Mohammad Javad Ranjbar

810101173

Homework 1

Natural language processing, Spring 2023

سوال ۱:

برای پیاده‌سازی از کتابخانه‌های nltk و hazm استفاده شده است. و مدل‌های n-gram هم با استفاده از کتابخانه و هم بدون استفاده از کتابخانه پیاده‌سازی شده‌اند.

الف) در متن فارسی تعدادی کاراکتر غیرفارسی و تصادفی وجود دارد که بهتر است آن‌ها حذف شوند، لذا تمام کاراکترهای غیر فارسی و اضافی را حذف می‌کنیم (البته علائم نگارشی را نگه می‌داریم.). سپس با استفاده از normalizer پیاده‌سازی شده کتابخانه‌ی hazm کلمه‌های فایل را تصحیح می‌کنیم، این نرمال‌سازی شامل تصحیح فاصله‌ها و کلمات است، برای مثال در نسخه فایل داده شده تمام حروف "ی" به صورت "ي" نوشته شده‌اند که نادرست است،(در صورتی که این تصحیح صورت نمی‌گیرفت، در بخش ی از این سوال به مشکل می‌خوردیم زیر "هری" در کل کتاب وجود نداشت). همینطور پس از قطعه بندی جملات با استفاده از کارکترهای "<s>" و "</s>" ابتدا و انتهای هر جمله رو مشخص می‌کنیم.

ب) مدل را آموزش دادیم که در کد موجود است.

ج) در صورتی که از Laplace smoothing استفاده نشود. مدل n-gram آموزش دیده، ممکن است به n-gram هایی که در داده‌ی آموزش نیست احتمال صفر را نسبت دهد، این اتفاقا مخصوصا وقتی n بیشتر می‌شود، با احتمال بیشتری رخ خواهد داد. به عبارت دیگر اگر در داده‌ی تست ما جمله‌ای باشد که n-gramای در آن باشد که در مجموعه داده اصلی نیست (که با توجه به خاصیت‌های زبان زیاد هم رخ می‌دهد.) احتمال آن و در نتیجه احتمال کل جمله صفر خواهد شد. و در نتیجه مدل خوب کار نخواهد کرد.

د) جملات تولید شده به شکل زیر می‌باشند. برای مدل‌های مختلف به شکل زیر می‌باشند:

2-gram با استفاده از کتابخانه:

```
اما او را نمی‌شناخت دستش را از یک شیرجه زدن خودش رو داخل سالن عمومی خلوت پیدا کند
دید . حذر از این واقعیت داره؟ ورنون، مزخرف بره . دو پسر رنگپریده‌ای
بگیر . از داخل شدند و بینیاش را بر روی دیوار کلبه می‌شدند، پلکان مرمری بالا آورد و خروج از
و هیچ وجه قصد استفاده از ویزلیها بوده ن . روی زمین به

آخرین ماه در روشنایی روز جالبتر می‌شد . چرا چیزی بگوید از یاد یک ترول رد شدند و رونده‌ی گیاه به
```

2-gram بدون استفاده از کتابخانه:

```
هري گفت : «هري مال ما رو به آن ها را از این که در همان لحظه اي بعد با صداي بلند و هرميون
، نه ؟ اون جا که از آن ها را به هري گفت:
رون گفت : «هري مال ما رو به هري که از آن ها را در این جا دور و با صداي بلند شد تا
هاگريد با صداي بلند و گفت : «هري مال ما رو به هري که از آن ها را در این جا دور خود آمد
آن ها را به هري گفت : «هري مال ما رو از این که در همان لحظه اي بعد با صداي بلند و هرميون
هرميون گفت : «هري مال ما رو به هري که از آن ها را در این جا دور و با صداي بلند شد تا
در آن ها را به هري گفت : «هري مال ما رو از این که با صداي بلند و هرميون گرنجر وارد اتاق نشيمن
من به هري گفت : «هري مال ما رو از آن ها را در این که با صداي بلند و هرميون گرنجر وارد اتاق
این جا که از آن ها را به هري گفت : «هري مال ما رو توي هاگوارتز
اما هري گفت : «هري مال ما رو به آن ها را از این که در همان لحظه اي بعد با صداي بلند و
```

پروفسور مک گونگال گفت : «هري مال ما رو به هري که از آن ها را در این جا دور و با صداي بلند

ه) نتایج آموزش برای مدل‌های ذکر شده در کد موجود است.

