به نام خدا

پروژه اول بازیابی اطلاعات

محمد جواد زندیه ۹۸۳۱۰۳۲

۵ اردیبهشت ۱۴۰۰

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر

۱. با ذکر مثال شرح دهید که در گام پیش‌پردازش چه عملیاتی انجام داده‌اید. همچنین دلیل انجام هر پردازش را ذکر کنید.

در این گام عملیات های زیر به ترتیب روی داکیومنت ها صورت گرفته است (روی محتوای content هر خبر):

۱. نرمال سازی: محتوای content هر خبر را در این بخش نرمال می‌کنیم. ۱- اعداد را از فارسی به انگلیسی تبدیل کنیم تا جست و جو در آن در مراحل بعدی بتواند صورت گیرد. ۲- بین علائم نگارشی و کلمات فاصله (space) قرار میدهد که در هنگام استخراج توکن کار به درستی صورت گیرد. ۳- فاصله ها را به نیم فاصله تبدیل می‌کند.

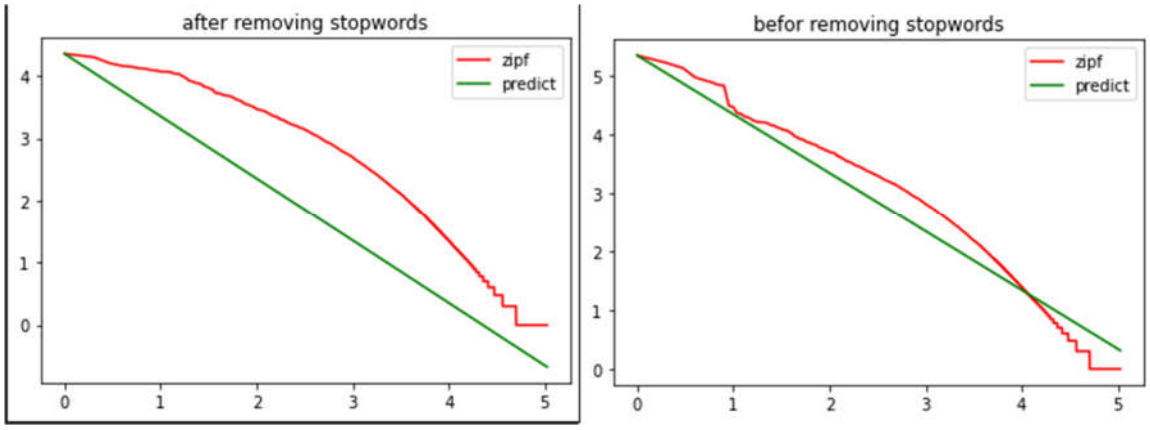
۲. استخراج توکن : در این بخش توکن های هر داکیومنت را پس از نرمال سازی استخراج می‌کنیم تا بتوان در مراحل بعد روی آنها پیش پردازش های دیگری انجام داد و در پاسخ به کوئری ها هم تطابق توکن های کوئری با آنها را بررسی کرد.

۳. حذف stop word ها: حذف کلمات پرتکرار مثل حروف اضافه. این کار به این دلیل صورت می‌گیرد که تکرار این کلمات در داکیومنت زیاد است و موجب می‌شود حجم positional index بسیار بزرگ شود و بعلاوه جست و در آن هم طولانی می‌شود و ساخت آن هم بیشتر طول خواهد کشید.

۴. ریشه یابی: ریشه یابی به منظور تطبیق کلمات هم ریشه است (البته شاید در زبان عربی خیلی مناسب نباشد). به عنوان مثال فعلا "رفتم" را به رفت&رو تبدیل می کند و معادل آنها است. پس کلمات هم ریشه را در واقع مشابه در نظر می‌گیرد.

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

قبل از حذف stopword ها، نمودار ما بیشتر به مقدار پیش بینی قانون zipf نزدیک بود در حالیکه بعد از حذف آنها همچنان قانون میتواند تا مقدار خوبی نمودار اصلی را پیش بینی کند. علت برقراری قانونن Zipf این است که نمودار ما به نمودار پیش بینی نزدیک است (البته در حالت بعد از حذف stopword ها کمتر شباهت وجود دارد).



۳. صحت قانون heaps را در دو حالت قبل و بعد از ریشه‌یابی بررسی کنید. برای بررسی این قانون لازم است با استفاده از اندازه‌ی واژه‌نامه و تعداد توکن‌ها در ۵۰۰، ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ سند اول، اندازه‌ی واژه‌نامه مربوط به کل اسناد تخمین زده شود. در نهایت اندازه‌ی واقعی واژه‌نامه و اندازه‌ی تخمینی در هر دو حالت مقایسه و تحلیل شود. آیا در هر دو حالت قانون برقرار است؟ چرا؟ (رسم نمودار برای هر حالت الزامی است.)

