

تبدیل زبان طبیعی به کوئری‌های SQL با استفاده از مدل‌های زبانی بزرگ

سوال ۱

در این تمرین، هدف طراحی و ارزیابی یک سامانه‌ی تبدیل جملات زبان طبیعی به کوئری‌های پایگاهداده^۱ با استفاده از مدل‌های زبانی بزرگ^۲ است. همچنین دانشجویان باید با استفاده از معماری تولید تقویت شده با بازیابی^۳ و تحلیل حساسیت نسبت به پارامترهای گوناگون، عملکرد مدل‌ها را مورد بررسی قرار دهند.

Mجموعه‌داده Spider: این مجموعه‌داده، شامل جملات زبان طبیعی به همراه کوئری‌های متناظر SQL و ساختار پایگاهداده مربوط به هر پرسش است.

برای سهولت کار از نسخه ۱ این مجموعه‌داده در این تمرین بهره می‌گیریم. اطلاعات بیشتر در خصوص این مجموعه‌داده، تابلو امتیازات و ایده‌های مطرح برای حل این مسئله می‌توانید لینک زیر را مطالعه کنید.

<https://yale-lily.github.io/spider>

مجموعه‌داده مذکور به فایل تعریف تمرین، پیوست شده است.

۱ - پیاده‌سازی مدل پایه:

با استفاده از یکی از نسخه‌های کوچک مدل gemma مانند نسخه‌های^۴ ۱۲ میلیاردی از کتابخانه^۵ transformers، مدلی طراحی نمایید که جمله‌ای مانند:

“List all students who have registered for more than three courses.”

را به کوئری معادل در SQL تبدیل کند. مدل را بر روی داده‌ی آزمون مجموعه Spider ارزیابی و خروجی‌ها را تحلیل نمایید.

۲ - تحلیل حساسیت به پارامترها:

اثر تغییر هر یک از پارامترهای زیر را بر عملکرد مدل به صورت جداگانه بررسی کرده و نتایج را به صورت جدول یا نمودار ارائه دهید:

SQL Query^۱
Large Language Models - LLMs^۲
Retrieval-Augmented Generation - RAG^۳

- استفاده از تنظیم دقیق^۴ بر روی مجموعه داده Spider
- تاثیر تعداد گام گرم کردن^۵
- تنظیم دمای تولید^۶ در فرآیند تولید متن
- مهندسی پرامپت^۷:

در این بخش، با استفاده از روش‌های گوناگون مهندسی پرامپت، تأثیر نوع طراحی ورودی مدل را بر دقت نهایی بررسی نمایید:

- حالت بدون نمونه^۸: فقط پرسش را در اختیار مدل قرار دهید.
- حالت تکنمونه‌ای^۹: یک مثال مشابه به همراه پرسش اصلی به مدل داده شود.
- حالت چندنمونه‌ای^{۱۰}: چند مثال از ورودی و خروجی‌های صحیح به عنوان راهنمایی به مدل داده شود.
- تغییر در طول متن ورودی: آیا حذف یا خلاصه‌سازی ساختار پایگاهداده (schema) باعث کاهش یا افزایش دقت می‌شود؟
- استفاده از توصیف‌های معنایی برای ستون‌ها به جای فرمت ساختاری خام نتایج را ثبت کرده و اثر هر روش را با مدل پایه مقایسه و تحلیل نمایید.

- افزودن معماری تولید تقویت‌شده با بازیابی^{۱۱} با استفاده از معماری RAG، سامانه را توسعه دهید به‌طوری‌که به جای گنجاندن مستقیم ساختار پایگاهداده در متن ورودی، ابتدا آن را ایندکس کرده و سپس در زمان پرسش از پایگاه بازیابی نمایید.
- از ابزارهایی مانند Chroma یا FAISS برای ایندکس‌گذاری استفاده نمایید.
- عملکرد مدل را پیش و پس از افزودن RAG مقایسه نمایید.
- اثر انواع مدل‌های تعبیه^{۱۲} مانند MiniLM و BGE را بررسی کنید.