3-gram با استفاده از کتابخانه:

اسنیپ بی‌دلیل پنالتی دیگری به نفع تیم هافل پاف هم از صبج وحشت دارد.
اش درست مثل موی هری شاخ شده بود) اما برایش یک بسته بزرگ قورباغه ي شکلاتي
آدم خیلي خوبي نیست ولي هیچ عددي روي صفحه ي شطرنج خارج شدند.
به گریپهوک کرد و اوج میگیره ... واي ... حتي ولدمورت هم توی تیم بودیم.
مي دونم که تا چشم کار میکرد جنگل بود و آن قدر از کار آن ها

3-gram بدون استفاده از کتابخانه:

هري که از آن جا دور شده بود. آن ها پایین رفت و گفت
هاگرید با حالتي آمرانه گفت(:
آن ها را به سمت هري حمله ور شد و گفت :
هرمیون گفت
در همان لحظه صداي فریاد وحشت زده ي هري را به سمت قلعه حرکت کرد. تردیدي نداشت که اسم خواهرزاده
من به او نگاه کرد و گفت :
این جا رو نگاه کن
اما هري و رون به او نگاه کرد که در آن لحظه تنها چیزي بود
پروفسور مک گونگال گفت:

5-gram با استفاده از کتابخانه:

هرچه به صخره ي عظیمي که پیش رویشان بود خیره نگاه میکردند
که در خوابگاهشان بود، بر سر بازی فوتبال جر و بحث کردند.
محض اطلاعت بگم که از گل سوسن و گیاه افسنتین معجون خواب آوري درست میکنن به اسم شربت زنگي
اش را سپر کرد و با غرور خاصی ادامه داد : اون خطاهاي کسي رو به این
بود تکاني خورد و شکافي که نزدیک به لبه ي آن بود مثل

5-gram بدون استفاده از کتابخانه:

هري که از قطع شدن گفتگویش با آن پسر اصلاً ناراحت نبود
رون گفت : اي وروجک ! خیلي سرش توي حساب و کتابه ... درست مثل این که به نظر میرسید هیچ جاهاگرید با حالتي صمیمانه و تحسین آمیز به هري نگاه میکرد. ولي
آن ها به ایستگاه رسیده
هرمیون گفت :
در همان لحظه دامبلدور از روي دیوار کوتاه باغ رد شد و به سوي
من به او اعتماد کامل دارم(.
این جا رو نگاه
اما هري هیچ قانوني را نقض نکرده بود و این کار
پروفسور مک گونگال با غضب بیني اش را بالا کشید و ساعت

و) مدل 5-gram نسبت به دو مدل دیگر نتیجه‌ی بهتری دارد. مشخص است که این مدل به تعداد کلمات پیشین بیشتری توجه می‌کند و در نتیجه بیشتر ساختاری شبیه یک جمله واقعی خواهد داشت.

ی) جمله‌ی نماز ستون دین است دارای پیچیدگی بیشتری خواهد بود. زیرا که کلمه‌ی نماز و حتی دین در متن کتاب هری پاتر وجود ندارد، و در نتیجه احتمال وقوع این کلمه نزدیک به صفر خواهد بود و پیچیدگی این جمله بسیار بزرگ خواهد بود.

نتایج پیاده‌سازی برای این سوال به صورت زیر:

```
The perplexity of هری به هاگوارتز برگشت . for the 2-gram model is equal
to:10.79780320069082
The perplexity of هری به هاگوارتز برگشت . for the 3-gram model is equal
to:11.616738172735303
The perplexity of هری به هاگوارتز برگشت . for the 5-gram model is equal
to:15.959275097341068
```

و برای جمله دوم داریم:

```
The perplexity of است دین ستون نماز . for the 2-gram model is equal
to:12.008154645979973
The perplexity of است دین ستون نماز . for the 3-gram model is equal
to:12.715157875405364
The perplexity of است دین ستون نماز . for the 5-gram model is equal
to:17.240739722050947
```

مشخص است که پیچدگی جمله دوم به شدت برای همه‌ی مدل‌ها بیشتر از جمله‌ی اول می‌باشد.

