Yahya Emara
M12871276

Assignment 3 – Capstone Assesment

The main goal of this project is focused on clustering and classifying text documents in visual semantic spaces. As an aspiring master's student in Machine Learning, especially NLP, this project is an eye opener and a preparation for me to start my graduate journey. As I intend to work with Dr. Ali Minai during my graduate studies, we chose to also work with him in collaboration for this project as our advisor. Furthermore, the project can become a stepping stone for me in machine learning, where I plan to integrate it into Dr. Minai's research and to be able to contribute to a bigger project.

Even though NLP is a field I am still exploring, I am not completely new to Data Analysis and Machine Learning. My experience with the Machine Learning (EECE 5137) course and the Intelligent Systems (EECE 5136) course have given me a good foundation in the field. These two classes have allowed me to understand building ML models, neural networks and classification. My experience in data analysis through courses is based on the Probability and Random Processes (EECE 5119) course, which allowed me to understand statistics and data analysis on a deep level which I consider to be essential for a strong Machine Learning background.

Throughout my co-op experiences, I have come across machine learning projects. First, before any of my co-ops, as a Data Analysis intern with Dr. Danny Wu where I was introduced through bio-informatics projects. Later, I have done 5 of my co-ops at Infinera as a software engineering intern. At Infinera, part of my responsibilities was developing data analysis scripts to compare different chips' efficiencies. I also worked on a Machine Learning based resume parser project with other co-ops where various text analysis and classification techniques including NLTK. This project allowed me to put my machine learning skills into practice and resulted in the tool now being used all across the company.

The project is a chance for me and my teammates to be able to develop our skills on a practical level through an idea that is solely based on a passion and interest of ours. Additionally, the project allows me to work in a team environment where we bear all the responsibility from the idea to the implementation, which is very helpful, especially as I plan to pursue a graduate degree. Our preliminary approach to this project is that we will start with research. We have contacted Dr. Ali and were able to find very useful resources including research papers and thesis done at his lab that can help us in gathering data as well as understanding more about word embedding and vectorization. Our expected result will be to develop a semantic space where documents such as sports articles, congress speeches, etc. can be classified correctly where it is not too vague and not too specific. We will measure our success not just by the ability of classification of documents, but also by being able to reduce the vector space dimensions as much as possible, which can be challenging