به نام خدا

درس پردازش زبان طبیعی

دکتر مرضیه داود آبادی

تمرین سری ششم

طراح تمرین: محمد شادفر

مهلت تحویل: 1404/03/16

نکات تکمیلی

1. پاسخ سوالات را به صورت کامل در یک فایل PDF و به همراه کدهای سوالات یک فایل قرار داده و تا زمان تعیین شده StudentNumber_FirstName_LastName_HW6.zipفشرده به شکل بارگذاری نمایید .

2. برای پیاده سازی ها زبان پایتون پیشنهاد می شود، لازم به ذکر است توضیح کد ها و نتایج بدست آمده، باید در فایل PDF آورده شوند و به کد بدون گزارش نمره ای تعلق نخواهد گرفت.

3. به ازای هر روز تاخیر 50 درصد از نمره تمرین کسر خواهد شد .

4. لطفا برای انجام تمرین زمان مناسب اختصاص داده شود و انجام آن را به روزهای پایانی موکول نکنید.

5. بد نیست منابع استفاده شده در حل هر سوال را ذکر کنید .

6.خلاقیت نمره اضافی دارد.

موفق باشید

بخش اول : راه‌اندازی اولیه و اتصال

1. **راه‌اندازی OpenWebUI با داکر:**
   - OpenWebUI را با استفاده از داکر اجرا کنید.
2. **سرو کردن مدل LLM با vLLM و اتصال به OpenWebUI**
   - یک مدل LLM را انتخاب کرده و با استفاده از vLLM آن را سرو کنید.
   - مدل سرو شده را به OpenWebUI متصل کنید. ملاحضات امنیتی برای سرو کردن مدل را در نظر بگیرید.

بخش دوم: مقایسه عملکرد مدل ها

3. **مقایسه عملکرد مدل:**
   - دو مدل LLM مختلف (به عنوان مثال، یک مدل کوچکتر و یک مدل بزرگتر) را انتخاب کنید.
   - هر دو را با استفاده از vLLMسرو کنید و آنها را به OpenWebUIمتصل کنید.
   - مجموعه‌ای از 3-5 اعلان (prompt) خاص (مانند خلاصه‌سازی، پاسخ به یک سؤال واقعی(fact check)، تولید متن خلاقانه) را تدوین کنید.
   - برای هر اعلان، پاسخ‌های دو مدل را بر اساس موارد زیر مقایسه کنید.
   - یک تحلیل کوتاه بنویسید که مشاهدات خود را خلاصه کرده و دلایل احتمالی تفاوت‌ها را توضیح دهید.

بخش سوم: کاوش عمیق‌تر در پیکربندی vLLM

4. **پارامترهای نمونه‌برداری (Sampling) و تأثیر آنها:**
   - تحقیق کنید و توضیح دهید که پارامترهای temperature، top_pو top_kدر استنتاج LLM چه کاری انجام می‌دهند.
   - پاسخ‌هایی با تنظیمات بسیار متفاوت برای temperature، top_pو top_kتولید کنید.
5. **کوانتیزاسیون (Quantization) و مبادلات آن:**
   - مفهوم کوانتیزاسیون مدل (مثلاً 8-بیتی، 4-بیتی) را تحقیق کنید.