پروژه درس بازیابی اطلاعات

امیرحسین سرور - ۹۷۳۱۰۲۸

گزارش فاز اول

۱) با ذکر مثال شرح دهید که در گام پیش پردازش چه عملیاتی انجام داده اید. همچنین دلیل انجام هر پردازش را ذکر کنید.

در اولین مرحله، پس از خواندن اسناد از فایل اکسل، آنها را در آرایهای به صورت خام ذخیره می کنیم. تابع پیشپردازش یا همان preprocess این آرایه از اسناد (documents) را به عنوان ورودی دریافت می کند و خروجی آن باید آرایهای باشد از آرایهی توکنهای موجود در هر سند (document_tokens). یک نکته قابل ذکر اینکه در این پروژه برای پردازش متون از کتابخانه هضم (hazm) استفاده شده است. برای شروع مراحل پیشپردازش، روی آرایه اسناد حرکت می کنیم و به ترتیب برای هر سند، عملیاتهای نرمالسازی، استخراج توکن، ریشه یابی و حذف کلمات پرتکرار را انجام می دهیم. در نهایت آرایه توکنهای ایجاد شده از هر سند را به آرایه خروجی اضافه می کنیم. این مراحل در کد به صورت

زير است:

```
def preprocess(documents):
    normalizer = Normalizer(token_based=True)
    lemmatizer = Lemmatizer()
    documents_tokens = []

for document in documents:
    normalized_doc = normalizer.normalize(document)

    tokens = word_tokenize(normalized_doc)
    final_tokens = []

    for token in tokens:
        token = lemmatizer.lemmatize(token)
        token = refine_token(token)

        if not token:
            continue
        if token in stopwords_list():
            continue

        final_tokens.append(token)

        documents_tokens.append(final_tokens)

return documents_tokens
```



به عنوان مثال، فرض کنیم میخواهیم اسناد زیر را پیشپردازش کنیم:

سند ۱: بُشقابِ مَن و رمـــان «خانه ی پدری» را بِگیر.

سند ۲: فاصله میان پیشوند ها و پسوند ها را، اصلاح می کند.

🔎 پس ورودی تابع به صورت زیر خواهد بود:

['بُشقابِ مَن، و «رمـــان خانه ی پدری» را بگیر.', 'فاصله میان پیشوند ها و پسوند ها را اصلاح می کند.']

حال هرکدام از اسناد را به ترتیب برداشته، ابتدا آنها را نرمال میکنیم (برای توضیح ساده تر فرض میکنیم هر دو سند همزمان در حال پردازشاند). میبینیم که اعراب حذف شدهاند، حروف کشیده به حالت اصلی برگشتهاند و نیمفاصله هم رعایت شده است که در نرمالسازی هم قصدمان انجام همین کارهاست :

. ['بشقاب من، و «رمان خانه\u200cی پدری» را بگیر.', 'فاصله میان پیشوند\u200cها و پسوند\u200cها را اصلاح می\u200cکند.']

* (در اینجا \u200c همان نیمفاصله است)

🖊 سپس توکنهای هر سند را استخراج میکنیم:

[['بشقاب', 'من', '،', 'و', '«', 'رمان', 'خانه\u200cء', 'پدری', '»', 'را', 'بگیر', '.'] ['فاصله', 'میان', 'پیشوند\u200cدها', 'و', 'پسوند\u200cدها', 'را', 'اصلاح', 'می\u200cدند', '.']]

در این مرحله با استفاده از تابع lemmetize در کتابخانه هضم، به ریشهیابی هر توکن میپردازیم تا کلمات و افعال معادل چندین بار به صورتهای مختلف در دیکشنری نهایی ظاهر نشوند. میبینیم که برخی از پسوندها مثل «ها» و «ی» از کلمات حذف شدهاند و افعال هم به صورت بنهای ماضی و مضارع درآمدهاند:

, ['.', 'و', '،', 'و', '«', 'رمان', 'خانه', 'پدر', '»', 'رו', 'گرفت#گير', '.']] ['فاصله', 'ميان', 'پيشوند', 'و', 'پسوند', 'را', 'اصلاح', 'کرد#کن', '.']]



