هدف پروژه

هدف این پروژه، طراحی و پیاده‌سازی سیستمی است که بتواند با استفاده از تکنیک‌های پردازش زبان طبیعی (NLP)، به شناسایی شباهت‌ها و برچسب‌گذاری خودکار در داده‌های متنی بپردازد.

دیتاست 60k Stack Overflow Questions

این دیتاست شامل اطلاعات مربوط به 60,000 سؤال از Stack Overflow است. این داده‌ها به دو مجموعه تقسیم شده‌اند:

- **Training Set:** برای آموزش مدل
- **Validation Set:** برای ارزیابی عملکرد مدل

در این پروژه، ستون‌های کلیدی دیتاست عبارتند از:

- **Title:** عنوان سؤال.
- **Body:** متن کامل سؤال که ممکن است شامل تگ‌های HTML و محتوای اضافی باشد.
- **Tags:** برچسب‌هایی که موضوع سؤال را مشخص می‌کنند.

فاز اول: Preprocessing

پیش‌پردازش داده‌ها یکی از حساس‌ترین و مهم‌ترین مراحل این پروژه است. در این مرحله، هدف اصلی آماده‌سازی متن‌ها برای تحلیل و مدل‌سازی است. شما باید مطمئن شوید که داده‌ها عاری از نویز و به شکلی مناسب برای استفاده در مدل‌های پردازش زبان طبیعی (NLP) باشند.

- متن‌ها ممکن است شامل محتوای اضافی و نویزهایی مانند تگ‌های خاص یا ساختارهای غیرضروری باشند که لازم است به دقت مدیریت و پاک‌سازی شوند.

- توجه در انجام این مرحله از پروژه تاثیر مستقیمی بر کیفیت خروجی شما خواهد داشت. در پیاده‌سازی آن دقت کافی را به خرج دهید.
- از تکنیک‌ها و روش‌های مطرح شده در کلاس درس و کلاس حل تمرین برای انجام این مرحله استفاده کنید. انتخاب این روش‌ها بر عهده شماست.

فاز دوم: Word2Vec & Similarity Retrieval

در این فاز، مدل بازیابی اطلاعات خود را پیاده‌سازی کنید. هدف، یافتن سؤالات مشابه از دیتاست پیش‌پردازش‌شده، برای یک پرسش ورودی (Query) است.

مراحل:

۱. **ایجاد بردارهای معنایی:**

- از مدل Word2Vec برای تولید بردارهای کلمه‌ای (Word Embeddings) استفاده کنید.

۲. **تحلیل بردارها در فضای سه‌بعدی:**

- بردارهای بدست آمده برای چندین نمونه تصادفی از کلمات (حداقل 100 کلمه) را در فضای سه‌بعدی ویژوالایز کنید.
- برای هر دیتاپوینت در این نمودار، متن خود کلمه را به عنوان label آن دیتاپوینت کنار آن قرار دهید تا تحلیل نمودارتان واضح تر شود.
- این کار را برای بردار داکیومنت ها(هر سطر از دیتاست) نیز انجام دهید.
- تحلیل و بررسی این دو نمودار در داکیومنت ارائه شود.

۳. **محاسبه شباهت:**

- از معیار Cosine Similarity برای محاسبه میزان شباهت میان بردارها استفاده کنید.

۴. **بازیابی سؤالات مشابه:**

- با دریافت یک query، نزدیک‌ترین سؤالات موجود در دیتاست به query دریافتی را پیدا کنید.

فاز سوم: Tagging

در این فاز، روی بخشی از دیتای validation ابتدا پیش‌پردازش‌های فاز اول اعمال می‌شود، سپس با استفاده از مدل KNN که روی داده‌های train آموزش دیده است، نزدیک‌ترین همسایه‌ها پیدا شده و تگ‌های پیشنهادی با متن اصلی شباهت‌سنجی می‌شوند؛ در نهایت تگ‌های تخصیص داده شده با تگ‌های واقعی مقایسه و دقت ارزیابی می‌شود. مراحل این فاز به شکل زیر است:

۱. **آماده‌سازی دیتای validation:**

- همانند فاز اول، تمام مراحل پیش‌پردازش برای سطرهای انتخابی از دیتای validation اجرا شود.
- به‌جای استفاده از کل دیتای validation می‌توان یک زیرمجموعه تصادفی (مثلاً 10% از داده‌ها) را انتخاب کرد.
- این زیرمجموعه باید به اندازه کافی متنوع باشد تا کارایی الگوریتم را به‌خوبی نشان دهد.

۲. **یافتن k همسایه نزدیک با KNN:**

- برای هر سطر از دیتای validation:
  - با استفاده از همان مدل word embedding که در فاز دوم استفاده شد، تبدیل متن به بردار (embedding) انجام می‌شود.
  - سپس این بردار به مدل KNN داده می‌شود تا k همسایه نزدیک آن از دیتای آموزشی (train) پیدا شود.

۳. **استخراج تگ‌ها از همسایه‌های نزدیک**

۴. **ارزیابی تگ‌های پیش‌بینی شده:**

- تگ‌های پیش‌بینی شده برای هر نمونه در دیتای validation با تگ‌های اصلی (موجود در ستون Tags) مقایسه شوند.
- اگر تگ‌های پیش‌بینی شده با لیست تگ‌های اصلی اشتراک داشته باشند، آن نمونه به‌عنوان "تگ‌گذاری درست" در نظر گرفته شود.
- با توجه به تگ های پیش‌بینی شده accuracy الگوریتم tagging تان را محاسبه کنید و گزارش دهید.

۵. **نمایش چند نمونه:**

- برای نشان دادن عملکرد الگوریتم، چند نمونه از پیش‌بینی‌های موفق و ناموفق به همراه توضیحات داخل داکیومنت ارائه شود.

نکات تکمیلی

- علاوه بر سورس کد پروژه، فایل مستندات نیز باید آپلود شود.
- نام اعضای گروه در فایل مستندات ذکر شود و فقط یکی از اعضا پروژه را آپلود کند.
- هر گونه شباهت نامتعارف بین کد شما و کد سایر گروه ها تقلب محسوب مي شود و نمره ای برای این پروژه دریافت نخواهید کرد.
- در صورت نوشتن داکیومنت تمیز (برای مثال با LATEX) نمره اضافه برای شما در نظر گرفته خواهد شد.
- فایل شامل سورس کد پروژه و مستندات را در قالب فایل zip و با نام شماره دانشجویی خود ذخیره و ارسال نمایید.
- در صورت داشتن هرگونه سوال می توانید با SoroushPasandideh0 و fatemeh_dehbashii در ارتباط باشید یا در گروه درسی مطرح کنید.

موفق باشید؛
تیم حل تمرین