به نام خدا

دوره کارآموزی هوش مصنوعی
پردازش زبان طبیعی

بخش اول پروژه - جمع آوری داده

محدثه رهنما - زینب تقوی

آبان الی دی ۹۹

جمع آوری و برچسب‌گذاری اطلاعات

به منظور جمع‌آوری داده از وب، مسئله به دو زیر مجموعه تقسیم شده‌است: جمع آوری اطلاعات و برچسب زدن (با توجه به این نکته که امکان دارد برخی از مطالب چندین برچسب داشته باشند)
مرحله ی اول : برای انجام این کار از موتور های جستجو استفاده شده است. به این معنا که به ازای هر برچسب موجود در فایل xlsx.Tag ،در سایت های مورد نظر و مرتبط با موضوع جستجو شده و لینکهای به دست آمده ذخیره سازی میشود. این کار با تعریف یک پروژه ی Scrapy و ذخیره ی اطلاعات(متن فارسی - عنوان - تگ) در فایل csv .انجام شد. برای سایت‌های متفاوت از یک چهارچوب ثابت کد استفاده شده تا برای افزودن سایت جدید تنها نیاز به اعمال تغییرات جزیی باشد و به راحتی بتوان سایت‌های بیشتری را مورد جستجو قرار داد.

مرحله‌ی دوم : فایل‌های csv . بدست آمده از هر سایت وارد مرحلهی دوم می‌شوند. در این مرحله لینک‌های واحدی که در جستجوها با برچسبهای مختلف ذخیره شده اند شناسایی میشود. برای نمونه ممکن است در وب سایتی، مطلبی با برچسب «پایتون» و «طراحی وب» وجود داشته باشد. به همین دلیل لینک این مطلب در هنگام جستجوی با دو برچسب مذکور استخراج شده و دو بار در فایل csv .ذخیره میگردد. بنابراین هدف از مرحله‌ی دوم شناسایی لینک‌های تکراری با برچسب‌های متفاوت و ادغام برچسب آنهاست. در مرحله با استفاده از قابلیت‌های کتابخانه ی pandas میتوان تگ‌هایی که عنوان یکسانی دارند را یافت و آنها را به صورت لیست در آورد و به صورت رشته‌ای به عنوان تگ جدید ذخیره کرد. لازم به ذکر است که تگ‌ها به آسانی قابل تبدیل مجدد به لیست هستند و نگه داری و بازیابی آن‌ها ساده‌تر است. پس از آن سایر ردیف‌ها با عنوان یکسان وغیرضروری را پاک می‌کنیم. با افزودن و ذخیره هریک از این dataframe ها به یک dataframe جامع‌تر, در نهایت یک مجموعه داده‌ی آماده خواهیم داشت. برای ایجاد ساده‌تر پروژها کد py.spider_make نوشته شده است. این کد به صورت خودکار برای هر سایت منتخب، یک پروژه‌ی scrapy ساخته و generate میکند. سپس اطلاعات مرتبط با scrapy agent-user- را در بخش py.setting آن اضافه میکند. برای این که تمام این کارها به صورت واحد انجام شود، اسکریپت main آماده شده است(و برای کارایی بهتر از کتابخانه‌ی argparse استفاده کردیم). پس از اطمینان از وجود پروژهی scrapy برای هر سایت منتخب، به ترتیب آنها را اجرا کرده و بخش اول مرتبط به spider کردن و سپس ذخیرهی script ها می‌پردازد. سپس تابع dataset_make اجرا شده که در واقع همان بخش دوم مسئله، یعنی تعیین نهایی تگها و جمعآوری آنها در یک فایل csv .را انجام می‌دهد.(برای مرتب سازی رشته‌ها از امکانات کتابخانه‌ی re استفاده کردیم)

سایتهای منتخب:

- virgool.io
- sokanacademy.com
- learn.com
- roocket.ir
- zerotohero.ir

پیش پردازش داده‌ها

در این مرحله لازم است تا متون جمع‌آوری شده مورد پردازش قرار گیرند، داده‌های اضافی و ناخواسته حذف گردند و متن برای مراحل بعدی آماده شود. بدین منظور داده‌های ذخیره شده در فایل dataset.csv خوانده شده و برای هر متن مراحل زیر به ترتیب اجرا می‌گردند. پیش پردازش با کتابخانه‌های هضم و پارسیوار انجام شده است.

۱. خطاهای املایی احتمالی تصحیح شده‌اند. (کتابخانه ی پارسیوار)

۲. صورتک (ایموجی) های احتمالی در متن شناسایی و حذف می‌گردند. این کار با یونی‌کدهای تعریف شده برای ایموجی‌ها انجام پذیر است.

۳. نیم فاصله‌های احتمالی در متن حذف شده و با فاصله جایگزین می‌شوند. نیم فاصله‌ها در تشخیص مرز کلمات چالش ایجاد می‌کنند. گرچه که در مراحل بعدی و شناسایی توکن‌ها، ابزار پیش پردازش در مواقع لزوم نیم فاصله‌ها را اضافه می‌کند که این مورد مشکلی در پردازش ایجاد نمی‌کند.

۴. یکسان‌سازی (نرمال سازی) برای یک دست کردن کارکترها و یونی‌کدهای عربی احتمالی یا کارکترهای اضافی ضروری است. (کتابخانه‌ی هضم)

۵. شناسایی توکن‌ها یا به نوعی کلمات نیز انجام می‌شود و متن به لیستی از کلمات متوالی تبدیل می‌شود.(کتابخانه‌ی هضم)

۶. علایم نگارشی و ایست واژه‌های فارسی از توکن‌ها حذف می‌گردد. این کار با فهرست ایست واژه‌های فارسی صورت می‌گیرد.

۷. در نهایت ریشه‌یابی با هدف یکسان‌سازی کلماتی که با مشتقات مختلف در متن ظاهر می‌شوند انجام شده است. (کتابخانه‌ی پارسیوار)