به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس پردازش زبان‌های طبیعی

تمرین ۱

اسفند ماه ۱۴۰۱

*فهرست

مقدمه

هدف از این تمرین آشنایی شما با مدل های زبانی n-gram و همچنین روش های مختلف Tokenization و مقایسه ی آنها می باشد. توکنایز کردن جملات اولین قدم برای انجام بسیاری از تسک ها در پردازش زبان طبیعی می باشد و با شکستن جملات به واحدهای کوچک تر امکان درک بهتر متن و انجام آنالیزهای بیشتر را فراهم می کند. مدل های زبانی n-gram نیز از جمله مدل های احتمالاتی هستند که سعی دارند به کمک روابط موجود در احتمال، کلمه بعدی را با در نظر گرفتن کلمات قبلی موجود در متن پیش بینی کنند.

در قسمت اول این تمرین لازم است تا شما یک مدل زبانی n-gram را پیاده سازی کنید و در بخش دوم به مقایسه چند Tokenizer بپردازید. برای انجام هر دو بخش نیز می توانید از کتابخانه های موجود استفاده کنید.

مجموعه داده

برای آموزش و ارزیابی روش های مختلف دو مجموعه داده در نظر گرفته شده است که در پوشه data فایل تمرین موجود است. داده ی اول متن کتاب هری پاتر(جلد اول) به زبان فارسی و داده ی دوم همین کتاب به زبان انگلیسی می باشد.

سوال ۱

همان طور که می‌دانید مدل های زبانی سعی دارند که احتمال وقوع دنباله ای از کلمات را پیش بینی کنند. یکی از ساده ترین مدل های زبانی، n-gram ها هستند. بنابراین می توان با در نظر گرفتن یک مجموعه داده چنین مدلی آموزش داد به طوری که بتواند احتمال وقوع یک کلمه را با استفاده از کلمات قبلی محاسبه کرده و به کمک آن یک متن نزدیک به زبان طبیعی تولید کند. در این بخش لازم است با انجام مراحل گفته شده و با داده‌ی مربوط به کتاب هری پاتر (فارسی) یک مدل زبانی را آموزش دهید:

الف) در ابتدا سعی کنید پیش پردازش های لازم مانند نرمالسازی داده، قطعه بندی جملات، اضافه کردن کاراکترهایی برای شروع و پایان جملات و ... را انجام دهید.

ب) پس از جدا سازی کلمات هر جمله bigram(n=2) های موجود در متن را به دست آورده و با به دست آوردن احتمالات مربوط به آنها، مدل زبانی را آموزش دهید.

(برای به دست آوردن احتمالات از روش add-1 Laplace smoothing استفاده کنید)

ج) در صورتی که از روش Laplace smoothing استفاده نشود چه مشکلی پیش می آید؟

د) به کمک مدل زبانی آموزش داده شده ۵ جمله جدید تولید کنید (طول جملات بین ۱۲ تا ۲۴ توکن باشد)

ه) قسمت ب را برای n=3 و n=5 تکرار کنید و مجددا به ازای هر مدل ۵ جمله جدید تولید کنید.

و) جملات به دست آمده برای سه مدل آموزش داده شده را مقایسه کنید. به نظر شما کدام مدل جملات بهتری تولید کرده است؟

ی) دو جمله زیر را در نظر بگیرید:

۱. هری به هاگوارتز بازگشت.

۲. نماز ستون دین است.

مدل زبانی شما Perplexity کدام جمله را بیشتر می داند؟ چرا؟ (نیازی به پیاده سازی نیست و توضیح کافی است)

سوال ۲

یکی از بخش های مهم در پردازش زبان طبیعی چگونگی انجام Tokenization بر روی داده های ورودی می باشد. روش ها و کتابخانه های متعددی برای انجام این کار وجود دارد. در این بخش به مقایسه سه روش پرداخته می شود:

۱. White Space Tokenization

۲.Spacy Tokenizer

۳.Subword Tokenization(BPE)

الف) در مورد روش کار هر یک از موارد بالا توضیح مختصر دهید.

ب) یک بار داده های فارسی و بار دیگر داده های انگلیسی را بر روی سه Tokenizer ذکر شده اعمال کنید. سپس جدول زیر را کامل کنید. (برای آموزش bpe از همان داده های فارسی یا انگلیسی استفاده کنید. در صورت دادن ورودی فارسی از داده های فارسی و برای ورودی انگلیسی از داده های انگلیسی برای آموزش استفاده کنید)

الگوریتم استفاده شده	تعداد توکن های خروجی برای کتاب هری پاتر	
	زبان فارسی	زبان انگلیسی
White Space		
Spacy		
BPE		

ج) هر یک از وروی های زیر را با سه روش گفته شده Tokenize کنید و نتایج به دست آمده را مقایسه کنید. استفاده از روش White Space Tokenization چه مشکلی دارد؟

en_input = This question is about tokenization and shows several tokenizer algorithms.Hopefully, you will be able to understand how they are trained and generate tokens.

fa_input = این سوال در مورد قطعه بندی جملات است و چندین الگوریتم توکنایز کردن متن را نشان می دهد. امیدواریم بتوانید نحوه آموزش آنها و تولید توکن ها را درک کنید.

ملاحظات (حتما مطالعه شود)

تمامی نتایج شما باید در یک فایل فشرده با عنوان NLP_CA1_StudentID تحویل داده شود.

- خروجی مورد انتظار تمرین فایل jupyter به همراه فایل گزارش است (گزارش می تواند داخل jupyter هم نوشته شود)
- خوانایی و دقت بررسی‌ها در گزارش نهایی از اهمیت ویژه‌ای برخوردار است. به تمرین‌هایی که به صورت کاغذی تحویل داده شوند یا به صورت عکس در سایت بارگذاری شوند، ترتیب اثری داده نخواهد شد.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه‌ی کدهای پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرا مجدد آنها نیاز به تنظیمات خاصی می‌باشد بایستی تنظیمات مورد نیاز را نیز در گزارش خود ذکر کنید.
- توجه کنید این تمرین باید به صورت تک نفره انجام شود و پاسخ‌های ارائه شده باید نتیجه فعالیت فرد نویسنده باشد (همفکری و به اتفاق هم نوشتن تمرین نیز ممنوع است). در صورت مشاهده تقلب به همه افراد مشارکت کننده، نمره تمرین صفر و به استاد نیز گزارش می‌گردد.
- در صورت بروز هرگونه مشکل با خانم ثمین مهدی زاده از طریق ایمیل زیر در ارتباط باشید:

  saminsani162@gmail.com

مهلت تحویل: ۲۲ اسفند ۱۴۰۱