سپس علائم نگارشی را از توکنها حذف می کنیم چرا که علائم نگارشی هم جزو توکنها حساب شدهاند که عملا هیچ فایدهای برای ما ندارند و فقط حجم دیکشنری را افزایش می دهند. اگر توکنی فقط شامل علائم نگارشی باشد (مثلا فقط نقطه یا ویر گول) باشد، در این قسمت حذف خواهد شد و به لیست توکنها اضافه نمی شود (دستور continue).

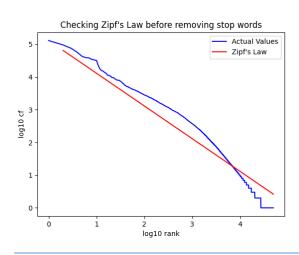
```
, ['بشقاب', 'من', 'و', 'رمان', 'خانه', 'پدر', 'را', 'گرفتگیر']]
['فاصله', 'میان', 'پیشوند', 'و', 'پسوند', 'را', 'اصلاح', 'کردکن']]
```

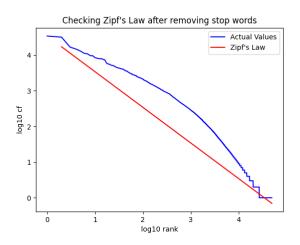
در مرحله آخر کلمات پرتکرار (Stop Words) را هم از توکنها حذف می کنیم، به این صورت که هر توکن اگر در مرحله آخر کلمات پرتکرار (تابع (stopwords_list() در کتابخانه هضم) وجود داشت، آن را به لیست توکنهای نهایی اضافه نخواهیم کرد (دستور continue) چون کلمات پرتکرار در اکثر اسناد موجود هستند و نگهداری آنها در جستجوی پرسمان کمک چندانی نخواهد کرد. پس درنهایت لیستی از لیست توکنهای نهایی، به صورت زیر خواهیم داشت:

```
[['بشقاب', 'رمان', 'خمانه', 'پدر', 'گرفتگیر'], ['فاصله', 'پیشوند', 'پسوند', 'اصلاح', 'کردکن']]
```

 ۲) صحت قانون Zipf را در دو حالت قبل از حذف کلمات پرتکرار از واژهنامه و بعد از حذف کلمات پرتکرار بررسی کنید. در صورت برقراری/ عدم برقراری این قانون در هر حالت، علت را شرح دهید.

با استفاده از تابع (check_Zipf)، نتایج زیر بدست می آیند:







Zipf است که از روی دیکشنری برداشته و توابع قرمز رنگ هم نشان دهنده رابطه که از روی دیکشنری برداشته ایم و توابع قرمز رنگ هم نشان دهنده رابطه است که میباشند. از آنجایی که در این قانون داریم K ایم الم در از آنجایی که در این قانون داریم K ایم الم دیکشنری دارد؛ پس می توان نتیجه گرفت اگر کلمات پرتکرار را حذف آن هم بستگی به تعداد تکرار پرتکرار ترین کلمه دیکشنری دارد؛ پس می توان نتیجه گرفت اگر کلمات پرتکرار حذف شده اند نکنیم، پرتکرار ترین کلمه، تکرار به مراتب بیش تری نسبت به پرتکرار ترین کلمه در زمانی که کلمات پرتکرار حذف شده دارد، چرا که در حالت دوم پرتکرار ترین کلمه لزوما یک کلمه پرتکرار نیست. بنابراین این رابطه زمانی که کلمات پرتکرار حذف نشده این موضوع هستند.

