دانشگاه علم و صنعت ایران

دانشکده مهندسی کامپیوتر

درس پردازش زبان طبیعی

پیاده سازی تولید متن با استفاده از بازیابی (RAG) بر روی گزارش‌های سازمانی

استاد:

دکتر مرضیه داوودآبادی

همیار استاد:

محمدامین عباسی

تیر ماه ۱۴۰۳

Retrieval-Augmented Generation (RAG)یک رویکرد پیشرفته در زمینه هوش مصنوعی است که با هدف افزایش کیفیت تولید متن از طریق دسترسی به منابع اطلاعاتی خارجی توسعه یافته است. این فناوری، مدل‌های پیشرفته تولید متن مانند مدل‌های زبانی بزرگ را با سیستم‌های جستجوی پیشرفته اطلاعات تلفیق می‌کند تا امکان دسترسی به اطلاعاتی دقیق و به‌روز فراهم آورد، و در نتیجه، به ارائه پاسخ‌هایی با کیفیت تر کمک کند. در این فرآیند، ابتدا سیستم بازیابی اطلاعات (IR) مجموعه‌ای از مستندات یا اطلاعات مرتبط با پرسش یا درخواست ورودی را شناسایی می‌کند. سپس، با استناد به این اطلاعات بازیابی شده، مدل تولید متن پاسخ یا متن مورد نظر را ایجاد می‌کند.این پروژه با هدف ارتقاء پاسخ‌دهی به سوالات مرتبط با گزارش‌های سازمانی توسط RAG طراحی شده است. استفاده از این تکنولوژی این امکان را فراهم می‌آورد که به سوالات کاربران بر اساس محتوای موجود در گزارش‌های سازمانی پاسخ‌هایی دقیق و کنونی ارائه دهیم. در طراحی این پروژه، انتخاب مدل زبانی مناسب برای کار با RAG به عهده تیم توسعه قرار دارد. پیشنهاد می‌شود از مدل LLaMA 3 برای تولید متن و LaBse برای جانمایی استفاده شود. برای بهره‌گیری از قدرت پردازشی GPU، می‌توانید از پلتفرم Kaggle استفاده کنید.

فایل های پروژه:

- input 1 و input 2 منبع متنی پروژه می‌باشد.
- questions 1 و questions 2 نمونه سوالات مدنظر پروژه می‌باشد.

علاوه بر این، تیم اجرایی باید در پایان پروژه یک گزارش جامع ارائه دهد که تمام جوانب پروژه را پوشش دهد. این گزارش باید شامل بخش‌های زیر باشد:

- مقدمه: توضیح کوتاهی درباره پروژه، اهمیت استفاده از تکنولوژی Retrieval-Augmented Generation (RAG)در پاسخ‌دهی به سوالات مبتنی بر گزارش‌های سازمانی و اهداف کلی پروژه.
- پیش‌زمینه تکنولوژیکی: بررسی اجمالی از تکنولوژی‌ها و مدل‌های زبانی مورد استفاده در پروژه، از جمله LLaMA 3وLaBse ، و چگونگی ترکیب آن‌ها با سیستم‌های بازیابی اطلاعات.
- روند پیاده‌سازی: توضیح مراحل پیاده‌سازی RAG برای پروژه، شامل جزئیات فنی، چالش‌ها و راه‌حل‌های اتخاذ شده در طول پروژه.
- نتایج و بررسی‌ها: ارائه تحلیلی از نتایج به دست آمده، از جمله کیفیت پاسخ‌ها، بهبودهای مشاهده شده نسبت به رویکردهای قبلی و هرگونه اطلاعات آماری مرتبط.

- موانع و چالش‌ها: بررسی چالش‌هایی که تیم در طول پروژه با آن‌ها مواجه شده است، از جمله محدودیت‌های فنی، مسائل مربوط به داده‌ها و چگونگی غلبه بر این موانع.
- آموخته‌ها و پیشنهادات: ارائه درس‌هایی که از اجرای پروژه آموخته شده و پیشنهادهایی برای تیم‌هایی که در آینده قصد دارند پروژه‌های مشابهی را اجرا کنند.
- منابع: لیستی از تمام منابع، از جمله مقالات، وب‌سایت‌ها و ابزارهایی که در طول پروژه مورد استفاده قرار گرفته‌اند.

نکات:

- حتما از کتابخانه‌های موجود برای RAG مانند LLaMAIndex یا LangChain استفاده کنید.
- تیم اجرایی پروژه می‌تواند تا ۳ نفر باشد.
- تاریخ تحویل پروژه ۱۷ تیر می‌باشد.
- برای آشنایی بیشتر با روند پیاده‌سازی RAG، می‌توانید از لینک زیر به عنوان منبع آموزشی استفاده کنید.
  https://medium.com/@tejaswi_kashyap/rag-processing-using-llamaindex-43d9786f9d8e