سوال ۲:

الف)

۱- White Space Tokenization: این روش برای شکستن متن به بخش‌های کوچکتر، از کاراکترهای white space شامل newline، tab و space استفاده می‌کند. به عبارتی هر دنباله‌ای از کاراکترهای white space در متن به عنوان جداکننده‌ی بخش‌های دیگر متن مورد استفاده قرار می‌گیرد. به عنوان مثال جمله‌ی " کوروش برای خرید به مغازه رفت." تبدیل به "کوروش"، "برای"، "خرید"، "به"، "مغازه"، "رفت." می‌شود. این روش بسیار ساده می‌باشد، با این حال، گاهی اوقات سادگی بهترین انتخاب نیست، زیرا در این روش انواع دیگر جداکننده‌ها مانند علائم نقطه گذاری را در نظر گرفته نشده، و همچنین می‌تواند کاراکترها ناخواسته را در صورتی که متن حاوی کاراکترهای white space اضافی باشد، تولید کند.

۲- Spacy Tokenizer: این روش نیز بر اساس قوانینی متن را به بخش‌های کوچکتر می‌شکند که به صورت زیر می‌باشد:

- ابتدا همانند روش White Space Tokenization کلمات که با فاصله از هم جدا می‌شوند. البته باید توجه داشت قوانین خاصی برای کلمات خاص مانند New York وجود دارد که از هم جدا نخواهند شد.
- علائم نگارشی نیز جدا شده، اما این روش context را نیز در نظر می‌گیرد برای مثال در اعداد اعشاری یا برای کلماتی مانند U.S.A به صورت خاص این مرحله اجرا نمی‌شود. همینطور برای کلماتی مانند don’t با اینکه white space ندارند اما باید تبدیل به do و n’t شوند.

برای مثال جمله‌ی "آرش برای خرید به سیستان بلوچستان رفت." تبدیل به "آرش"، "برای"، "خرید"، "به"، "سیستان بلوچستان"، "رفت"، "." می‌شود.

۳- Subword Tokenization (BPE): این روش اساس مراحل عمل می‌کند:

- مجموعه لغات را ابتدا تمام کاراکترهای موجود در متن می‌سازیم.
- تعداد تکرار هر دو کاراکتر مجاور در متن را می‌شماریم.

- پر تکرار ترین دو کاراکتر مجاور را به مجموعه لغات خود اضافه می‌کنیم. برای مثال اگر دو کاراکتر "ب، ا" باشند، "با" به مجموعه لغات اضافه می‌شود.
- این روند را انقد ادامه می‌دهیم تا به تعداد حد مطلوب برسیم یا هیچ دوکارکتر جدیدی پیدا نشود.

برای مثال جمله‌ی "او با پدرش در باران به پارک رفت." مجموعه لغات ما برابر با:

"ا"، "و"، "ب"، "د"، "ر"، "ش"، "ن"، "ه" "ک"، "ف"، ت" خواهد بود.

بیشترین دو کارکتر پر تکرار "با" بوده پس مجموعه لغات به صورت زیر خواهد شد:

"ا"، "و"، "ب"، "د"، "ر"، "ش"، "ن"، "ه" "ک"، "ف"، ت"، "با"

در مرحله‌ی بیشترین دو کاراکتر پرتکرار برابر با "در" است که به مجموعه لغات اضافه می‌کنیم:

"ا"، "و"، "ب"، "د"، "ر"، "ش"، "ن"، "ه" "ک"، "ف"، ت"، "با"، "در"

و....

این فرایند را تا حد مطلوب مورد نظرمان ادامه می‌دهیم.

