به نام خدا

Question Answering

شرح Task

A) مدل Llama-3.2-1B-bnb-4bit، روی مجموعه داده PQuAD برای تسک پرسش‌پاسخ Fine-Tune کنید (در صورت محدودیت سخت افزاری امکان استفاده از دیتاست PersianQA هم دارید). نتایج معیارهایF1-Score و Exact Match (EM) گزارش دهید. در فرآیند آموزش برای کاهش مصرف منابع استفاده از تکنیک‌هایی مانند LoRA یا QLoRA مجاز است اما از روش های Zero-Shot و Few-Shotاستفاده نکنید.

B) یک مدل امبدینگ مثل paraphrase-multilingual-MiniLM-L12-v2 را روی داده‌های پروژه Fine-Tune کرده، سپس از روش RAGبرای بازیابی اطلاعات و پاسخ‌دهی به سوالات پیاده سازی کنید. توجه شود فایل pdf پیوست شده را در فرآیند Chunking یک بار به صورت word-based و بار دیگر به صورت Sentence-based پیاده سازی نمایید. برای مدل بازیابی از TF-IDF یا BM25 استفاده شود. در انتها نتایج معیارهای F1-Score و Exact Match (EM) را محاسبه و باهم مقایسه نمایید. همچنین از نظر Precision، Recallو Hit@kتحلیل کنید.

C) میزان تشابه معنایی (Semantic Similarity) بین جواب مدل و جواب اصلی را با معیارهای Cosine Similarity و MRR محاسبه کنید.

D) یک مدل امبدینگ مثل distiluse-base-multilingual-cased-v2 ، روی داده‌های پروژه Fine-Tune کرده و عملکرد فرآیند بازیابی را بررسی نمایید. سپس میزان تشابه معنایی را با معیارهای Cosine Similarity و MRR محاسبه کنید.

E) چند مدل امبدینگ مثل multilingual-e5-base و هر مدل مناسب دیگری ، روی داده‌های پروژه Fine-Tune کرده و عملکرد فرآیند بازیابی را بررسی و مقایسه ای از مدل ها ارائه کنید. برای بهبود کیفیت فضای امبدینگ روش های مختلف Pre-Processing را پیاده سازی نمایید. جهت ارتقای سرعت و کارایی فرآیند بازیابی بردارها، استفاده از پایگاه‌های داده‌ی برداری مانند FAISS، Chroma، LanceDB و سایر پایگاه داده های مناسب نیز توصیه می‌شود. در انتها تحلیل نتایج بر اساس معیارهای Cosine Similarity، MRR، Precision، Recallو Hit@k، ملاک ارزیابی شما در این بخش خواهد بود.

F) برای بهترین مدل QA، یک واسط کاربری طراحی نمایید که در آن کاربر بتواند با وارد کردن متن سوال، پاسخ تولیدشده توسط را مشاهده نماید. استفاده از Gradio، Streamlit، Fast-API و سایر ابزار های مناسب، مجاز است.

ساختار تحویل Task

- کدها در قالب یک فایل زیپ یا لینک به یک مخزن گیت‌هاب باشند.
- مستندات در قالب PDF یا Markdown نوشته شوند.
- پیش‌پردازش داده‌ها، معماری مدل، ارزیابی عملکرد، و چالش های مواجه شده شرح داده شود.
- پیاده سازی ها باید قابلیت تست گرفتن توسط مصحح را داشته باشند.
- اگر دسترسی به GPU و سخت‌افزار محدود دارید استفاده از Google Colab و یا Kaggle مجاز است.

منابع مورد نیاز :

- لینک دیتاست ها :

https://huggingface.co/datasets/Gholamreza/pquad
https://huggingface.co/datasets/SajjadAyoubi/persian_qa

- لینک مدل ها :

https://huggingface.co/unsloth/llama-3-8b-bnb-4bit

https://huggingface.co/Qdrant/bm25

https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

https://huggingface.co/intfloat/multilingual-e5-base

- برای تنظیم مجدد فضای امبدینگ :

https://medium.com/llamaindex-blog/fine-tuning-embeddings-for-rag-with-synthetic-data-e534409a3971

https://github.com/run-llama/finetune-embedding#steps-for-runnin

راه ارتباطی برای پرسش و پاسخ :

- از طریق ایمیل : amin.rezanejad.edu@gmail.com
- از طریق تلگرام : @Amin_Rezaneajd

همیشه در پناه خدا موفق باشید

امین رضانژاد