گزارش پیاده‌سازی RAG ساده روی داده‌های خبری فارسی

محمد سینا محلاتی

۳۰ آذر ۱۴۰۴

چکیده

در این پروژه یک نسخه‌ی ساده از سامانه‌ی Retrieval-Augmented Generation (RAG) برای پاسخ‌گویی به پرسش‌های فارسی روی داده‌های خبری فارسی پیاده‌سازی شد. روند کار شامل آماده‌سازی داده، چانک‌کردن متن‌ها، ساخت بردارهای معنایی و لغوی، ایندکس‌سازی با FAISS و در نهایت تولید پاسخ با یک مدل زبانی بود. ارزیابی به‌صورت دستی انجام شد و نتایج نشان داد روش‌های بازیابی معنایی و ترکیبی نسبت به روش صرفاً لغوی خروجی بهتری دارند.

۱ مقدمه

ایده‌ی RAG این است که مدل زبانی به‌جای تکیه‌ی کامل بر دانش درونی خود، ابتدا چند قطعه متن مرتبط را از یک پایگاه بازیابی می‌کند و سپس پاسخ را با تکیه بر همان شواهد تولید می‌کند. این رویکرد به‌خصوص برای سوال‌هایی که پاسخ «دقیق» دارند (عدد، نام، بازه، محل و ...) مفید است و احتمال پاسخ‌های حدسی را کمتر می‌کند.

هدف این پروژه، ساخت یک خط پایه‌ی ساده و قابل فهم از RAG برای اخبار فارسی بود؛ به‌طوری‌که اجزای اصلی سیستم (پیش‌پردازش، بازیابی و تولید پاسخ) شفاف و قابل تغییر باشند.

۲ آماده‌سازی داده

۱.۲ منبع داده

داده‌ی اولیه به صورت فایل CSV از مجموعه‌داده‌ی خبری فارسی تهیه شد. سپس با pandas بارگذاری شد و برای اجرای سریع‌تر، یک زیرمجموعه‌ی کوچک‌تر در فایل news_subset.csv ذخیره شد.

۲.۲ پاک‌سازی و نرمال‌سازی

برای افزایش کیفیت متن‌ها، چند مرحله‌ی ساده انجام شد:

- حذف ردیف‌هایی که متن خبر خالی بود یا طول خیلی کمی داشت (حداقل ۵۰ کاراکتر).
- نرمال‌سازی ساده‌ی برخی حروف عربی/فارسی (مثل تبدیل ي به ی و ك به ک).

۳ ساخت خط پایه‌ی RAG

۱.۳ چانک‌کردن متن

هر خبر برای بازیابی دقیق‌تر به قطعه‌های کوچک‌تر تقسیم شد. تنظیمات چانک‌کردن:

- اندازه‌ی چانک: 800 کاراکتر

- همپوشانی: 150 کاراکتر

خروجی چانک‌ها در قالب جدولی با ستون‌های chunk_id, doc_id, text, category, date ذخیره شد تا هم رهگیری سند اصلی ممکن باشد و هم تحلیل نتایج بازیابی ساده‌تر شود.

۲.۳ مدل‌های Embedding

برای بازیابی سه نوع نمایش برداری تولید شد: معنایی، لغوی و ترکیبی.

- **بردار معنایی (Semantic Embedding):** برای نمایش معنایی چانک‌ها از مدل sentence-transformers/paraphrase-multilingual-mpnet-base-v2 استفاده شد. بردارها به نوع float32 تبدیل شدند و سپس با نرمال‌سازی L2 آماده‌ی استفاده در بازیابی شدند.
- **بردار لغوی (Lexical Embedding):** ابتدا از TF-IDF با ngramهای ۱ و ۲ استفاده شد. سپس برای تبدیل بردارهای sparse به بردار چگال و قابل ایندکس شدن، کاهش‌بُعد با TruncatedSVD انجام شد (بُعد نهایی 256).
- **بردار ترکیبی (Hybrid):** بردار معنایی و بردار لغوی (بعد از نرمال‌سازی L2) با هم concatenate شدند و در انتها بردار ترکیبی نیز نرمال‌سازی شد تا هم‌زمان اطلاعات معنایی و لغوی وارد فرآیند بازیابی شود.

۳.۳ ایندکس‌سازی با FAISS

برای هر یک از سه حالت بازیابی (semantic، lexical، hybrid) یک ایندکس مجزا با FAISS IndexFlatIP ساخته شد. چون بردارها L2-normalize شده بودند، شباهت Inner Product عملاً معادل شباهت Cosine Similarity عمل می‌کرد.

