گزارش سیستم پاسخگویی به سوالات دارویی با معماری RAG

۱ .پیش‌پردازش داده‌ها

- **استخراج متن :**از کتابخانه‌هایی مانند PyPDF2 جهت استخراج متن از فایل PDF استفاده شد.
- **تمیزسازی متن :**حذف علائم نگارشی، اعداد و فواصل اضافی، و تبدیل تمامی حروف به حروف کوچک برای یکسان‌سازی داده‌ها.
- **حذف توقف‌واژه‌ها :**استفاده از یک لیست از کلمات پرکاربرد (stopwords) برای حذف کلمات غیرضروری جهت کاهش نویز در داده‌های ورودی.
- **تقسیم‌بندی به چانک‌ها :**تقسیم متن استخراج‌شده به بخش‌های کوچکتر (چانک‌ها) به منظور بهبود دقت بازیابی اطلاعات.

۲ . معماری مدل

- **مدل‌های بازیابی اطلاعات:**
  - **BM25:**
    - پارامترهای $k1=1.6$ و $b = 0.7$ تنظیم شدند تا حساسیت به فراوانی کلمات و طول اسناد بهینه شود.
  - **TF-IDF:**
    - با تنظیم گزینه `sublinear_tf=True`، افزایش تعداد تکرار یک کلمه تأثیر لگاریتمی داشته و نویز کاهش پیدا میکند.
  - **ترکیب مدل‌ها :**
    - با توجه به اینکه هر کدام از این مدل‌ها به تنهایی پاسخگویی مطلوبی نداشتند، با وزن‌دهی ترکیبی مثلاً ٪۵۰ BM25 و ٪۵۰ (TF-IDF) سعی شد تا بهترین نتایج بازیابی حاصل شود.
- **مدل زبانی:**
  - از مدل **Llama-3.1-8B-4bit**به همراه تکنیک کوانتیزه‌سازی ۴-bit استفاده کردم تا با منابع محدود، بتونم پاسخ‌های دقیقی تولید کنم.
  - مدل زبانی بعد از دریافت پرسش کاربر و چانک‌های بازیابی‌شده، با استفاده از یه prompt بهینه‌شده، پاسخ نهایی رو تولید میکند.

۳ .ارزیابی عملکرد

- **Exact Match (EM):**
  - نسبت تطابق دقیق پاسخ تولید شده با پاسخ مرجع اندازه‌گیری میشود.
  - در نتایج اولیه، میانگین EM حدود ۰.۲۰ گزارش شده است که نشان‌دهنده پایین بودن تطابق دقیق است.
- **F1-Score:**
  - معیار F1 بر مبنای همپوشانی کلمات بین پاسخ تولید شده و پاسخ مرجع محاسبه میشود.
  - میانگین F1-Score حدود ۰.۵۰ بوده که نشان از عملکرد متوسط در همپوشانی واژگان دارد.

برای ارزیابی، یک دیتاست تست خودم نوشتم که شامل ۱۰ سوال به همراه جواب مرجع است. این ارزیابی‌ها بهم نشون دادن که اگرچه از نظر F1 عملکرد قابل قبولی داریم، اما تطابق دقیق (EM) نیاز به بهبود بیشتری دارد تا سیستم بهتر بتواند پاسخ‌های دقیق‌تری ارائه دهد.

۴ .چالش‌های مواجه‌شده در مسیر انجام Task

•تنظیم و ترکیب مدل‌های بازیابی:

۱. هر کدام از مدل‌های BM25 و TF-IDF به تنهایی جواب مناسبی ندادن؛ بنابراین احساس می‌کردم باید با ترکیبشون، نتیجه بهتری بگیرم.
۲. پیدا کردن وزن‌دهی مناسب برای ترکیب نتایج این دو مدل از اهمیت زیادی برخوردار بود تا بتونم چانک‌های مرتبط‌تری استخراج کنم.

•هماهنگی بین مدل‌های بازیابی و مدل زبانی:

۱. اصلی‌ترین چالش، ساخت یه prompt مناسب برای ترکیب نتایج بازیابی شده و تولید پاسخ‌های دقیق توسط مدل زبانی بود.
۲. همچنین تنظیم پارامترهایی مثل temperature و top_p برای بهبود دقت و مرتبط بودن پاسخ‌ها، مهم بود.

•اجرای واسط کاربری Gradio

۱. اجرای gr.Interfaceدر Colab گاهی با خطای قطع شدن اتصال (Disconnected) به دلیل محدودیت زمان اجرا مواجه می‌شد. (هنگام اتصال، از شکن برای دیدن Interface استفاده کنید)

- **تولید پاسخ با Llama-3**

۱. مدل Llama-3 برای داده‌های پزشکی فارسی **فاین‌تیونینگ نشده** بود و در مواردی پاسخ‌های عمومی یا **غیرمرتبط** تولید می‌کرد.
۲. **تکرار جملات** در پاسخ‌های طولانی (مثال: پاسخ به سوال "جایگزین کلرفنیرامین")

توجه : این پروژه در محیط Google Colab پیاده‌سازی شده و نسخه‌های نصب شده کتابخانه‌ها متناسب با coda 12.4 Colabمی‌باشند.