بسم‌تعالی

گزارش تمرین اول درس پردازش زبان طبیعی
گروه ۹
اعضای گروه: علیرضا حسین‌خانی، صالح شجاعی و محمدجواد سعیدی

بخش اول - استخراج دادگان

در این تمرین به سراغ استخراج دادگان مرتبط با غذاهای استان‌های گروه ۳ رفتیم که شامل ۶ استان گیلان، آذربایجان غربی و شرقی، اردبیل، کردستان و زنجان می‌شود.
پس از بررسی سایت‌ها و منابع موجود اقدام به خزش کردن داده‌های ۳ سایت ،https://fa.wikibooks.org https://ghazaland.com/ و https://roostanet.ir/ نمودیم و برای هرسایت خزش‌گر مربوطه‌ش را نوشتیم.
علت انتخاب این سایت‌ها هم اطلاعات کامل و با جزئیاتشان بود تا بتوانیم مواد مورد و طرز تهیه غنی‌ای را خزش کرده‌باشیم.
در فایل‌های پیوست این گزارش ۳ فایل مربوط به خزشگر به نام‌های:

crawler_roostanet.py
crawler_ghazaland.py
crawler_wiki_book.py

مشاهده می‌کنید که شامل کدهای خزشگرهای این ۳ سایت است.

باتوجه به اینکه فرمت نمایش هرکدام از غذاها و جزئیاتی که بیان می‌کردند در هر یک از سایت‌های خزش‌شده متفاوت بود و ما نیز می‌خواستیم به فرمت خواسته شده در صورت تمرین غذاها را ذخیره‌سازی نیاز به یک عملیات transformation میان jsonهای ذخیره شده از هر سایت و قالب مدنظر داشتیم که برای انجام این کار از مدل‌های زبانی بزرگ کمک گرفتیم.
باتوجه به اینکه چند صد غذای خزش شده داشتیم و استفاده از UIهای LLMهای بزرگ دشواری‌هایی داشت اقدام به استفاده از APIهای مدل‌های زبانی بزرگ کردیم.
با استفاده از https://metisai.ir/ که به امکان استفاده از APIهای مدل‌های زبانی بزرگ را می‌داد یک توکن خریداری کردیم و برای ساختاردهی داده‌های خام از آن استفاده کردیم که کد آن را در فایل:

llm.py

مشاهده می‌کنید که ابتدا از پکیج openai استفاده کرده و یک کلاینت نوشتیم و سپس با کمک پرامپتی که به LLM خواسته‌ها و محدودیت‌های ما را منتقل می‌کرد به LLMها درخواست زدیم و jsonهای خروجی را ذخیره کردیم که در بخش outputs می‌توانید به تفکیک هر استان خروجی‌های موردنظر را مشاهده کنید.
همچنین مواردی نظیر meal_type و occasion در سایت‌هایی که خزش کرده بودیم مشخص نشده‌بود و این فیلدها را نیز به کمک LLM و پرامپتی که به آن داده بودیم مشخص کردیم.

در پایان این بخش حدود ۳۱۳ غذای مربوطه استخراج و جمع‌آوری و تمیز شد که به تفکیک هر استان در زیر مشاهده می‌کنید:
گیلان: ۱۰۱ غذا
زنجان: ۲۲ غذا
کردستان: ۴۵ غذا
اردبیل: ۳۵ غذا
آذربایجان شرقی: ۸۰ غذا
آذربایجان غربی: ۳۰ غذا

بخش دوم - برچسب‌گذاری دادگان