SEG3904 Project Proposal

*Project Title:*

*Network Traffic Classification for Detection of malicious Botnets*

Student:

Supervisor: Miguel Garzon – mgarzon@uottawa.ca

**Overview**

The purpose of this project is to analyze multiple classification models for detecting of malicious botnets. Before the project is started there will be a literature review on the subject. This will help in gathering as much information on the subject before starting. At the start of this project, proper data mining techniques will be utilized to collect the appropriate data. Then, the collected data will be prepared, preprocessed, and discretized for ease of classification. Finally, different classification models will be used, and results will be analyzed. At the end of this research project, a report will be produced that discusses in detail the different classification models used and which of these performed more accurately in classifying malicious botnets over data communication networks.

Learning Outcomes:

At the end of this project, the student will have learned to:

● Understand and implement Data Mining techniques.

● Learn how to use Spark and Hadoop file distributed system to store, retrieve, edit, and analyze data.

● Learning how to prepare and preprocess data for proper analysis

● Learn how to implement and analyze different classification models.

● Learn how to analyze and interpret data packets in general and more specifically for cybersecurity threats.

Technologies:

* Python is a general use high level programming language.
* Pyspark is a Python API that gives access to SPARK.
* Scikit-learn is a machine learning library for Python.
* Docker is a containerization software to perform operating system level virtualization

Resources:

● Various literature works that will be discovered during the literature review.

● Technology websites mentioned above.

* William Stallings, Data and Computer Communications, 9th edition, Prentice-Hall, 2011.

**Deliverables**

|  |  |
| --- | --- |
| **Deliverable** | **Weight** |
| Literature review report | 15% |
| Environment setup | 10% |
| Implementation code: Part a) Data Collection Module | 10% |
| Implementation code: Part b) Data Cleansing, Preprocessing, and Discretization | 20% |
| Implementation code: Part c) Classification of data | 20% |
| Implementation code: Part c) Complete code | 5% |
| Final Report with final results, analysis of results, challenges, code structure overview, self­assessment of learning, and possible future work items. | 20% |

**Work Plan (135 hours)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Meet?** | **Action** | **Hours** |
| 1 | Y | Project plan and expectations | 10 |
| 2/3/4 | Y | Performing Literature review | 35 |

|  |  |  |  |
| --- | --- | --- | --- |
|  5 | Y | Create Report of literature review |  15 |
|  6/7/8 | Y | Classifying network Traffic |  35 |
| 9/10/11 | Y | Compiling data and Creating report |  40 |