CMP-681: Information Retrieval Term Project Proposal: URL Crawler Toolkit

1. Project Topic

As is known, the data plays a crucial role in the success of supervised based machine learning approaches since they fundamentally attempt to fit one or more learnable decision function(s) derived from the input features and their corresponding labels. However, according to our best knowledge, in the domain of URL based phishing detection there is not a publicly available benchmark dataset that is being used by the suggested methods in the literature by means of evaluation and comparison. As a result, in this term project we aimed to develop an URL Crawler Toolkit as a Software Track project that would collect URL's and provides a reliable legitimate website dataset for future researches.

There are several link crawler implementation on the Internet yet most of them are targeted on crawling whole content from given URL. Even a few of them are mainly purposed to extract links from given website, yet they are extracting these links from single URL for each iteration. Therefore, none of the open sourced project could be used in our implementation since we have intention to make a software that could work in parallel in order to reduce required among of time to extract high numbers of links to setup a dataset. Since we are going to implement this Software in our own, inevitably, we would not be able to compare or test our implementation directly with another project. But, we would provide performance criteria and results by using some Diagnostic Tools.

2. Project Type

The selected project type is Software Track for this term project.

3. Project Members

This project is mainly and only focused personal by Fırat Coşkun Dalgıç. However, if there are some students who has not found a topic or group members, they are welcome to join this contribution and they are going to be coordinated by myself. The contact information of members and their roles are given below:

- Coordinator: Fırat Coşkun Dalgıç, firatcoskundalgic@gmail.com, +905059523084

4. Purposed Software Implementation

Link Crawler Toolkit software algorithm is given in Figure 1 and would be consist of two main parts which are Information (URL) Crawling and Information Retrieval mechanism respectively.

In the first part, we are going to extract both external and internal hyperlinks from each given URL by crawling DOM model of that website. Afterwards, each extracted candidate links are going to be filtered by using pre-defined search criteria in which these criteria are assumed to be Top Level Domain, Number of Domain in Subdomains, Path Level but it could be extended during project development. Once, all parallel batch processing is completed, we would have remaining URLs named Search Results and are going to move the second part.

In the second phase, firstly we are going to retrieve full information from candidate URLs such as domain name, WhoIS Information, subdomains, TLD, query parameters if exists. Then, we are going to filter results by applying Storage Filters where those filters could be the Sample per Domain (the maximum number of samples could be stored in database for each domain), Sample per TLD (the maximum number of samples could be stored in database for each Top Level Domain) and those could be expanded during the development.