Project 2

Each team has two members. Every student should submit the code and report (in pdf format) on Blackboard system before 23:59 May 21. Report can be written either in English or Chinese. Name your report as `studentID_Name.pdf` and your code as `studentID_name.zip`. **Remember to include the project contribution of each team member in the report.**

Q1 Question and answering system

Design and build a question answering system with Pyspark. The data set is from SQuAD v2.0.

(1) [2 points] Write python code to process data. Use the official dev set as test set, and split the original training set into training set and validation set (5000 samples). Prepare the data according to the requirements of ML model training.

(2) [4 points] Read the documents/examples/source codes of Huggingface Transformers. Write the bash script and python code to train the QA model using a pretrained DistilBERT model. As no GPU is available in the server, you can use pytorch-cpu to debug your code and train the model for a few hours. The pretrained DistilBERT model is available at here.

(3) [6 points] Deploy the trained QA model with Spark-NLP, and predict the data of the test set in a streaming processing manner using Kafka. You can also use the finetuned DistilBERT model here.

(4) [3 points] Read the tutorial and illustrate one possible method that can support open-domain question answering. The code is not required for this task.

Q2 Startup Analyses

Select two startups you are interested that are related to AI or data science and write an analyses report. No minimal page limit. [6 points]

You can follow the SWOT pattern and can include the following in your report:

- The key strength and niche of the company.
- Are there any threats to the company?
- What do you learn from the public talks or interviews of the entrepreneur/company?
- Are there any similar oppotunites in China that are inspired by the US companies?

Note: You can also write the report with ChatGPT/New bing, etc. In this case, please claim the AI system you use in your report. Write all the questions/prompts you use to chat with the bot. **Also list at least three cases where ChatGPT/New Bing fail or work out of your expectation.**

References

- AI 100: The most promising artificial intelligence startups of 2022
- Generative AI Startups in 2023
- Product Hunt
- Hacker News – Y Combinator

Presentation

[4 points] Presentations are on May 26. Submit the ppt files after the presentation. Each team has 12 minutes.