قانون heap’s در هر دو حالت برقرار است زیرا حالت خطی بودن نمودار برقرار است. دقت در هر دو حالت در حدود ۷۳ درصد می باشد.

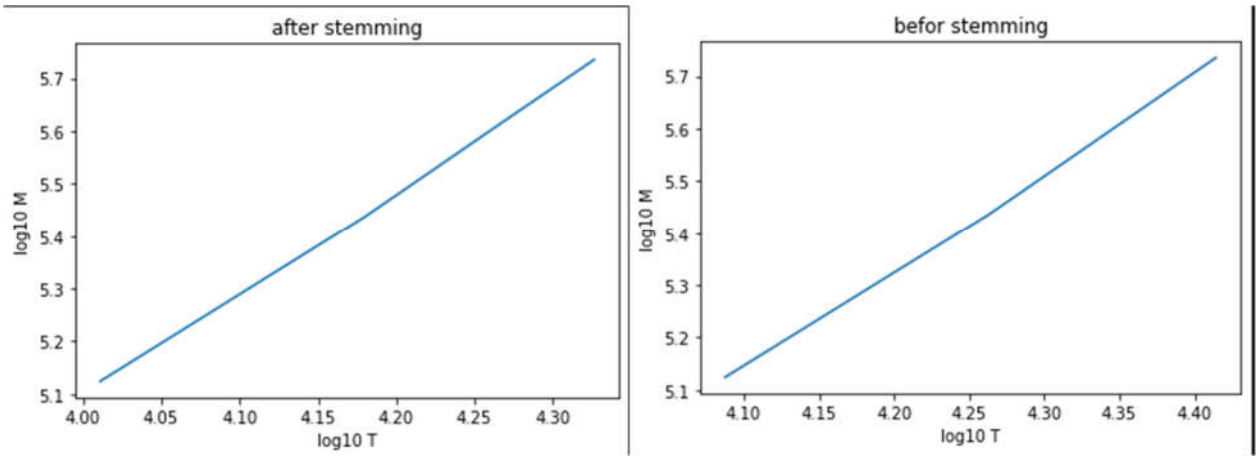
```
(befor stemming)
b = 0.5345475740357929 k = 22.3041540772557
actual vocab size = 105839
predicted vocab size = 81097
accuracy=  0.7662298396621283


(after stemming)
b = 0.516209223836649 k = 23.220813574661733
actual vocab size = 87354
predicted vocab size = 63730
accuracy=  0.7295601804153216
```

مقدار b و k را از روی مقادیر طول واژه نامه و تعداد توکن ها با ۵۰۰ و ۲۰۰۰ داکیومنت بدست می آوریم (نحوه بدست آوردن در کد به صورت کامنت مشخص شده است)

با بدست آوردن این دو مقدار می توان با داشتن تعداد توکن ها، طول واژه نامه را بدست آورد (همچنین دقت هم حاصل تقسیم مقدار واقعی بر مقدار پیش بینی شده است)

خطی بودن نمودار ها نیز نشان دهنده این است که قانون heap’s برقرار است.



۴. حداقل سه مورد از مواردی که در ریشه‌یابی با چالش روبرو بودید را ذکر کنید. (بطور مثال کلماتی که نیازی به ریشه‌یابی ندارند اما طبق روند ریشه‌یابی از دست می‌روند.)

۱. گاهی در پیدا کردن پیشوند ها دچار اشتباه می‌شود و مثلا حرف ب ابتدایی کلمات را عنوان پیشوند درنظر می‌گیرد و برعکس. مثلا کلمه "ببر" را نمی‌تواند تشخیص دهد که حیوان ببر مورد نظر است و یا فعل امر بِبَر و یا فعل امر بِبُر. در کل یعنی به اعراب ها دقت نمی‌کند.

۲. خیلی دقت بالایی در ریشه یابی ندارد و دایره پوشش آن کم است مثلا باید کلمات "رفتن"، "رفت"، "رفتم"، "خواهم رفت" و ... همه را مثلا معادل با "رو" بگیرد اما عملا چنین کاری را نمی‌کند و یه سری از کلمات خاص را پوشش می‌دهد.

۳. در حذف پسوند ها هم به خوبی عمل نمی‌کند. مثلا انتظار داریم که "شیراز" و "شیرازی" از یک ریشه باشند و در جست و جو هر دو مشابه باشند درحالی که هیچ عملیات روی آنها انجام نمی دهد.

