پردازش زبان های طبیعی

مدرسین: دکتر فیلی، دکتر یعقوب زاده

تمرین کامپیوتری اول– Tokenization

حامد باغبانی (Baghbani.hamed@ut.ac.ir)

مهلت ارسال: ۲۲ اسفند

یکی از بخش های مهم در پردازش زبان های طبیعی چگونگی انجام عملیات Tokenization بر روی داده‌های ورودی می باشد. در این تکلیف می خواهیم دو روش متداول Tokenization و جزئیات مربوط به آن ها را بررسی کنیم. دو روش مد نظر روش های زیر می باشند:

۱) Byte Pair Encoding (BPE) Algorithm

۲) WordPiece Algorithm

در **گام اول** هر کدام از دو روش گفته شده را با یکدیگر مقایسه کرده و تفاوت ها و شباهت های هر کدام را تشریح کنید. جهت آشنایی بیشتر با روش های گفته شده می توانید از مقاله موجود در لینک [1] استفاده کنید. به طور خاص روش BPE را خودتان از صفر تا صد پیاده سازی کرده و بر روی مثالی که در ادامه آمده است فرایند اجرای این الگوریتم را به صورت مرحله به مرحله تشریح کنید (فرایند های تعریف Vocabulary از روی کاراکترهای متن ورودی، محاسبه فرکانس رخداد جفت ها و سپس ادغام پرتکرار ترین جفت و ... را بر روی مثال ورودی با جزییات کامل نشان دهید). همچنین نشان دهید پیاده سازی شما چگونه کلمه "lowest" را به عنوان یک کلمه Out of Vocabulary توکنایز می‌کند.

Corpus

low	lower	newest
low	lower	newest
low	widest	newest
low	widest	newest
low	widest	newest

در **گام دوم** با استفاده از کتابخانه hugging face[2] هر کدام از دو روش گفته شده را پیاده سازی کنید. تعداد توکن‌های استخراج شده توسط توکنایزرها را برای نمونه متن ورودی داده شده و کیفیت آن ها بررسی کنید. این کار را بر روی تمام

[1] https://huggingface.co/docs/transformers/tokenizer_summary
[2] https://huggingface.co/docs/tokenizers/python/latest/quicktour.html

مدل ها که به ترتیب بر روی دیتاست اول و دوم (هر کدام به طور جداگانه) آموزش داده شده اند را بررسی کرده و تفاوت‌های آن ها را با توجه به روش مورد استفاده تحلیل کنید همچنین نشان دهید تاثیر حجم داده های آموزشی بر روی هر الگوریتم چگونه بوده است. در ادامه عملکرد هر الگوریتم را بر روی توکن های OOV[3] مثال داده شده نیز بررسی کنید.

یک نمونه متن ورودی جهت ارزیابی توکنایزرهای آموزش داده شده:

```
This is a deep learning tokenization tutorial. Tokenization is the first step in a deep learning NLP pipeline. We will be comparing the tokens generated by each tokenization model. Excited much?! 😍
```

در **گام سوم** تعداد توکن های هر کدام از دو الگوریتم داده شده را بر روی متن کتاب گوتنبرگ بدست بیاورید و در جدولی به صورت زیر گزارش کنید:

ردیف	نام الگوریتم استفاده شده برای توکنایز	تعداد توکن های خروجی الگوریتم برای کتاب کوتنبرگ	
		توکنایزر آموزش داده شده بر روی کتاب گوتنبرگ	توکنایزر آموزش داده شده بر روی کل داده های ویکی پدیا
۱	Byte Pair Encoding (BPE)		
۲	WordPiece		

مجموعه دادگان:

برای آموزش مدل ها از دو مجموعه داده زیر استفاده کنید. داده اول متن یک کتاب از گوتنبرگ می باشد (با توجه به کوچک بودن این داده توصیه می شود در ابتدا از این کتاب برای پیاده سازی استفاده کنید) و مجموعه دادگان دوم نیز مجموعه داده ویکی پدیا انگلیسی می باشد.

http://www.gutenberg.org/cache/epub/16457/pg16457.txt

https://s3.amazonaws.com/research.metamind.io/wikitext/wikitext-103-raw-v1.zip

[3] Out of Vocabulary

ملاحظات (حتما مطالعه شود):

- تمامی کد ها و گزارش مربوطه بایستی در یک فایل فشرده با عنوان NLP_CA1_StudentID تحویل داده شود.
- کدهای نوشته شده برای هر بخش را با نام مناسب مشخص کرده و به همراه گزارش تکلیف ارسال کنید. همه کد های پیوست گزارش بایستی قابلیت اجرای مجدد داشته باشند. در صورتی که برای اجرای مجدد آنها نیاز به تنظیمات خاصی می باشد 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 وجود هر گونه سوال، می توانید با دستیار آموزشی مربوطه در تماس باشید.