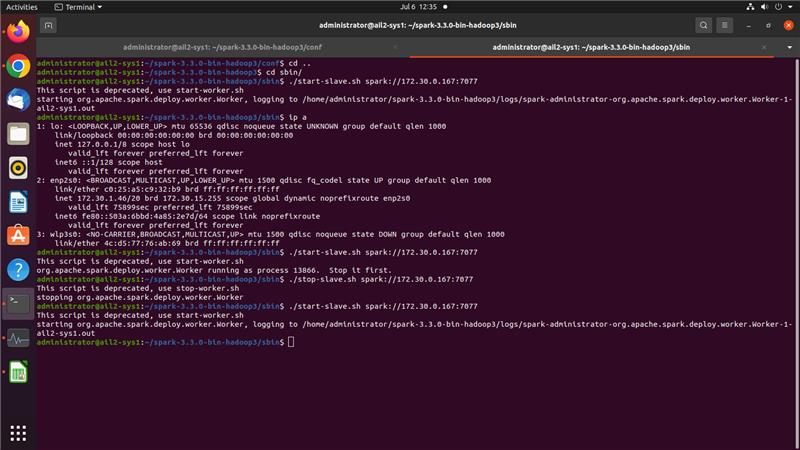
No of features are 12

No of rows are 359217

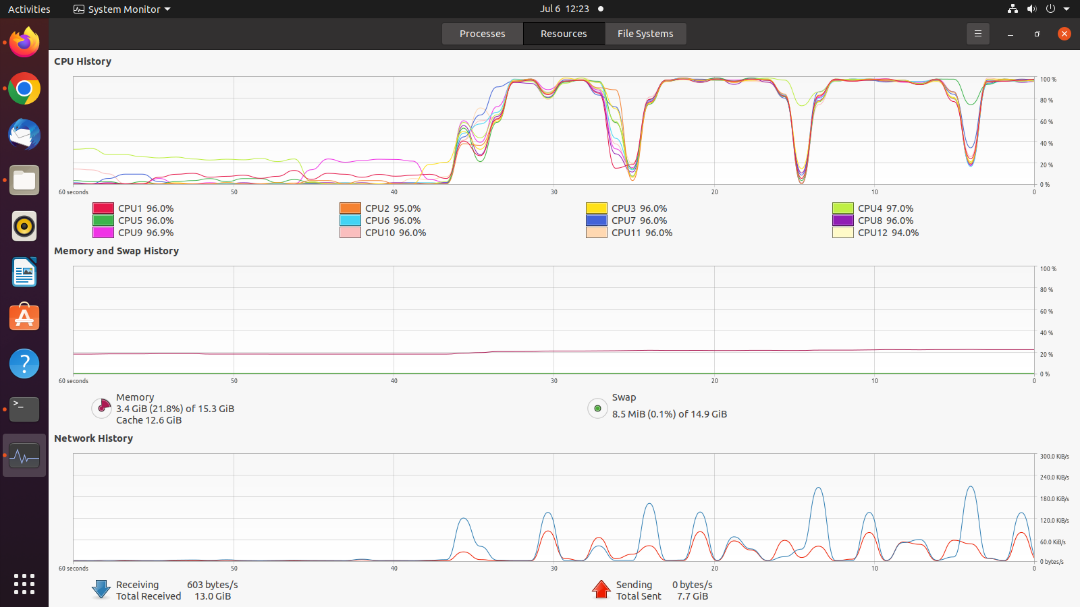
**System Workflow**

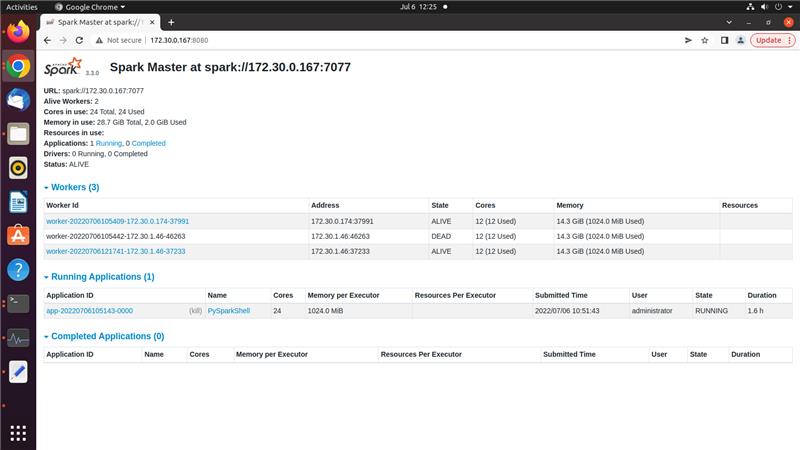
1. We begin by loading the 5.1 Gb dataset into systems spark bin folder .
2. The cluster configured by the ip address of master is started as slave and master based on the system to enable scheduling and processing of data across the cluster.
3. We run the pyspark shell and import the feature extraction and machine learning algorithms.
4. First step is to load the csv file into a data frame in csv format and check the schema.
5. We then perform typecasting on integer values mapped as string in csv back to integer to be used in vector assembler.
6. We clean the data by removing rows with null values and removing label which might have been altered by mistake of delimiters due to content being a large string.
7. We remove the extra features from the obtained dataset to only have original features a feature vector of all features combined and the labels.
8. We then split the train and test data as 70 30 percent .
9. We input the labels and training features into the NaiveBayes and LinearSVC and use classification evaluators and give input as predicted and actual labels to predict accuracy f1 score etc.
10. We write the output from classifier into a folder on the bin folder of spark.