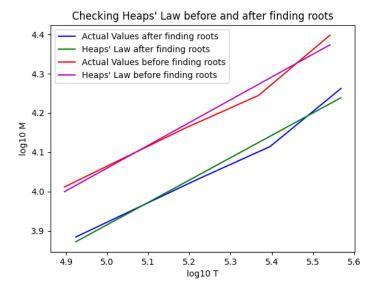
۳) صحت قانون heaps را در دو حالت قبل و بعد از ریشه یابی بررسی کنید . برای بررسی این قانون لازم است با استفاده از اندازه ی واژه نامه و تعداد توکنها در ۵۰۰، ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ سند اول، اندازه ی واژه نامه مربوط به کل اسناد تخمین زده شود. در نهایت اندازه ی واقعی واژه نامه و اندازه ی تخمینی در هر دو حالت مقایسه و تحلیل شود. آیا در هر دو حالت قانون برقرار است؟ چرا؟

در ابتدا، دیکشنری مربوط به ۵۰۰، ۵۰۰، ۱۵۰۰ و ۲۰۰۰ سند اول را به دو صورت، قبل و بعد از ریشهیابی کلمات، میسازیم و سپس اندازه آن و همچنین تعداد توکنهایی را که از این اسناد بدست آمدهاند را بدست میآوریم. نتیجه این آمار به صورت زیر خواهد بود:

	After finding roots				Before finding roots			
N	500	1000	1500	2000	500	1000	1500	2000
Т	48020	165259	248753	369820	78804	155059	233399	347643
M	7664	10726	13016	18292	10274	14516	17570	24998

اکنون با استفاده از این آمار و تابع (check_heaps)، در هر دو حالت نمودار مقادیر واقعی را رسم می کنیم و بعد از آن با استفاده از این نقاط خطی که BestFit بر این نقاط است را بدست می آوریم که همان رابطه ی لگاریتمدار و خطی قانون هیپس است و آن را رسم می کنیم. نمودارها به صورت زیر می شوند:





سپس با استفاده از توابع بدست آمده، سعی می کنیم اندازه دیکشنری نهاییمان را با استفاده از تعداد تو کنهای موجود در کل اسناد، پیشبینی کنیم:

در شکل بالا، اندازه تخمینی واژهنامه در دو حالت قبل و بعد از ریشهیابی، و همینطور اندازه واقعی آن نشان داده شده است. همانطور که مشخص است، در حالتی که کلمات را ریشهیابی کردهایم، قانون هیپس نتیجه دقیق تری به ما می دهد. چرا که بسیاری از کلمات معادل، به عنوان یک توکن در دیکشنری آمدهاند و اندازه دیکشنری را به طور غیرمعقول زیاد نمی کنند؛ و همینطور اینکه اندازه واقعی واژهنامه با اندازه تخمینی آن در حالت بدون ریشهیابی اختلاف بیشتری دارد.



 ع) حداقل سه مورد از مواردی که در ریشه یابی با چالش روبرو بودید را ذکر کنید (بطور مثال کلماتی که نیازی به ریشه یابی ندارند اما طبق روند ریشه یابی از دست می روند).

چالش اول: برای ریشهیابی کلمات، در کتابخانه هضم دو تابع وجود داشت. یکی stemmer و یکی stemmatizer. بعد از امتحان کردن هردو، متوجه شدم که lemmatizer به طور دقیق تری کلمات را ریشهیابی می کند. پس سعی کردم از آن استفاده کنم اما عملکرد آن به شدت کندتر از stemmer بود؛ چرا که با وجود اینکه دقیق تر بود، اما بخاطر جستجویی که درون خودش انجام می داد سرعت برنامه را پایین می آورد. اما stemmer در زمان معقول تری ریشهیابی می کرد ولی دقت پایین تری را نتیجه می داد. در واقع یک trade-off بین سرعت و دقت.

چالش دوم: بعد از استفاده از تابع stemmer، متوجه شدم که بعضی جاها کلمات را اشتباهاً ریشه یابی می کند در حالی که احتیاجی به ریشه یابی نداشتند، و این کار طبعاً دقت را پایین می آورد. به عنوان مثال، گاهی به اشتباه پسوندها و پیشوندها را حذف می کند در کلمه «اصفهان»، «ان» آن را تصور اینکه نشانه جمع است، حذف می کند در حالی که اینکار در کلماتی مثل درختان یا سربازان مجاز است نه در این کلمات.

چالش سوم: برای ریشه یابی افعال، این توابع معمولاً هرچیزی بغیر از بن ماضی یا مضارع آن را از کلمه حذف می کردند. مثلا فعل «رفتیم» به «رفت» تبدیل میشد. اینکار باعث میشد کلماتی مثل «تصمیم» هم به عنوان فعل در نظر گرفته شده و به «تصم» تبدیل شوند که کلمهای بی معنی است.

