پردازش زبان‌های طبیعی

نیم‌سال دوم ۱۴۰۲

مدرس: دکتر احسان الدین عسگری

دانشکده علوم ریاضی

تمرین سری دوم

نام و نام‌خانوادگی: علی شفیعی-امیراحمد شفیعی- ندا فلاح

شماره دانشجویی: ۹۸۱۰۴۲۰۶،۹۹۱۰۴۰۲۷،۹۸۱۰۰۲۲۶

در ابتدا کتاب‌خانه‌های لازم را نصب و ایمپورت میکنیم. سپس دیتاست corpus را از روی فایل میخوانیم. این دیتاست شامل لغات فارسی است که از کورپس هضم بدست آمده‌است. سپس ریشه کلمات موجود در دیتاست را به دیتاست اضافه میکنیم، همچنین به انتهای اسم‌ها و صفت‌ها در دیتاست پسوندهایی را اضافه میکنیم زیرا گاها lemmatizer در حذف آنها به خوبی عمل نمیکند.
سپس در قسمت بعدی یک دیکشنری از ابزارهای لازم میسازیم.
سپس در کلاس foreignWordDetector و در تابع detect آن فرایند تشخیص کلمات خارجی را انجام میدهیم. برای اینکار ابتدا جملات متن ورودی را استخراج میکنیم، سپس آن‌ها را نرمالایز میکنیم، با کمک لمتایزر کتابخانه دادما ، ریشه کلمات را بدست می‌آوریم و با آن‌ها جایگزین میکنیم.
از انجایی که فعل‌ها صرف‌های مختلف دارند و لمتایزر گاها برای آن‌ها دچار مشکل میشود و همچنین از آنجایی که دیتاست همه فعل‌ها را ندارد ، با کمک مدل posTagger کتابخانه هضم ، فعل‌ها را تشخیص میدهیم و آن‌ها را از لیست کلمات حذف میکنیم.
سپس برای کلمات بدست آمده بررسی میکنیم که آیا این کلمه در کورپس وجود دارد یا نه، اگر وجود نداشت آن را به عنوان کلمه انگلیسی در نظر میگیریم و به دیکشنری خروجی اضافه میکنیم.
سپس در تابع run یک نمونه از foreignWordDetector میسازیم و با ورودی گرفتن جمله از کاربر کلمات خارجی آن را استخراج میکنیم.
ایده دیگری که این تمرین به سمت آن رفتیم، استفاده از فونتیک کلمات انگلیسی و ساختن فرم فارسی کلمات انگلیسی با استفاده از آن‌هاست.
برای اینکار ابتدا دیکشنری انگلیسی را لود میکنیم و کلمات انگلیسی را در لیست words و فونتیک متناظر آن‌ها را در لیست phonicsList ذخیره میکنیم.
سپس با استفاده از یک دیکشنری که در آن شکل فارسی برای هر فونتیک نوشته شده‌است ، شکل نوشتاری فارسی کلمات را بدست می‌آوریم.
پس از پایان اینکار در لیست words ، کلمات انگلیسی و در لیست ،persianized شکل فارسی نوشتاری آن‌ها قرار گرفته‌است.
در تابع changeEnglishWord یک متن را به عنوان ورودی میگیریم و کلمات انگلیسی آن را به کمک ایdetector که ساختیم پیدا میکنیم و آن‌ها را به کمک کتابخانه googletrans به شکل انگلیسی میبریم و سپس ترجمه آن‌ها را بدست می‌آوریم و به جای آن‌ها جایگزین میکنیم.
یک ایده هم این است که به کمک همان لیست نوشتار فارسی کلمات انگلیسی که آن‌ها را بدست آوردیم ، هر کلمه خارجی پیدا شده را در متن ورودی را در آن لیست پیدا کنیم و شکل انگلیسی آن‌را به این شکل بدست آوریم و سپس آن‌را ترجمه کنیم. البته در جستجو کلمه در لیست persianized باید به این نکته توجه کنیم که گاها بعضی کلمات به درستی به فارسی نوشته نمیشوند برای مثال کلمه computer به صورت کمپیوتر خوانده میشود ولی در فارسی کامپیوتر نوشته میشود، برای رفع این مشکل میتوان از distance edit استفاده کنیم و برای مثال هزینه insert شدن ”آ” را کم در نظر بگیریم زیرا بسیاری از اوقات فونتیک ”شوا” در فارسی به درستی رعایت نمی‌شود و به جای آن ”آ” نوشته میشود.
با اینجور ایده‌ها میتوان شکل انگلیسی کلمات لیست persianized را بدست آورد، که از آنجا که کتابخانه googletrans به خوبی نتیجه میدهد و همچنین کمبود زمان، اینکار را انجام ندادیم.

گاهی ممکن است برخی کلمات هم در زبان فارسی وجود داشته باشند هم در زبان انگلیسی. برای اینکه فرم درست کلمات

را تشخیص دهیم، یک بار کلمه را فارسی در نظر میگیریم و بار دیگر کلمه را انگلیسی در نظر میگیریم و معادل فارسی اش را به کمک روش گفته شده جایگزین میکنیم. حال به کمک cosinesimilarity و با مدل زبانی هضم، میبینم که کدام جمله محتمل تر است.