Internship Report for February - March 2018

Yash Saboo

2015-2019 Batch

T.Y.C.S.

15070121170

**MORE DETAILS THE PROJECT**

**Problem Statement:**

To develop a Google Chrome Web Extension which can detect phishing/imitation websites and alert users about the same based on previously browsed websites.

**Technology Used and Technical Details:**

**As mentioned in the previous report, “**I was assigned to explore, exclusively, the algorithms and methods that will be used by the anticipated product and work on proof-of-concept.” To achieve the same, following technologies were used:

1. PHP for algorithm implementation.
2. Composer, MySQL and Apache Server in the integration with wamp64 for PHP execution.
3. Git to fork, clone and push all the project on Symantec Bitbucket System.
4. Python to download the dataset from an online repository.

These technologies were learnt and implemented, as when required, using various resources available on Internet. I enrolled in few online courses and tutorials on Coursera and CodeAcademy to understand it practically.

**TASKS DONE**

**Task 1: Algorithm Exploration and POC Development**

The algorithm exploration and POC development went hand-in-hand by following the following flow:

* **Choosing Algorithm and Technique**
  + The overall details of the algorithms used are bound by a confidentiality clause but it basically involved image hashing to find the similarity between two images.
  + Scholastic and technical writings - Blogs, research papers, previous projects and Symantec Wiki - were referred to absorb the algorithms.
  + Investigated the ways in which the algorithms can be tweaked with respect to the project application.
* **Data Selection**
  + The dataset was downloaded from an online repository, PhishTank, containing a massive collection of known phishing websites. It has information regarding the website and screenshots of the phishing sites.
  + All these were paired and arranged neatly with a streamlined ID system, which would make it quite simple to automate access to the screenshots.
  + A Python script was written to automate mass download the information of identified phishing websites from phishtank.com and, also, group them in folders based on the website that is being imitated.
* **Data Preprocessing**
  + Running the Python script resulted in the creation of a repository, which was later manually filtered to get only viable screenshots (quite a few of the screenshots were either blanks or incorrectly listed).
  + This manual filtering also helped in understanding the data that is being dealt with and, thus, in later stages, helped in understanding where the algorithm is likely to fail in some cases.
* **Development**
  + The three algorithms were implemented rudimentarily in PHP by an open source developer. The project was available on GitHub to fork and clone using Git.
  + The project was forked and the PHP scripts of algorithms were amended to make it work more efficiently with respect to the phishing concept that project was supposed to contrivance.
  + Composer, MySQL and Apache Server, integrated in wamp64, were used to execute the PHP scripts.
  + The apache server’s scripts had to be changed to allow proper functioning of the PHP scripts.
  + The algorithms went through multiple revisions.
* **Testing** (March)
* **Benchmarking** (March)

Currently, I am amidst my Development stage.

Additionally, all these files are uploaded regularly on the Bitbucket of Symantec using Git.

**Task 2: Learning HTML, CSS and JavaScript**

As my project is supposed to be deployed as a chrome extension, I must learn Chrome Extension Development and the languages involved in it. I accomplished the latter part by enrolling and successfully completing the course, with 100% Grade, on Coursera called, “Programming Foundations with JavaScript, HTML and CSS”. It was selected among other tutorials because the course materials’ projects and assignments included functions and events based on Image Modifications.

Certificate Link: <https://www.coursera.org/account/accomplishments/verify/G8QX4WKUVLAS>

**FUTURE WORK**

At the end of February, I was assigned to test the algorithm and document the benchmarking. If successful in first testing, I’ll move on to benchmarking directly or, else, I will develop and revise the algorithms again. Once the benchmarking has been documented, I must present a presentation on the same to the team. Simultaneously, I’ll be also learning Chrome Extension Development exclusively. They will be described in the next report (February - March).

Thank You.