پردازش زبان طبیعی

فاز اول پروژه

فاطمه قلی‌زاده - ۹۷۵۲۱۴۴۱

بهار ۱۴۰۱

داده‌های جمع‌آوری شده

برای انجام این پروژه دو مجموعه داده جمع‌آوری شده‌است:

۱. مجموعه زوج سوالات مشابه: این مجموعه شامل بیش از ۴۰۰ هزار زوج سوال است که با توجه به تشابه یا عدم تشابه به یکدیگر، برچسب زده شده‌اند. این مجموعه داده از طرف Quora آماده شده است و در دسترس عموم قرار دارد. برای همین صرفا دانلود می‌شود.
۲. سوالات استخراج شده از سایت Quora: با استفاده از ابزار selenium و beautifulsoup، تعدادی سوال از موضوعات معینی از سایت Quora استخراج شده است. این مجموعه شامل متن سوالات و لینک آن ها است و برچسبی ندارد. شیوه استخراج به این شکل است که ابتدا لیست معینی از لینک‌های تعدادی از موضوعات سایت quora در اختیار خزشگر قرار می‌گیرد. خزشگر ابتدا از هر صفحه تا حد امکان سوال جمع‌آوری می‌کند. سپس به طور بازگشتی وارد صفحه هر سوال شده و سوالات مرتبط با آن را نیز استخراج می‌کند.

فرمت ذخیره داده ها به این شکل است که ابتدا فایل خام داده‌ها در مسیر data/raw ذخیره می‌شود. داده‌های زوج سوالات شامل یک فایل tsv است و داده‌های استخراج شده، یک فایل json است که شامل لیستی از سوالات به همراه لینک آن‌ها است (برای هر سوال، مشخص شده است که از چه لینکی به آن رسیده‌ایم).

پیش‌پردازش

برای هر کدام از داده‌ها پیش‌پردازش در ۵ مرحله زیر انجام شده‌است:

۱. تبدیل حروف بزرگ به حروف کوچک: با استفاده از متد lower این کار را انجام می‌دهیم.
۲. حذف داده‌های تکراری، در مجموعه استخراج‌شده جملات تکراری حذف می‌شوند و در مجموعه دانلودشده، زوج سوالات تکراری حذف می‌شوند. برای سادگی ابتدا هر مجموعه داده به DataFrame تبدیل می‌شود و با استفاده از متد drop_duplicate این کار انجام می‌شود.
۳. حذف علائم نگارشی و نشانه‌های خاص: با استفاده از ماژول re هر چیزی به جز حروف الفبای انگلیسی حذف می‌شود.
۴. تفکیک کلمات و جملات: با استفاده از ابزار nltk و توابع word_tokenize و sent_tokeize متون خام به جملات و کلمات تفکیک می‌شود.
۵. حذف ایست‌واژه: با استفاده از nltk این کار انجام می‌شود.

خروجی هر مرحله از مراحل بالا به ترتیب در مسیرهای زیر ذخیره می‌شوند:

1. تبدیل حروف بزرگ به کوچک: data/lower
۲. حذف داده‌های تکراری: data/no_duplicate
۳. حذف علائم نگارشی و نشانه‌های خاص: data/no_special
۴. تفکیک کلمات و جملات: data/tokenize
۵. حذف ایست‌واژه: data/no_stopwords

ذخیره داده‌ها به صورت تجمیعی است؛ یعنی هر مرحله از پیش‌پردازش خروجی همه مراحل پیش از خود را داراست. آرگمان‌هایی برای انجام هر مرحله تعریف شده‌اند تا در صورت نیاز بتوان هر مرحله را جداگانه و مستقل از سایر مراحل انجام داد. توضیحات بیشتر در README.md داده شده‌است.

آمار

در حین مراحل پیش‌پردازش و نیز تحلیل داده‌های جمع‌آوری‌شده، موارد زیر تهیه می‌شوند:

۱. تعداد داده‌های خام: در ابتدای پیش‌پردازش تعداد داده‌های خام محاسبه و نوشته می‌شود.
۲. تعداد جملات: در حین اجرای کد تحلیل تعداد جملات بدست می‌آید.
۳. تعداد کلمات: در حین اجرای کد تحلیل تعداد کلمات بدست می‌آید.
۴. تعداد کلمات منحصربه‌فرد: در حین اجرای کد تحلیل تعداد کلمات منحصربه‌فرد بدست می‌آید.
۵. هیستوگرام کلمات پرتکرار: پس از شمارش تعداد کلمات، نمودار میله‌ای کلمات پرتکرار در هر مجموعه داده ساخته می‌شود و در مسیر data/word_count_plots ذخیره می‌شود.