Fine-Tuning^۴

Warm-up Steps^۵

Temperature^۶

Prompt Engineering^۷

Zero-shot^۸

One-shot^۹

Few-shot^{۱۰}

Embedders^{۱۱}

۵ - ارزیابی سامانه:

برای سنجش عملکرد مدل تبدیل زبان طبیعی به SQL، از دو معیار کلیدی زیر استفاده نمایید. برای هر پرسش، کوئری تولید شده توسط مدل را با کوئری مرجع مقایسه کرده و مقدار هر دو معیار را محاسبه نمایید. در پایان، میانگین هر معیار را بر روی کل مجموعه آزمون گزارش دهید.

- دقت تطابق کامل^{۱۲}:

در این معیار، بررسی می‌شود که آیا کوئری تولید شده توسط مدل، به صورت دقیق (با لحاظ ترتیب و ساختار) برابر با کوئری مرجع است یا خیر. اگر رشته‌ی کوئری تولیدی دقیقاً با کوئری هدف یکسان باشد، امتیاز ۱ و در غیر این صورت امتیاز ۰ دریافت می‌کند.

فرمول محاسبه برای کل داده‌ها به صورت زیر است:

$$EM = \frac{\text{تعداد پاسخ‌های کاملاً صحیح}}{\text{تعداد کل نمونه‌ها}}$$

- امتیاز شباهت ژاکارد^{۱۳}:

این معیار برای مقایسه دو مجموعه از توکن‌ها به کار می‌رود. در این تمرین، ابتدا هر کوئری به مجموعه‌ای از توکن‌ها (مثلاً کلمات یا عبارات جدا شده با فاصله) تبدیل می‌شود. سپس ضریب ژاکارد بین کوئری تولید شده و کوئری مرجع محاسبه می‌گردد:

$$JS(Q_1, Q_2) = \frac{|Q_1 \cap Q_2|}{|Q_1 \cup Q_2|}$$

که در آن Q_1 و Q_2 به ترتیب مجموعه توکن‌های کوئری تولید شده و کوئری مرجع هستند. امتیاز ژاکارد عددی بین ۰ و ۱ است. مقدار ۱ به معنای تطابق کامل و مقدار نزدیک به صفر به معنای عدم شباهت است.

در گزارش نهایی، برای هر مدل و تنظیم مختلف، میانگین Exact Match و میانگین Score را ارائه کرده و آن‌ها را در قالب جدول و نمودار مقایسه نمایید.

۶ - تحلیل مفهومی:

- مقایسه معماری RAG با تنظیم دقیق مستقیم مدل‌های زبانی
- چالش‌های تبدیل زبان طبیعی به SQL در مقایسه با سایر وظایف تولید متن

Exact Match Accuracy^{۱۲}
Jaccard Similarity Score^{۱۳}

- خطاهای رایج مانند تولید کوئری‌های ناصحیح یا غیرقابل اجرا (hallucination)
- راهکارهای پیشنهادی برای بهبود دقت و پایداری سامانه

در انجام تمرینات به نکات زیر توجه فرمایید.

- ۱ - پیاده‌سازی با زبان Python و چارچوب کاری PyTorch انجام شود. استفاده از torchvision برای بارگذاری داده‌ها مجاز است.
- ۲ - در صورت استفاده از مدل‌های از پیش آماده یا معماری‌های مقاله‌ای، منابع را به‌طور کامل ذکر کنید.
- ۳ - ارسال فایل فشرده شامل کد و گزارش با نام StudentID_HW08.zip در موعد مقرر الزامی است. در غیر این صورت نمره تمرین لحاظ نخواهد شد.
- ۴ - کپی‌برداری از دیگر دانشجویان یا استفاده مستقیم از کدهای آماده بدون درک کامل، موجب حذف کل نمره تمرین خواهد شد.