۵. پاسخ به پرسمان در حالت‌های زیر:

الف) یک پرسمان از کلمات ساده و متداول (مانند تحریم‌های آمریکا علیه ایران، در نتایج بازیابی‌شده انتظار می‌رود اسنادی که کلمات تحریم، آمریکا، علیه و ایران را دارند در بالای لیست و اسنادی که برخی از کلمات را ندارند در رتبه‌های پایین‌تر لیست قرار داشته باشند.)

خروجی در فایل output1 در کنار پروژه می‌باشد.

نکته: در هر بخش فقط تیکه های کمی از متحوا قرار داده شده است تا بتوان صحت کلی را بررسی کرد و به طور کامل تمامی بخش های همخوانی با کوئری قرار داده نشده است.

```
Query:  تحریم‌های آمریکا علیه ایران
----------------------------------------------------

Title:  آمریکا درآمد کشورها را مسدود می‌کند و ژست انسان‌دوستی می‌گیرد

URL:  https://www.farsnews.ir/news/14001121000750/آمریکا-درآمد-کشورها-را-مسدود-می‌کند-و-ژست-انسان‌دوستی-می‌گیرد

Content:
سخن 29 بهمن رهبری خط اصلی دشمن یادآور هدف اصلی تحریم تنگنا اقتصادی هدف قرار اذهان فکر عمومی تمسک انواع
----------------------------------------------------

Title:  جلیلی: آمریکا را از تحریم‌کردن پشیمان کنیم/ دانشجویان می‌توانند حصار تحریم را بشکنند

URL:  https://www.farsnews.ir/news/14000916000573/جلیلی-آمریکا-را-از-تحریم‌کردن-پشیمان-کنیم-دانشجویان-می‌توانند-حصار

Content:
قرار شهادت رسید&رس جلیلی تاکید دشمن کشور عرصه علمی فناوری تحریم کرد&کن تحریم شکست&شکن جنگ دانشجو جنبش دانشجویی نقش ویژه
----------------------------------------------------

Title:  مخبر: دولت برای حمایت از سیاست‌های جمعیت اهتمام جدی دارد

URL:  https://www.farsnews.ir/news/14000808000371/مخبر-دولت-برای-حمایت-از-سیاست‌های-جمعیت-اهتمام-جدی-دارد

Content:
انقلاب اسلامی ابعاد جمله اقتصادی فرهنگی نظامی ملت ایران فشار تحریم قرار فضل الهی نتوانسته اهداف دست جمعیت تبلیغات اقدامات
----------------------------------------------------

Title:  یادداشت| ماسک‌هایی که بوی رفاقت نمی‌دهد

URL:  https://www.farsnews.ir/news/14000815000496/یادداشت|-ماسک‌هایی-که-بوی-رفاقت-نمی‌دهد

Content:
توجه تقدیر مسئول مدیر جمهوری اسلامی ایران چشم استوار ایران تحریم ظالمانه اقدامات مسئول کشور غلبه ویروس کرونا مشاهده نکته
----------------------------------------------------

Title:  زیگزاگ اصلاح‌طلبان مقابل آمریکا/ دیروز مزاحم برجام، امروز طرف اصلی مذاکرات

URL:  https://www.farsnews.ir/news/14000812000209/زیگزاگ-اصلاح‌طلبان-مقابل-آمریکا-دیروز-مزاحم-برجام-امروز-طرف-اصلی

Content:
آمریکایی تن میز مذاکره ترک مذاکره هسته‌ای اروپایی هدف رفع تحریم اصاله برجام سر امّا اصلاح‌طلبان انتخابات مجلس ریاست‌جمهوری شورا
----------------------------------------------------
```

ب) یک پرسمان با عملگر NOT (مانند تحریم‌های آمریکا ! ایران، انتظار می‌رود اسنادی که شامل دو کلمه تحریم و آمریکا هستند اما کلمه‌ی ایران را ندارند در نتایج بازیابی‌شده وجود داشته باشند.)

