DT211 Project Proposal Form 2014/2015

Student Name: Shane Quinn	**Student Number:** **C11759215**
Project Title	Using Natural Language Processing to Map Character Interactions
Summary (approx 200 words)	The aim of this project is to create a program that will analyze the text of a novel or short story using Natural Language Processing (NLP) techniques, such as Named Entity Recognition (NER) and Coreference Resolution. Using these techniques the program will learn to differentiate or recognise which words are the names or pronouns referring to specific characters. The program will then create a graph of the interaction between the various characters over the course of the source text.
Background (and References)	Natural Language Processing (NLP) is a field of computer science that deals with how computers interact with human language. There a large range of task covered by NLP. This project will use several of these tasks: Information Extraction: This often involves extracting data from unstructured sources and putting a structure to it. Part of Speech Tagging (PoS): PoS is concerned with tagging parts of a sentence in terms of noun, adjectives, verbs etc. This technique will allow for identification of ‘entities within the text’ . Named Entity Recognition (NER): NER is used to find and classify elements or entities in the text into pre-defined categories such as the names of persons, organizations, locations, quantities etc. Coreference Resolution: This is the task of finding all expressions that relate to a particular entity. For example, the pronoun ‘she’ might refer to a previous expression such as ‘the woman’ or ‘Ann’, but would not relate to ‘Barry’. ‘Natural Language Processing with Python’ Chapter 7: Extracting Information from Text http://www.nltk.org/book/ch07.html ‘Jumping NLP Curves: A Review of Natural Language Processing Research’

	http://sentic.net/jumping-nlp-curves.pdf ‘Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules’ http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00152
Proposed Approach	**Research** ● Feasibility study. ● NLP research. ● Develop Python skills. **Design Stage** ● Define system requirements. ● Develop a system architecture design. **Implementation Stage** ● Develop method for processing raw text. ● Develop method for categorizing and tagging words. ● Develop and train system to extract information from text. ● Identify and define entities to be tracked by the system. ● Implemented graphing of the collected data. **Testing Stage** ● Validation and verification of system requirements. ● Final testing.
Deliverables	● Raw text Processing. ● Sentence analysis and word tagging. ● Information extraction from raw text. ● System training. ● Graphing functionality.
Technical Requirements	● Python ● Natural Language Toolkit (NLTK or equivalent) ● Graphing Library

Project Reviews – Please include reviews of two of LAST years projects from your programme.

Project 1 **Title: Source Code Visualisation** **Student: Ivan Bacher** Description (brief): Analyzing static source code files can be a time consuming and tedious task, even for experienced developers. Often visualization techniques such as UML diagrams are used to speed up the process. The aim of this project is to implement 2D/3D visualization mechanism based on the program's source code. What is complex in this project ● 3D visualization of data that has be pulled from an unstructured source What technical architecture was used ● Code Parser ● Javascript Frameworks (AngularJS and ExpressJS) ● WebGL Explain key strengths and weaknesses of this project, as you see it. **Strengths** ● Useful tool for visualising source code ● More automated that previous techniques at code visualization (eg. UML) **Weaknesses** ● Badly structured code may affect the visualizations	
Project 2 **Title: Third Level Education Course Suggestion Web Application** **Student: Glen Holmes** Description (brief): This web application uses complex algorithms to help the decision process for leaving	

certificate students when they are choosing a third level college course. What is complex in this project: ● Algorithms to decide what course ● Specificity due to students interests What technical architecture was used ● Web Application Explain key strengths and weaknesses of this project, as you see it. **Strengths** ● Very complex ● If incorporated correctly could be very widely used **Weaknesses** ● Hard to complete in the given time frame as it seems to be quite a large project ● AI characteristics could skew the data and lead a student to the wrong course	
Proposal Sign off: Lecturer Comments	
Student Signature	**Date**
Lecturer Signature	**Date**