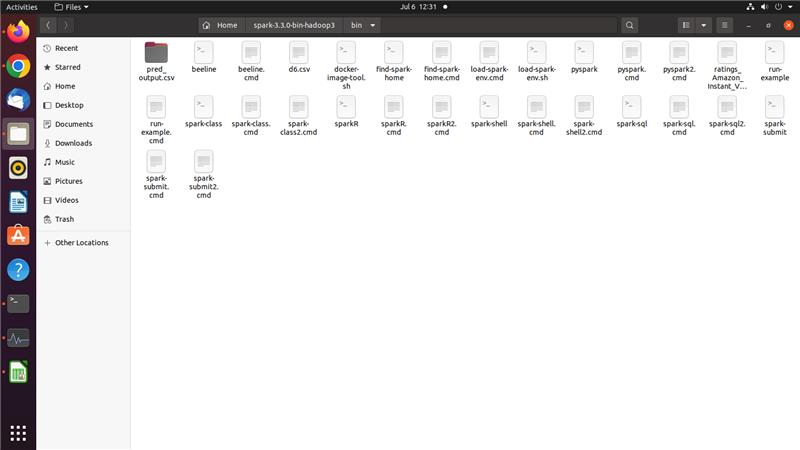
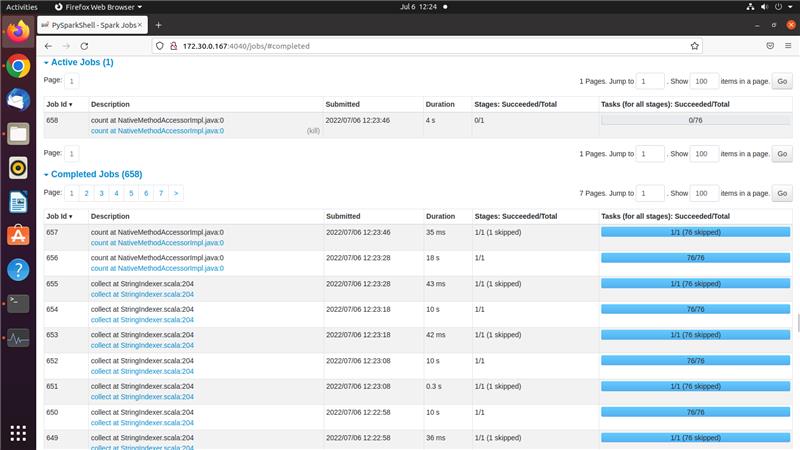
**Execution**

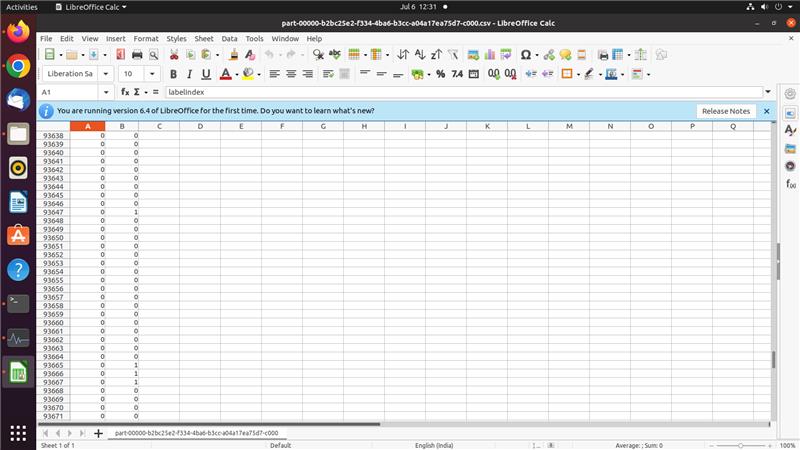
****

****

****

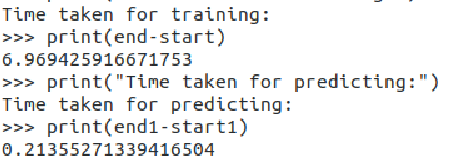
****

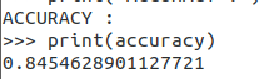
****

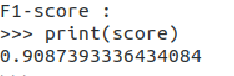
****

**Conclusion**

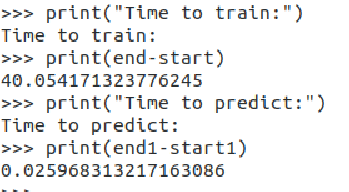
**Naïve Bayes**

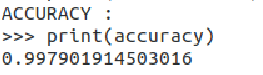


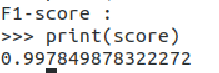




**LinearSVC**







**Conclusion:**

We have implemented feature extraction and machine learning using naïve bayes and LinearSVC on a csv dataset on websites and classified them into benign and malicious based on their attributes. We have tried to optimize the results by preprocessing of data and we have obtained an accuracy of 84 and 99 percent for Naïve Bayes and LinearSVC respectively.

**References:**

[1] Do Xuan, Cho, Hoa Dinh Nguyen, and Victor Nikolaevich Tisenko. "Malicious URL detection based on machine learning." *International Journal of Advanced Computer Science and Applications* 11.1 (2020).

[2] Vemuri, Pujitha. "Detecting malicious shortened URLs using machine learning." (2018).

[3] Cui, Baojiang, et al. "Malicious URL detection with feature extraction based on machine learning." *International Journal of High Performance Computing and Networking* 12.2 (2018): 166-178.

[4] Patgiri, Ripon, et al. "Empirical study on malicious URL detection using machine learning." *International Conference on Distributed Computing and Internet Technology*. Springer, Cham, 2019.

[5] Joshi, Apoorva, et al. "Using lexical features for malicious URL detection--a machine learning approach." *arXiv preprint arXiv:1910.06277* (2019).