Project 2

Every student should submit the code, report (in pdf format) and ppt on Blackboard system before 23:59 June 2. Report can be written either in English or Chinese. Name your report as `studentID_Name.pdf` and your code as `studentID_name.zip`.

Q1 Question and answering system

Design and build a **generative** question answering system. The training data set is from SQuAD v2 Dataset (also located in the server path `/shareddata/data/project2`).

(1) [2 points] Write python code to process data. Use the context and question as input, and the answer as output. Use the official validation set as test set, and split the original training set into training set and validation set (5000 samples for valid set, the rest for train set). Prepare the data according to the requirements of model training (Can refer to the original T5 and Flan-T5 paper for data format). You can use either Pyspark or pure python code.

(2) [5 points] Write the bash script and Ray-train python code to train the QA model by further funetuning the Flan-T5-small model. As no GPU is available in the server, you can use pytorch-cpu to debug your code, train the model for a few hours and save a checkpoint. Note that the validation set is used for designing the hyperparameters and selecting the model checkpoint. You can also refer to the training examples in the huggingface repo. You can also rent the GPU server in AutoDL.

(3) [4 points] Deploy the finetuned QA model with Spark-NLP, and answer the questions of the test set in a streaming processing manner using Kafka. You can search the spark-nlp model here. Note that the deployed model is not the original Flan-T5-small model.

(4) [2 points] Read a survey paper about retrieval-augmented generation and illustrate one possible method that can support open-domain question answering with system diagram and pipeline introduction. The code is not required for this task.

Q2 Startup Analyses

[4 points] Select one startup you are interested from the following candidates. Suppose you are an investor who is interested in the startup. Write an analyses report about the startup and their product. More than one page.

> Startup candidates: Ray, Hudi, Iceberg, Snowflake, Clickhouse, Doris

Follow the SWOT pattern and include the following in your report:

- The key strength and niche of the company and their products.
- The detail description and analyses of the product.
- The future trend of the targeted market.
- Are there any threats to the company?
- What do you learn from the public talks or interviews of the entrepreneur/company?

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. You are responsible to check the content and make sure that the content is correct. 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.**

Q3 Paper Reading

[4 points] Read one of the following papers and write a report more than one page. Follow the suggestions here to organize your report.

- Spark SQL: Relational Data Processing in Spark
- Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing
- Ray: A Distributed Framework for Emerging AI Applications
- Ray 2.0 Architecture whitepaper
- Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores
- AlpaServe: Statistical Multiplexing with Model Parallelism for Deep Learning Serving

Presentation

[4 points] Presentations are on June 4 & June 5. Each student has 12 minutes. 1~2 related questions will be asked after the presentation.