۵) پاسخ به پرسمان در حالتهای زیر. در هر مورد، تیتر خبر بازیابی شده را به همراه جمله(هایی) که حاوی عبارت پرسمان بودهاند، گزارش کنید. همچنین در هر مورد با ذکر جزئیات شرح دهید که آیا سند بازیابی شده به پرسمان کاربر مرتبط هست یاخیر؟

الف) یک پرسمان از کلمات ساده و متداول تککلمهای (بینالملل)

پرسمان: انتخابات



ب) یک پرسمان از عبارات ساده و متداول دو کلمهای (دانشگاه امیر کبیر)

پرسمان: فدراسیون کشتی

```
>> Please Enter your Query: ﴿ الله المعافل ال
```



پ) یک پرسمان از عبارات ساده و متداول چند کلمه ای (دانشگاه صنعتی امیر کبیر، سازمان ملل متحد، جمهوری اسلامی ایران)

پرسمان: بانک رفاه کارگران

```
>> Please Enter your Query: الله المال ال
```

ت) پرسمان دشوار و کمتکرار تککلمهای (ژیمناستیک)

◄ پرسمان: ژیمناستیک



ث)یک پرسمان دشوار و کمتکرار دوکلمهای (واکسن آسترازنکا)

پرسمان: واکسن آسترازنکا

برای هرکدام از پرسمانهای بالا، جمله(هایی) را که حاوی پرسمان بودهاند برای اولین خبر بازیابی شده نشان میدهیم:

- 🔪 پرسمان: انتخابات
- تیتر خبر: عدم حضور سوریان در انتخابات اتحادیه جهانی کشتی
 - 🗡 متن خبر حاوی پرسمان:
- حمید سوریان امسال در انتخابات اتحادیه جهانی کشتی شرکت نمی کند. قرار است در حاشیه مسابقات انتخابات اتحادیه جهانی کشتی نیز برگزار شود.
- ارتباط خبر به پرسمان: سند به انتخابات جهانی کشتی مربوط بوده و کاربر هم به دنبال نتایج انتخابات بوده است. با اینکه کاربر مشخص نکرده چه انتخاباتی مدنظرش است، سند بازیابی شده به پرسمان مرتبط است.



- پرسمان: فدراسیون کشتی
- 🗡 تیتر خبر: دبیر: رفیقبازی کردم یا نه در نروژ نتیجه گرفتیم/ مطمئن باشید سال بعد بهتر نتیجه می گیریم
 - 🗸 متن خبر حاوی پرسمان:

رئیس فدراسیون کشتی گفت: اگر رفیق بازی کردم، نتیجه آن را در مسابقات جهانی نروژ دیدیم . رئیس فدراسیون کشتی ادامه داد.

رئیس فدراسیون کشتی در پاسخ به سوال دیگر تصریح کرد.

ارتباط خبر به پرسمان: سند مربوط به صحبتهای رییس فدراسیون کشتی است که با پرسمان کاربر هم انطباق خوبی دارد و مرتبط است.

- **پرسمان:** بانک رفاه کارگران
- تیتر خبر: بررسی ساماندهی و بهبود وضعیت مالی بانک رفاه کارگران در جلسه ستاد هماهنگی اقتصادی دولت متن خبر حاوی پرسمان:

در جلسه ستاد هماهنگی اقتصادی دولت، درباره ساماندهی و بهبود وضعیت مالی بانک رفاه کارگران بحث و تصمیم گیری شد.

مجلس شورای اسلامی برای ساماندهی امور بانک رفاه کارگران، مقرر شد ضمن سرعت بخشیدن به بازپرداخت بدهی بدهی دولت به سازمان تامین اجتماعی و صندوق بازنشستگی کشور، اقدامات لازم برای بازپرداخت بدهی سازمان تامین اجتماعی به بانک رفاه کارگران نیز به عمل آید.

طرح "میزکار کارگزاری" بانک رفاه کارگران کلید خورد.

