|  |  |
| --- | --- |
| **Title** | **Risk perception in Cyber security environments**  **using NLP** |
| **Student name:** | **Gerard Tio Nogueras** |
| **Supervisor name:** | **Dr Frey Sylvain** |
|  | |
| ***Aims/research question and Objectives*** | |
| **Project aims**  The main aim of the project is to find ways to properly perceive risk in cyber security utilising Natural Language Processing algorithms.  **Project objectives**  The objectives of the project:   * Learn the NLP pipeline and familiarize with NLTK and other libraries. * Carry out a pre-processing on the scripts available to prepare them for the classifiers. * Start an iterative process to find the appropriate classifier:   + Choose a classifier   + Carry out an extensive range of test to perceive the risk in the scripts and compare the results using know variables in NLP to quantify the precision, and other quantifiers.   + Evaluate the results and consider new parameters to calibrate the classifier.   + Save all the process and start again with a new classifier until reaching a satisfying rate of success. * Write a report of all the process and the results found during the project.   **Research motivation**   * **D&D (decision and disruption)**   + Board game simulating a cyber security environment and analysing the behaviour of the players. It analyses their risk-based decision and process, taking in consideration their different technical backgrounds. Unfortunately, the results of these analysis cannot be accepted because of the experiment (a game) is not scalable. * **NLP Gap**   + Motivation started with a need to prove the results of the research on D&D with a low amount of experiments that could be helped with the use of NLP. It became an interesting topic when we realize there was a gap in the field of NLP for risk perception and possible interesting scientific and commercial outcomes.   **Possible outcomes**   * **Commercial** * **Scientific** | |

|  |
| --- |
| ***Summary of proposed research and analysis methodology*** |
| **Problem and Starting point**  A period of learning will be needed to attack the project. As well as a familiarization with the testing process. Other than that, the gameplay speeches have already been transcribed and these scripts that will be used to train the classifiers and to carry out the test, have already been pre-processed by professional transcribers.  **The approach**  We will be using an iterative approach to test out the different possibilities as follow:   1. Learning Period:    1. Machine Learning and NLP theory    2. Practice with simple sets and training with NLTK 2. Data Preparation:    1. The transcripts will need to be adjusted to be absorbable by the classifier. 3. Testing different classifiers:    1. Choose an approach    2. Implement the classifier    3. Test out the classifier    4. Analyse the results and go back to 3.a. with the necessary adjustments    5. Possible change of library and back to 3.a. 4. Document all the project in a final report.   **DIFFERENT COMPONENTS OF THE PROJECT**  **Learning Period**  Machine Learning: Very good tutorials available online (udacity)  NLTK: the library is very well documented and has a tutorial on their main page that seems very complete.  **Data Preparation**  After the learning period, the necessary adjustments will be done to the transcripts to suit our needs and to prepare them for our approach.  **Testing different classifiers**  There is a widely method when it comes to testing classifiers for NLP related problems and assess their accuracy and strength. It consists in using 90% percent of the available data sets and using them as training for the classifier to test out the remaining 10%. The process is repeated with a different 90% of the data sets until all the combinations have been carried out. Each step is assessed by variables such as the accuracy or the recall.  These variables allow to recalibrate the classifier or change the approach to improve the scores.  This method is repeated until the scores reach a certain predefined threshold where the classifier is applicable for other similar problems. |

|  |
| --- |
| ***Research plan – Gantt chart or Pert chart*** |
| **Pert Chart** |

|  |
| --- |
| ***Ethical statement*** |
| **Licensed Software**  The project will be using licensed software and the project lead will be responsible to ensure any software installed or used is:   * Used within the agree license terms of agreement * Installed within the license terms of agreement * Uninstalled at the end of the project and any copies of license key numbers handed back and copies destroyed * Record kept of all licenses used, including any relevant information such licenses keys * Any software media is handed back at the end of the project |

|  |
| --- |
| ***Legal and commercial aspects*** |
| **Commercial**  It is not intended that the project will produce any intellectual property as all software used in the project is commercially available. However, when working on site in data centres anyone working on the project will need to comply with confidentiality and sign non-disclosure agreements as required if working on company premises and particularly in any data centres.  **Legal**  Agreement with any commercial organisation on sharing outcomes of this project must be agreed before starting engagement. The testing of the access control model will not include any vulnerability tests of the cloud infrastructure environment, unless with clear written permission of the cloud vendor and following any change control procedure as required. |