به نام خداوند عید آفرین خداوند نوروز و آیین و دین

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

مبانی پردازش زبان و گفتار

استاد: دکتر حمیدرضا برداران کاشانی

دستیاران آموزشی:

مهرداد قصابی

علی مامن پوش

محمدامین مولوی زاده

زمستان ۱۴۰۳ - بهار ۱۴۰۴

بخش اول: پرسش‌ها

۱- علائم نقطه گذاری را چه هنگام باید به عنوان توکن مجزا در نظر بگیریم و در چه هنگام باید آن‌ها را حذف کنیم؟

۲- مزایا و معایب توکن بندی مبتنی بر کلمات و توکن بندی مبتنی بر کاراکتر را نام ببرید؟

۳- به طور کامل توضیح دهید در صورتی که فقط از توکن بندی مبتنی بر فاصله استفاده کنیم چه مشکلاتی ممکن است به وجود آید؟

۴- اگر کلمه‌ای مانند watched را به واحدهای زیر کلمه‌ای توکنایز نکنیم چه مشکلاتی پیش می‌آید؟

۵- درباره‌ی روش‌های تعبیه سازی کلمات(word embedding) Glove و Elmo تحقیق کنید و تفاوت‌های آن‌ها را شرح دهید؟

۶- دلیل حذف ایست واژه‌ها(stop words) در پیش پردازش متون چیست؟

بخش دوم: پیش پردازش

قدم اول در وظایف NLP پیش پردازش متن است. در این تمرین با دو مجموعه داده فارسی (hp_fa.txt) و انگلیسی (hp_en.txt) کار می‌کنیم. برای مجموعه داده فارسی از کتابخانه Hazm و برای مجموعه داده انگلیسی از کتابخانه nltk استفاده کنید.

۱- مراحل زیر را بر روی مجموعه داده فارسی اعمال کنید.

۱-۱- فضاهای خالی اضافه را حذف کنید، متن را به جملات آن تجزیه کنید و سپس متن را normalize کنید.

۱-۲- جملات را به کلمات آن توکن‌بندی کنید.

۱-۳- علایم نگارشی را حذف کنید.

۱-۴- ایست واژه‌ها را از درون متن حذف نمایید.

۱-۵- ایموجی‌های موجود در متن را حذف کنید.

۱-۶- فرآیند لم‌سازی را بر روی متن اعمال کنید.

۲- مراحل زیر را بر روی مجموعه داده انگلیسی اعمال کنید.

۲-۱- فضاهای خالی اضافه را حذف کنید و متن را به جملات آن تجزیه کنید.

۲-۲- حروف بزرگ را به حروف کوچک تبدیل کنید.

۲-۳- جملات را به کلمات آن توکن‌بندی کنید.

۲-۴- اعداد و URL ها را حذف کنید.

۲-۵- علایم نگارشی و ایست واژه‌ها را حذف کنید.

۲-۶- ابر کلمات(Wordcloud) را برای متن پیش پردازش شده رسم کنید.

بخش سوم: سیستم تبدیل متن به اعداد

هدف از این بخش از تمرین، طراحی یک سیستم تبدیل متن به عدد است در این بخش باید شما سیستمی طراحی کنید که شما در آن محدوده اعداد صفر تا یک میلیارد را به حروف در آن وارد کنید و در خروجی با توجه به حروف وارد شده به شما عدد مورد نظر به شما نمایش داده شود.(برای مثال در صورت وارد شدن سه میلیون و سیصد و پنجاه و سه باید در خروجی ۳۰۰۳۵۳ نمایش داده شود.)

بخش چهارم: تصحیح خطاهای املایی

هدف از این بخش پیاده‌سازی الگوریتمی برای تصحیح خطاهای املایی است. این الگوریتم یک جمله را به عنوان ورودی دریافت می‌کند و بهترین پیشنهادها را برای هر کلمه غلط املایی به عنوان خروجی ارائه می‌دهد. برای این بخش از فایل Vocabulary.txt به عنوان دایره لغات استفاده کنید.

مراحل انجام کار:

انتخاب جمله: اولین جمله از چکیده مقاله انتخابی در بخش دوم تمرین را انتخاب کنید.

ایجاد جمله غلط املایی: جمله انتخاب شده را در این وب سایت کپی و غلط‌های املایی ایجاد کنید.(در صورت باز نشدن سایت از فیلترشکن استفاده کنید)

استفاده از جمله غلط املایی به عنوان ورودی الگوریتم: الگوریتم تصحیح خطا را بر روی جمله غلط املایی اجرا کنید و خروجی آن را نمایش دهید.

بدیهی است که تمامی موارد باید به صورت مشروح در فایل گزارش فنی آورده شود.

بخش پنجم: تشخیص اخبار جعلی

هدف این بخش از تمرین، توسعه یک سیستم هوشمند است که بتواند به صورت خودکار اخبار جعلی را از اخبار واقعی تشخیص دهد. این سیستم با استفاده از داده‌های متنی آموزش داده می‌شود و می‌تواند به عنوان یک ابزار کمکی برای کاربران، سازمان‌ها و رسانه‌ها در تشخیص اخبار جعلی مورد استفاده قرار گیرد.

۱- با توجه به دوفایل csv قرار داده شد در فایل zip تمرین، شما دو فایل با نام‌های True.csv و False.csv در اختیار دارید که باید تقسیم بندی داده‌های خود را به صورت زیر انجام دهید:

- Train: ۷۰٪

- Validation: ۱۵٪
- Test: ۱۵٪

۲- برای انجام پیش پردازش، مراحل ۲-۱ تا ۲-۵ ذکر شده در بخش دوم تمرین را در این‌جا نیز اعمال کنید.

۳- با استفاده از texts_to_sequences در کتابخانه‌ی Keras (که بخشی از TensorFlow است) برای تبدیل متن به دنباله‌های عددی استفاده کنید و سپس از Padding استفاده کنید تا همه‌ی دنباله‌ها طول یکسانی داشته باشند.

۴- مدل KNN را روی داده‌های Train آموزش دهید.

۵- مدل را روی داده‌های Validation ارزیابی کنید و معیارهای عملکرد مانند دقت، Precision، Recalو F۱-Score را محاسبه کنید.

۶- در صورت نیاز، پارامترهای مدل را تنظیم کنید.

۷- در نهایت، مدل را روی داده‌های Test ارزیابی کنید تا عملکرد نهایی آن مشخص شود.

۸- آموزش و ارزیابی مدل SVM:

- بار دیگر، مراحل ۴ تا ۸ را با استفاده از مدل SVM تکرار کنید.
- نتایج حاصل از KNN و SVM را مقایسه و تحلیل کنید.
- دلایل تفاوت عملکرد مدل‌ها را بررسی کنید.

خروجی مورد انتظار:

جدول مقایسه‌ای معیارهای عملکرد (Accuracy, Precision, Recall, F۱-Score) برای مدل‌های KNN و SVM

تحلیل نتایج و ارائه پیشنهادات برای بهبود مدل‌ها.

نکات تحویل

۱- پاسخ خود را در پوشه ای به اسم NLP_NAME_FAMILY_HW1 و در قالب zip بارگذاری نمایید.

۲- این پوشه باید حاوی موارد زیر باشد:

- کد نوشته شده در قالب یک فایل jupyter notebook
- فایل گزارش فنی در قالب یک فایل PDF

۳- لازم به ذکر است که رعایت قوانین نگارشی حائز اهمیت است.