ارتباط خبر به پرسمان: سند مربوط به بررسی وضعیت بانک رفاه کارگران است که با پرسمان کاربر هم مطابقت دارد.



- پرسمان: ژیمناستیک
- تیتر خبر: خیرخواه: برخی به دنبال فلج کردن ژیمناستیک هستند/ با بایکوت فدراسیون موفقیتها بیشتر شد متن خبر حاوی پرسمان:

سرمربی تیم ملی ژیمناستیک در خصوص عدم اعزام به مسابقات جهانی گفت: برخی دنبال این هستند تا ژیمناستیک را فلج کنند و حرف خود را به کرسی بنشانند.

رضا خیرخواه در گفتوگو با خبرنگار ورزشی خبرگزاری فارس، درباره عدم اعزام تیم ملی ژیمناستیک هنری به مسابقات جهانی ژاپن اظهار داشت: واقعا بابت این موضوع متاسفم. فدراسیون از مدتها قبل درخواست اعزام تیم ملی ژیمناستیک را به شورای برون مرزی وزارت ورزش داده بود و بارها این مسئله را پیگیری کرد اما متاسفانه این شورا هیچ پاسخی نداد.

مهدی احمدکهنی بهترین نتیجه تاریخ ژیمناستیک کشور را کسب کرده بود

برخی به دنبال این هستند که به هر نحو ممکن ژیمناستیک ایران را فلج کنند تا از این طریق حرفشان را به کرسی بنشانند.

سرمربی تیم ملی ژیمناستیک هنری گفت: از آقای سجادی خواهش می کنم صدای اهالی ژیمناستیک کشور را بشنود. اگر قرار است مجمع فدراسیون تصمیم گیرنده باشد، اجازه بدهند که همین مجمع برای فدراسیون تصمیم بگیرد. برخی تصمیمات باعث شده بسیاری از مدیران موفقی که داشتیم عطای حضور در ورزش را به لقایش بسپارند. شاید اگر الان ریاست فدراسیون در اختیار کسی بود که برایش فرقی نمی کرد رئیس کدام فدراسیون است، اینگونه اقدامات علیه ژیمناستیک کشور انجام نمی شد اما خدا جای حق نشسته است. دیدیم که هر چه بیشتر ژیمناستیک را بایکوت کردند، موفقیتها بیشتر شد.

چنین اتفاق تاریخی نشان داد که ژیمناستیک ایران به کلی متحول شده است

ژیمناستیک ثابت کرد به لحاظ مدال آوری می تواند جزو ۱۰ رشته ورزشی کشورمان باشد.

سرمربی تیم ملی ژیمناستیک هنری خاطرنشان کرد: شنیدهام که برخی از نمایندگان مجلس شورای اسلامی قصد دارند به اقداماتی که علیه فدراسیون شده، ورود کنند. امیدوارم این اتفاق هر چه سریعتر رخ دهد. کاری نمانده که علیه فدراسیون نشود. از حذف ژیمناستیک از المپیاد استعدادیابی و جلوگیری از اعزام ملیپوشان به مسابقات جهانی تا قطع بودجه فدراسیون.

ارتباط خبر به پرسمان: سند مربوط به چالشهایی در حوزه ورزش ژیمناستیک است که مطابقت کامل با پرسمان دارد.



- پرسمان: واکسن آسترازنکا
- تیتر خبر: بررسی ساماندهی و بهبود وضعیت مالی بانک رفاه کارگران در جلسه ستاد هماهنگی اقتصادی دولت متن خبر حاوی پرسمان:

برآوردی که در زمینه احتمال لخته شدن خون در واکسن آسترازنکا انجام شده، نشان میدهد که این مورد در یک نفر از هر ۱۰۰ هزار جوان دیده شده است.

«گاردین»، روزنامه انگلیسی در بخش دانش خود نوشته است: امکان ایجاد لخته خون بر اثر تزریق واکسن آسترازنکا تا چه حد است؟

ارتباط خبر به پرسمان: سند مربوط به بررسی خطرات احتمالی تزریق واکسن آسترازنکا بوده که با پرسمان کاربر هم مطابقت دارد.