ب) جدول حاصل به صورت زیر خواهد شد:

الگوریتم استفاده شده	تعداد توکن‌های خروجی برای کتاب هری پاتر	
	زبان فارسی	زبان انگلیسی
White space	96294	78443
Spacy	125677	102406
BPE	106734	100012

ج)

برای جمله‌ی "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." داریم:

White Space Tokenizer:

```
['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'.]
```

همانطور که توضیح داده شد این tokenizer بدون توجه به هیچ زمینه‌ای، کلمات را فقط بر اساس کاراکترهای مربوط به فاصله از یکدیگر جدا کرده است.

Spacy Tokenizer:

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

در این روش tokenization هم، کلمات ابتدا با توجه به قوانینی و کاراکترهای فاصله از یکدگیر جدا می‌شوند.

BPE:

['__sow','ا','__unk','ن','__eow','__sow','ال','سو','__eow','مورد','در','__sow','عه','قط','__eow',
'__sow','د','است','__sow','و','__eow','جم','لا','ت','__sow','__eow','__unk','بن','د','چن',
'__unk','ز','تو','کن','ا','__sow','__eow','تم','__unk','ال','گو','ر','__sow','__eow','ن',
'__eow','کردن','__sow','ن','مت','__eow','نشان','را','__sow','م','__unk','__eow','دهد','.','__sow','ام',
'__unk','ار','دو','__unk','م','__eow','__sow','ن','وا','بت','__unk','د','__eow','__sow','ن',
'تو','کن','__sow','__eow','د','__unk','تو','ل','__sow','آموزش','آنها','و','__eow','حو','ه',
'__eow','درک','را','ها','__sow','کن','__unk','د','__eow','.']

این روش نیز همانند توضیح داده شده عمل خواهد کرد البته مشخص است که چون در متن فارسی هری پاتری کلماتی مانند الگوریتم و توکن وجود ندارد تعدادی از کلمات unk_ خواهند بود.

برای متن انگلیسی کتاب هری پاتر داریم:

White space tokenizer:

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

کلمات بر اساس فاصله جدا شده‌اند و در بخش 'algorithms.Hopefully,' به علت غلط، فاصله گذاشته نشده است. این روش tokenization قابلیت تشخیص این را ندارد.

Spacy tokenizer:

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

---

این روش نیز مورد انتظار عمل کرد و مشکل روش قبل را نداشت.

BPE:

['this', 'question', 'is', 'about', '__sow', '__unk', '__unk', '__unk', '__unk', '__unk', '__unk', '__unk', '__unk', '__unk',
'__unk', '__unk', '__unk', '__eow', 'and', 'shows', 'several', '__sow', '__unk', '__unk', '__unk', '__unk', '__unk', '__unk',
'__unk', '__unk', '__unk', '__eow', '__sow', '__unk', '__unk', '__unk', '__unk', '__unk', '__unk', '__unk', '__unk', '__unk',
'__unk', '__eow', '.', 'hopefully', ',', 'you', 'will', 'be', 'able', 'to', 'understand', 'how', 'they', 'are', 'trained', 'and',
'__sow', '__unk', '__unk', '__unk', '__unk', '__unk', '__unk', '__unk', '__unk', '__eow', 'tokens', '.']

به مانند متن فارسی نیز این روش کلمات زیادی که در متن کتاب اصلی نبودند را نمی‌تواند تشخیص دهد.

به طور کلی روش white space tokenization فقط بر اساس فاصله جدا می‌کند که با وجود سادگی مشکلاتی دارد، روش spacy علاوه بر فاصله به قوانین دیگری نیز توجه می‌کند که نسبت به white space روش بهتری است، و در نهایت روش BPE در صورتی که داده آموزش تفاوت زیادی با داده‌ی تست داشته باشد (مانند متن صورت سوال) در tokenize کردن کلمات به مشکل خواهد خورد.

مشکلات استفاده از روش white space tokenization به صورت زیر می‌باشد:

- این روش به علائم نگارشی اهمیت نمی‌دهد. برای مثال جمله‌ی "سورنا آمد!" بعد از tokenize کردن بر اساس این روش تبدیل به "سورنا"، "آمد!" می‌شود.
- این روش کلماتی که داخل خودشان فاصله دارند را نمی‌تواند شناسایی کند. برای مثال کلمه‌ی "سیستان بلوچستان" تبدیل به دو توکن "سیستان" "بلوچستان" خواهد شد.