

# به نام خدا

## موضوع پروژه:

پیاده سازی تولید متن با استفاده از بازیابی (RAG)

بر روی گزارش های سازمانی

## استاد:

دکتر مرضیه داوودآبادی

## همیار استاد:

محمدامین عباسی

## اعضای گروه:

سینا علی نژاد ۹۹۵۲۱۴۶۹

نوید ابراهیمی ۹۹۵۲۱۰۰۱

امیررضا ویشه ۹۹۵۲۲۲۲۱

تیرماه ۱۴۰۳

## - مقدمه

تکنولوژی RAG با ترکیب مدل‌های بازپایی و تولید محتوا، امکان پاسخگویی دقیق‌تر به سوالات را فراهم می‌کند. این تکنولوژی با استفاده از اطلاعات موجود در گزارش‌های سازمانی، پاسخ‌هایی مناسب و مرتبط با محتوای گزارش ارائه می‌دهد. همچنین، RAG باعث افزایش دقت و کاهش خطا در پاسخ‌ها می‌شود. در نهایت، استفاده از RAG می‌تواند به بهبود کیفیت خدمات سازمان کمک کند.

## - پیش‌زمینه تکنولوژیکی

مدل Qwen/Qwen2-7B-Instruct یک مدل زبان بزرگ است که با استفاده از معماری تبدیل‌کننده با فعال‌سازی SwiGLU، توجه QKV و توجه گروهی ساخته شده است. این مدل با پیش‌آموزش بر روی مقدار زیادی از داده‌ها و پس‌آموزش با استفاده از تنظیمات نظارت شده و بهینه‌سازی ترجیح مستقیم، عملکرد بسیار خوبی در فهم زبان، تولید زبان، توانایی چندزبانه، برنامه‌نویسی، ریاضیات، استدلال و غیره از خود نشان داده است. همچنین، این مدل تا ۱۳۱،۰۷۲ توکن را پشتیبانی می‌کند.

مدل sentence-transformers/LaBSE یک مدل BERT مبتنی بر جمله است که برای تعبیه جمله برای ۱۰۹ زبان آموزش داده شده است. فرآیند پیش‌آموزش شامل مدل‌سازی زبان ماسک‌دار با مدل‌سازی زبان ترجمه است. این مدل برای دریافت تعبیه چندزبانه جمله و بازپایی دومتنی مفید است.

نحوه ترکیب این دو مدل بدین صورت است که مدل LaBSE ارتباط هر سوال با Chunkهایی که از پیش ساخته را پیدا می‌کند و k Chunk شبیه را برمی‌گرداند. در ادامه این اطلاعات به عنوان Context به مدل داده می‌شود. مدل Qwen با استفاده از این داده‌ها و داده‌های خودش جواب نهایی را تولید می‌کند.

## - روند پیاده‌سازی

در ابتدای کار مرحله import داده‌ها به پروژه است. قبل از وارد کردن داده‌ها به پروژه یک عملیات preprocessing توسط chatGPT بری Document ۱ انجام دادم (به دلیل نامرتب بودن زیاد آن). در ادامه بعد از ورود داده‌ها، همه متون را با استفاده از Normalizer Hazm نرمال کردم تا خروجی بهتری داشته باشیم.

برای سوالات نیز پس از تغییر نام ستون سوالات به questions، فقط داده‌های همین ستون را نگه داشتیم و بقیه داده‌های فایل‌های سوال را حذف کردم. همچنین سطرهایی که داده NaN هم داشتند را نیز حذف کردم.

بخش LangChain: در این بخش که می‌توان گفت بخش اصلی RAG پروژه هست ارتباط هر سوال به Chunk‌های متون را یافتیم. به این صورت که هر متن به یک سری Chunk تبدیل شد و LangChain با استفاده از مدل LaBSE ارتباط سوال را با این Chunk‌ها می‌یافت و k چانک شبیه را برمی‌گرداند که در این پروژه ما تعداد Chunk را برابر ۴ گذاشتیم.

بخش Text Generation: در این بخش به علت محدودیت دسترسی به مدل llama از مدل Qwen استفاده کردیم. بدین صورت که چون پروژه از ما تاثیر RAG در تولید جواب نهایی را می‌خواست، یکبار جواب را بدون Context و یکبار همراه با Context به مدل دادیم و جواب را تولید کردیم.