۴.۳ مدل زبانی (LLM) و تولید پاسخ

برای تولید پاسخ نهایی از یک مدل دستورمحور استفاده شد. مدل به‌کاررفته:

Qwen/Qwen2.5-3B-Instruct

پرامپت طوری طراحی شد که پاسخ حتماً فارسی باشد و تا حد ممکن بر پایه‌ی متن‌های بازیابی‌شده نوشته شود. همچنین اگر شواهد کافی در متن‌های بازیابی‌شده وجود نداشت، مدل باید دقیقاً جمله‌ی «اطلاعات کافی در متن‌های بازیابی‌شده نیست.» را برگرداند تا از حدس‌زدن جلوگیری شود.

۴ ارزیابی

۱.۴ سوال‌ها و روش ارزیابی

یک فایل JSONL شامل ۱۵ سوال تهیه شد. هر خط شامل چهار فیلد: id, question, reference_answer, doc_ids.
در این پروژه doc_ids به شناسه‌ی سند (ستون id در news_subset.csv) اشاره دارد، نه شناسه‌ی چانک.
برای هر سوال، سیستم در سه حالت semantic، lexical و hybrid اجرا شد و خروجی‌ها به‌صورت دستی برچسب «قابل قبول/غیرقابل قبول» گرفتند.

۲.۴ نتایج کمی

نتایج به تفکیک روش بازیابی در جدول ۱ آمده است.
به طور کلی، روش‌های معنایی و ترکیبی عملکرد نزدیک به هم و بهتر از روش لغوی داشتند و نرخ کلی پاسخ‌های قابل قبول برابر 44.44% شد.

روش بازیابی	نرخ پاسخ قابل قبول
hybrid	53.33%
lexical	26.67%
semantic	53.33%
کل	44.44%

جدول ۱: نرخ پاسخ‌های قابل قبول در ارزیابی دستی

۵ تحلیل کیفی

در نمونه‌های موفق، معمولاً متن مرتبط در نتایج top-k بازیابی شده و مدل توانسته پاسخ دقیق‌تری تولید کند (به‌خصوص در سوال‌های عددی و عبارتی). در مقابل، در نمونه‌های ناموفق معمولاً مشکل از مرحله‌ی بازیابی بوده است؛ یعنی چانک مرتبط وارد مجموعه‌ی بازیابی‌شده نشده و مدل هم طبق پرامپت از حدس‌زدن خودداری کرده و «اطلاعات کافی نیست» برگردانده است. بنابراین کیفیت خروجی، وابستگی زیادی به کیفیت retrieval و چانک‌کردن دارد.

۶ محدودیت‌ها و پیشنهادهای بهبود

- **چانک‌کردن مبتنی بر کاراکتر:** ممکن است جمله‌ی کلیدی یا عدد مهم در مرز دو چانک قرار بگیرد. چانک‌بندی مبتنی بر جمله/پاراگراف با overlap مناسب می‌تواند بهتر باشد.
- **تنظیمات بازیابی:** افزایش top_k (مثلاً از ۵ به ۱۰) برای سوال‌های دقیق می‌تواند شانس رسیدن به متن درست را بالا ببرد.
- **ترکیب بهتر لغوی/معنایی:** به‌جای الحاق ساده‌ی بردارها، می‌توان روش‌های ترکیب رتبه مثل rank fusion را آزمایش کرد.
- **بهبود تولید پاسخ:** مدل بزرگ‌تر یا تنظیم دقیق‌تر تولید می‌تواند پاسخ‌های ناقص/کلی را کمتر کند، البته به شرط اینکه همچنان پاسخ به شواهد محدود بماند.

۷ جمع‌بندی

در این پروژه یک خط پایه‌ی ساده از RAG روی اخبار فارسی ساخته شد. در بخش embedding از مدل sentence-transformers/paraphrase-multilingual-mpnet-base-v2 برای نمایش معنایی و از TF-IDF + TruncatedSVD برای نمایش لغوی استفاده شد و یک حالت ترکیبی نیز ساخته شد. تولید پاسخ با مدل Qwen/Qwen2.5-3B-Instruct انجام گرفت. ارزیابی دستی نشان داد روش‌های semantic و hybrid با نرخ 53.33% بهتر از روش صرفاً lexical عمل می‌کنند و نرخ کلی پاسخ‌های قابل قبول برابر 44.44% بوده است.