ML2 Quiz 1 / Survey Team Answers

Question 1 - Please write down your team # and the title of the project

Team #13 - Movie Script Semantic Search

Question 2 - How many members do your team have

3

Question 3 - What GPU resource do you plan to use?

Google Colab

Question 4 - Please describe the data that you plan to use. (e.g., source, how many images, how many training images, how many testing images etc.)

We are using a movie scripts corpus of almost 3,000 movie script texts and annotations by structural elements. There is metadata for each movie including elements such as movie title, actors, awards, year, etc.
Source: https://www.kaggle.com/datasets/gufukuro/movie-scripts-corpus

Question 5 - Have you downloaded the data and take a look at it?

Yes - the data is in a large number of text files. For each movie there are several files: raw script, dialog excerpts by character, and metadata

Question 6 - Please show the link of the code that you would like to build on (e.g., Github link, HugginFace link, CoLab link?) or you plan to build it from scratch.

Code link
https://github.com/kstathou/vector_engine

Related tutorial
https://towardsdatascience.com/how-to-build-a-semantic-search-engine-with-transformers-and-faiss-dcbea307a0e8

Question 7 - Have you checked your code to see if it is available.

Yes - see link above. This utilizes a pre-trained sentence transformer model, specifically the distilbert-base-nli-stsb-mean-tokens model.

Question 8 - What would be the final evaluation metrics of your project? (Accuracy, F1 score, AUC etc.?)

> Normalized Discounted Cumulative Gain (NDCG). It'sa popular method for measuring the quality of a set of search results

Question 9 - Do you have any difficulties or questions? You can put it here.