Assignment 1

1 Introduction

This assignment focuses on the ChatGPT Embeddings API, a tool developed by OpenAI. This API allows users to extract vector representations, also known as embeddings, from any textual data. These embeddings encapsulate the underlying semantics and structure of the text, which can then be used for numerous machine learning tasks such as text classification, sentiment analysis, and information retrieval, among others.

2 Tasks

The tasks for this assignment are as follows:

1. Download a dataset from Kaggle.com that interests you in terms of what it is trying to predict (classification, sentiment analysis, prediction, etc). Make sure the dataset contains **at least one** text feature column. This data will be used to explore the functionality and application of the ChatGPT embeddings API.

2. Use the ChatGPT embeddings API to generate embeddings from the text feature in the dataset. Then use these embeddings to build a prediction model for a relevant output (label) in the dataset. Please refer to the documentation for detailed guidelines on how to use the embeddings API for various cases.

3. Before you begin coding, create a Confluence page outlining your project plan in one document. This should include a brief introduction of the Kaggle dataset you have chosen, your plan for generating embeddings, and your strategy for building the prediction model.

4. Create at least three Jira tickets that outline the steps you plan to take to complete this project. The tickets should cover the stages of the project from data acquisition, embedding generation, and model

training and evaluation. As you go through the stages, you should change the tickets from To-do, to In-progress and then complete.

3 Deliverables

Upon completion of this assignment, you are required to submit the following:

1. A video recording demonstrating the prediction model you have created. Run some live examples of how the model is able to predict or analyse whatever you are trying to do. Not that video presentations must be live DEMOS of the code and model - they should not be you just going over documents or code.

2. A python notebook containing the code used to download the Kaggle dataset, generate the embeddings using the ChatGPT embeddings API, and build the prediction model.

3. A PDF and link to your Confluence page detailing your project plan.

4. A report on the Jira tickets you created, including a brief description of what each ticket covers and how it contributes to the overall project. Show a screenshot of these tickets in Jira (so it is clear that you actually did create them)

5. A brief report discussing your results and any insights or challenges you encountered during the project.

4 Marking Scheme

The grading for this assignment will be distributed as follows:

- **Video Recording (30%)**: This includes the clarity and quality of your explanation of the process and results. You must speak during your demo (in English).
  Video presentations must actually show a demo of your model. Video recordings that just browse over the code (without running it) and are not clearly demonstrating the model working will severely impact the mark received. If you cannot get things working, just explain what you

think went wrong, don't try to hide it!

If there is no video presentation, there is no marks given for the assignment.

- **Confluence Page (20%)**: This includes the thoroughness and clarity of your project plan as outlined on your Confluence page with proper headings. Sample project plans are available from Atlassian, you can use a tempalte that you like.
- **Jira Tickets (20%)**: This includes the completeness and clarity of your Jira tickets, as well as how well they capture your project process.
- **Code and Correct Problem Solution (30%)**: This includes the correctness of your code and the proper algorithm used for the problem at hand. For example, you should not be using linear regression to predict a categorical output.

As mentioned above,failure to submit a video recording or submitting a video recording without any verbal explanation will result in a zero in assignment (no part marks for submitting code etc). The reason for this is that the video is pivotal in demonstrating that you have done the work. If you cannot get a complete model working, just explain what did work and where there was issues, this still gets you most of the marks.