خروجی در فایل output2 در کنار فایل پروژه قرار دارد.

```
Query: تحریم های آمریکا ! ایران
--------------------------------------------------

Title: عدالت آمریکا پوشالی و دروغ است

URL: https://www.farsnews.ir/news/14001020000855/عدالت-آمریکا-پوشالی-و-دروغ-است

Content:
چالش لازم مسیر صبر ایستادگی هرچند دشمن آزار اذیت جنگ تحریم معاون پژوهشی مرکز اسلامی انگلیس آمریکا گفت&گو منافع اقتضا

--------------------------------------------------

Title: محو رژیم صهیونیستی از آرمان‌های نظام اسلامی حذف نشده است

URL: https://www.farsnews.ir/news/14001222000379/محو-رژیم-صهیونیستی-از-آرمان‌های-نظام-اسلامی-حذف-نشده-است

Content:
پر نقش مجلس نادیده گرفت&گیر توان مجلس تصریح آمریکا لغو تحریم تصویب کنگره آزادسازی تبادلات ارزی فروش نفت پذیرفت&پذیر مذاکره

--------------------------------------------------

Title: لغو تحریم‌ها تنها ضرورت احیای برجام است

URL: https://www.farsnews.ir/news/14001110000419/لغو-تحریم‌ها-تنها-ضرورت-احیای-برجام-است

Content:
دهه فجر بیان سیاست منطقه‌ای آیت‌الله رئیس سرآغاز خوبی خنثی‌سازی تحریم توسعه رابطه کشور همسایه اقدام راهبردی انتظار دولت جدید

--------------------------------------------------

Title: مسائل اقتصادی کشور را معطل نتیجه مذاکرات نمی‌کنیم

URL: https://www.farsnews.ir/news/14001203000383/مسائل-اقتصادی-کشور-را-معطل-نتیجه-مذاکرات-نمی‌کنیم

Content:
تکیه توان داخلی ایجاد ثبات رشد اقتصادی کشور مسیر خنثی‌سازی تحریم گام بردار قوی اثر تحریم فراهم کوچکی‌نژاد خاطرنشان رفع

--------------------------------------------------

Title: ادامه تحریم‌های سیاسی علیه المپیک پکن/ژاپن هم به صف منتقدان پیوست

URL: https://www.farsnews.ir/news/14001003000306/ادامه-تحریم‌های-سیاسی-علیه-المپیک-پکن-ژاپن-هم-به-صف-منتقدان-پیوست

Content:
دعوت کمیته بین‌المللی المپیک پارالمپیک حضور ماتسونو پاسخ سوال تحریم دیپلماتیک اصطلاح خاصی توصیف نحوه حضور کرد&کن تصمیم ژاپن

--------------------------------------------------
```

پ) یک پرسمان با عملگر عبارت (مانند "کنگره ضدتروریست"، انتظار می‌رود اسنادی که شامل عبارت کنگره ضدتروریست در نتایج بازیابی‌شده وجود داشته باشند؛ بعبارت دیگر موقیت مکانی کلمات در این حالت مهم است.)

فقط یک سند بازیابی شد. فایل output3 مشاهده شود.

```
Query:  "کنگره ضدتروریست"
--------------------------------------------------

 Title:  توضیحات یک منبع آگاه درباره وقفه مذاکرات وین

 URL:  https://www.farsnews.ir/news/14001222000450/توضیحات-یک-منبع-آگاه-درباره-وقفه-مذاکرات-وین

 Content:
کاربرد نظامی نیزبودند گسترش تاریخ 6 نوامبر سال 1987م نماینده کنگره واکنش خرید&خر نفت اداره انرژی آمریکا ایران لایحه مجلس

--------------------------------------------------
```

ت) یک پرسمان پیچیده (مانند "تحریم هسته‌ای" آمریکا ! ایران، انتظار می‌رود اسنادی که شامل عبارت تحریم هسته‌ای و کلمه‌ی آمریکا هستند اما کلمه‌ی ایران را ندارند در نتایج بازیابی‌شده وجود داشته باشد.)

هیچ سندی بازیابی نشد.

```
Query:  "تحریم هسته‌ای" آمریکا ! ایران"
--------------------------------------------------
no result
```

به جای آن کوئری زیر را بررسی می کنیم: (خروجی در output4 قرار دارد)

"روزنامه پرتغالی" طارمی ! گل

```
Query:  "روزنامه پرتغالی" طارمی !گل"
--------------------------------------------------

 Title:  طارمی برای گل زدن به ترکیب پورتو بازخواهد گشت+عکس

 URL:  https://www.farsnews.ir/news/14001201000727/طارمی-برای-گل-زدن-به-ترکیب-پورتو-بازخواهد-گشت-عکس

 Content:
ترکیب اصلی پورتو بازخواهد گشت اتفاق عجیب بازی لاتزیو رخ روزنامه پرتغالی مدعی ستاره کشور دلیل بازی نکردن دیدار لاتزیو


 Content:
هفته بیست لیگ پرتغال پورتو موریرا میهمان موریرنسه بازی یار طارم اهمیت پورتو خواست&خواه فصل 2018 - 2019 فصل 2020

--------------------------------------------------
```

ث) یک پرسمان کلمات نادر (مانند اورشلیم ! صهیونیست، خروجی مورد انتظار این قسمت مشابه با قسمت ب می‌باشد با این تفاوت که کلمات استفاده شده در پرسمان از کلمات نادر هستند.)

این عبارت هم بازیابی نشد.

```
Query:  اورشلیم ! صهیونیست
--------------------------------------------------
no result
```