## - نتایج و بررسی‌ها

همانطور که در فایل qa مشخص است، جواب‌های تولیدی توسط تکنولوژی RAG دارای دقت بیشتری نسبت به جواب‌های بدون این تکنولوژی هستند. مثال زیر نشان‌دهنده همین موضوع است:

Question: با سلام احترام خدمت شما من سال پیش در منطقه ۱۴ در ریزدانه ملکی خریداری کردم به متراژ ۷۲ متر چون تشکیل پرونده من برای تخریب و نوسازی بعد از برج ۵ بود طبقه چهارم و ۲+ به این پرونده تعلیق نگرفته می‌خواستم بدونم صبر کنم شهرداری و شورای شهر تصمیم به برگردوندن به روال قبل رو داره آیا امیدی هست یا بازم گرفتاریم شماره پرونده ۱۴۰۱۰۲۷۵۶ ممنون از پاسخ شما

Answer\_with\_RAG:

با توجه به اطلاعات موجود، نمی‌توان به صورت دقیق و مطمئن اعلام کرد که آیا شهرداری تهران و شورای عالی شهرسازی و معماری ایران تصمیم به برگردوندن به روال قبل یا نه را گرفته‌اند. اما بر اساس متن مورد نظر، می‌توان گفت که:

۱. شهرداری تهران مجاز است باغات و فضای سبز شهر تهران را برای تبدیل به فضای سبز عمومی در چارچوب قوانین و مقررات موضوعه و با حفظ حقوق مالکانه تملک نماید.
  ۲. باغات پیوسته و یکپارچه که به صورت انبوه و در کنار هم قرار گرفته و در نقشه‌های طرح تفصیلی شهر تهران، به عنوان پهنه محسوب شده‌اند، مشمول مفاد فصل اول سند ضوابط و مقررات طرح تفصیلی شهر تهران است.
  ۳. در فرایند بازبینی طرح تفصیلی، شهرداری تهران ملزم به تعیین سهم و ساماندهی کم و کیف عملکردها و کاربری‌های خدماتی و اداری در زیرپهنه‌های S ۱ و S ۲، ظرف مدت یکسال است.
  ۴. در فرآیند تبدیل اراضی کشاورزی به باغ، اگر فاقد کاربری مصوب خدمات و فضای سبز عمومی باشد، طبق مفاد دستورالعمل اصلاحی ماده ۱۴ قانون زمین شهری، بلامانع می‌باشد.
  ۵. مالکینی که قصد ساختمانسازی ندارند، می‌توانند از ضوابط تفکیک برای اراضی مزروعی استفاده نمایند.
- با توجه به این اطلاعات، اگر پرونده شما تعلیق شده و می‌خواهید بررسی و تصمیم‌گیری درباره آن ادامه پیدا کند، باید از این موارد در نظر گرفته شود:

- ممکن است نیاز به تجدیدنظر در قوانین و مقررات فعلی وجود داشته باشد.
- ممکن است نیاز به تطبیق با قوانین و مقررات جدید وجود داشته باشد.

:Answer\_without\_RAG

با عرض سلام،

در مورد سوال شما، باید توجه داشت که فرآیند تخریب و نوسازی در شهرها و مناطق مختلف متفاوت است. بنابراین، برای پاسخ به سؤال شما، لازم است که اطلاعات دقیق تری درباره وضعیت فعلی پرونده شما در شهرداری و شورای شهر محل زندگی شما در اختیار داشته باشیم.

اما در کل، اگر پرونده شما برای تخریب و نوسازی در حال حاضر در مرحله انتظار است، امیدوارم که در نهایت تصمیم به برگرداندن به روال قبلی اتخاذ شود. اما این موضوع بستگی به قوانین و مقررات محلی، اهمیت پروژه و نیز تصمیمات شخصی افراد مسئول دارد.

به عنوان یک نکته مهم، در این نوع مسائل، از طریق مشاوره با مراجع حقوقی محلی و یا با مشاوره مستقیم با مسئولان شهرداری می‌توانید بهترین راه حل را برای خود پیدا کنید.

با تشکر از پرسیدن سوال شما، امیدوارم که این پاسخ به شما کمک کند.

با سپاس،

[نام شما]

## - موانع و چالش‌ها

- ۱- مورد اول برای استفاده از مدل LLaMa بود. این مدل محدودیت استفاده دارد و سازندگان آن باید دسترسی استفاده از مدل را بدهند که برای ما این درخواست رد شد.
- ۲- محدودیت GPU: تعداد و سوالات ۱۳۹ عدد بود و ران کردن این میزان سوال به میزان زیادی GPU نیاز داشت که با هماهنگی با TA محترم این میزان به ۲۰ عدد کاهش یافت.
- ۳- متون داده‌ها: تعدادی از متن‌ها از حالت عادی خارج بودند. برای مثال یک حرف از آن کلمه بیش از حد کشیده شده بود و تحلیل آن را برای مدل دشوار می‌ساخت که در آخر این چالش را با chatGPT برطرف ساختیم.

## - آموخته‌ها و پیشنهادات

ترکیب مدل Text Generation با RAG و LangChain تجربه‌ای بود که پیش از این نداشتیم. در واقع قبلاً اینکار را با Embedding و Cosine Similarity انجام داده بودیم ولی به مانند این پروژه نبود. پیشنهاد نیز اگر اینکار را با چند مدل زبانی انجام دهیم و آنها را ترکیب یا نتایج آنها را با هم مقایسه کنیم می‌توان به موارد جالبی رسید.

## - منابع

<https://python.langchain.com/v0.2/docs/tutorials/rag/>

ChatGPT