Ryan Rubadue

Professor Annexstein

CS 5001

7 September 2022

Individual Capstone Assessment

The purpose of this capstone project from my individual perspective focuses on understanding and classifying textual documents. From a very high-level, the project's aim is to take in information in a textual form, analyze the information, and develop classifications/associations on the data. In order to achieve this goal, a trained model will be utilized on one or multiple sufficiently sized data sets related to a specific subject area. An example subject area could be sports; for which an application might be classifying how related words/phrases are to certain individual sports. From this analysis, several visualizations of the data will be able to be displayed. In a more academic sense, achieving this classification has several implications on the ability to generate 'meaning' from text and improve how we are able to represent what is contained in documents beyond simply the words within them.

While I have not been exposed to any textual/semantic analysis throughout my CS courses, there are several courses I have taken that I think will prove useful throughout the project. One such course is Machine Learning (CS5137) where I gained knowledge related to classifying and fitting data along with other fundamental concepts of machine learning. These concepts will help in the classification portion of the assignment as we will be using some sort of machine learning algorithm. Another course I believe will prove useful in this project is Software Engineering (EECE-4029). Throughout the following two semesters, our group will need to employ strategies and components of the formal software engineering process. An additional course I am currently enrolled in is Requirements Engineering (CS5127), which will help with properly forming the requirements and features the project should contain.

During my Co-Ops semesters I have had the opportunity to work for two employers. My first three semesters of Co-Op I spent with Northrop Grumman as a 'Computer Science Co-Op'. During this time, I gained experience working with Python, Java, and C# on several projects including web applications and desktop applications. The skills I learned during these semesters include foundational software/programming skills and a wide-variety of fundamentals for different software projects. My final two Co-Ops were spent at an architecture firm called SHP. During this time I worked in C# and gained a variety of both front and backend experience working on plugins for a popular 3d modeling program. I believe these experiences have helped prepare me to assist in leading a larger-scale software project and will allow me to be successful in this course.

There are several aspects of this project that make it particularly appealing to myself. One such aspect is the machine learning and analysis aspect of the project. The application area of generating meaning from text documents is extremely interesting to me and I believe could result in wide-ranging meaningful applications. Another reason I am interested in this project are the potential libraries that could prove useful throughout the project duration. While I have significant experience working with Python, I have largely not utilized any machine learning or data science libraries in the way that they would be for this application. In terms of a preliminary solution, I think it would be helpful to start small and work upwards. Throughout the project duration we can work to classify larger text components and continuously improve our classification algorithm, thus a good starting piece would be successfully classifying singular words in text documents.

By the end of this project I expect to be able to generate meaningful analysis and classifications for text documents. To expand on this, I believe our project should be able to associate subcomponents of documents it parses with a specific topic within the subject area. As a stretch goal, the project could automatically generate some text based on a small portion of text passed in related to the subject area (autocompleting what the program believes might be said along with the word/phrase passed in). In order to self-evaluate my contributions I will set personal goals throughout the semester alongside the team goals the group sets. Additionally, I will look for feedback from my other members and faculty advisor on how I am doing throughout the term.