گزارش فاز اول پروژه بازیابی اطلاعات

روژینا کاشفی ۹۸۳۱۱۱۸ – لیلی برکتین ۹۸۳۱۰۷۴؟

سوال اول

در گام پیش‌پردازش از سه عملیات استفاده شده است:

1- Normalization
در نرمال سازی واژه‌ها به شکل استاندارد تبدیل می‌شوند برای مثال تشدید و تنوین حذف می‌شود و حروف عربی به فارسی تبدیل می‌شود و نیم‌فاصله اصلاح می‌شوند.

2- Tokenization
در این مرحله واژه‌ها را به وسیله فاصله از هم جدا می‌کنیم.

3- Stemming
در این مرحله ریشه کلمات را به‌جای شکل‌های مختلف کلمات نگه‌داری می‌کنیم.

4- Removing Stop words
در مرحله حذف ایست واژهها، واژههاي پر تکرار که عموما معناي خاصی ندارند حذف می شود .براي حذف ایست واژهها از لیست ایست واژههاي کتابخانه هضم استفاده شده است.

به عنوان مثال متن خبر اول به صورت :

به گزارش خبرگزاری فارس، کنفدراسیون فوتبال آسیا (AFC) در نامه ای رسمی به فدراسیون فوتبال ایران و باشگاه گیتی پسند زمان قرعه کشی جام باشگاه های فوتسال آسیا را رسماً اعلام کرد. بر این اساس 25 فروردین ماه 1401 مراسم قرعه کشی جام باشگاه های فوتسال آسیا در مالزی برگزار می شود. باشگاه گیتی پسند بعنوان قهرمان فوتسال ایران در سال 1400 به این مسابقات راه پیدا کرده است. پیش از این گیتی پسند تجربه 3 دوره حضور در جام باشگاه های فوتسال آسیا را داشته که هر سه دوره به فینال مسابقات راه پیدا کرده و یک عنوان قهرمانی و دو مقام دومی بدست آورده است. انتهای پیام/

- **Normalization**

تغییرات پس از normalize کردن را با رنگ زرد مشاهده می‌کنیم. در نرمالایز کردن نیم‌فاصله‌ها درست شده است. نقطه انتها هر کلمه را از کلمه نهایی جدا شده است.

به گزارش خبرگزاری فارس، کنفدراسیون فوتبال آسیا (AFC) در نامه‌ای رسمی به فدراسیون فوتبال ایران و باشگاه گیتی پسند زمان قرعه کشی جام باشگاه‌های فوتسال آسیا را رسما اعلام کرد . بر این اساس 25 فروردین ماه 1401 مراسم قرعه کشی جام باشگاه‌های فوتسال آسیا در مالزی برگزار می‌شود . باشگاه گیتی پسند بعنوان قهرمان فوتسال ایران در سال 1400 به این مسابقات راه پیدا کرده است . پیش از این گیتی پسند تجربه 3 دوره حضور در جام باشگاه‌های فوتسال آسیا را داشته که هر سه دوره به فینال مسابقات راه پیدا کرده و یک عنوان قهرمانی و دو مقام دومی بدست آورده است . انتهای پیام/

- **TOKENIZATION**

در این مرحله داکیومنت کلمه کلمه می‌شود و سپس می‌توانیم بررسی کنیم ایا هر کلمه stopword هست یا نه و کلمات را به ریشه خود برگردانیم.

['به', 'گزارش', 'خبرگزاری', 'فارس', '،', 'کنفدراسیون', 'فوتبال', 'آسیا', '(', 'AFC', ')', 'در', 'نامه\u200cای', 'رسمی', 'به', 'فدراسیون', 'فوتبال', 'ایران', 'و', 'باشگاه', 'گیتی', 'پسند', 'زمان', 'قرعه', 'کشی', 'جام', 'باشگاه\u200cهای', 'فوتسال', 'آسیا', 'را', 'رسما', 'اعلام', 'کرد', '.', 'بر', 'این', 'اساس', '25', 'فروردین', 'ماه', '1401', 'مراسم', 'قرعه', 'کشی', 'جام', 'باشگاه\u200cهای', 'فوتسال', 'آسیا', 'در', 'مالزی', 'برگزار', 'می\u200cشود', '.', 'باشگاه', 'گیتی', 'پسند', 'بعنوان', 'قهرمان', 'فوتسال', 'ایران', 'در', 'سال', '1400', 'به', 'این', 'مسابقات', 'راه', 'پیدا', 'کرده', 'است', '.', 'پیش', 'از', 'این', 'گیتی', 'پسند', 'تجربه', '3', 'دوره', 'حضور', 'در', 'جام', 'باشگاه\u200cهای', 'فوتسال', 'آسیا', 'را', 'داشته', 'که', 'هر', 'سه', 'دوره', 'به', 'فینال', 'مسابقات', 'راه', 'پیدا', 'کرده', 'و', 'یک', 'عنوان', 'قهرمانی', 'و', 'دو', 'مقام', 'دومی', 'بدست', 'آورده', 'است', '.', 'انتهای', 'پیام/', ' ']

- **STOP WORDS & STEMMING**

مشاهده می‌کنیم که ویرگول یک stop word است که حذف شده است. یا کلمه هر نیز یک کلمه پرتکرار است که حذف شده است.
در مرحله ریشه‌یابی افعال به صورت ریشه نوشته می‌شود.

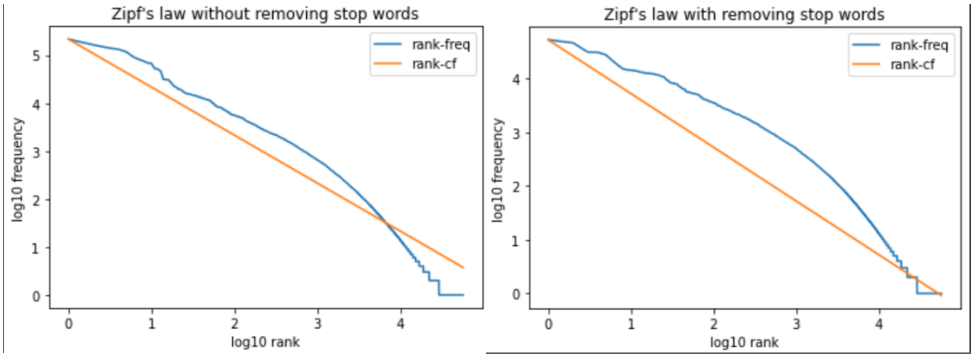
['گزارش', 'خبرگزاری', 'فارس', 'کنفدراسیون', 'فوتبال', 'آسیا', 'AFC', 'نامه', 'رسمی', 'فدراسیون', 'فوتبال', 'ایران', 'باشگاه', 'گیتی', 'پسند', 'زمان', 'قرعه', 'کشید&کش', 'جام', 'باشگاه', 'فوتسال', 'آسیا', 'رسما', 'اعلام', 'اساس', '25', 'فروردین', 'ماه', '1401', 'مراسم', 'قرعه', 'کشید&کش', 'جام', 'باشگاه', 'فوتسال', 'آسیا', 'مالزی', 'برگزار', 'شد&شو', 'باشگاه', 'گیتی', 'پسند', 'بعنوان', 'قهرمان', 'فوتسال', 'ایران', 'سال', '1400', 'مسابقات', 'اس', 'گیتی', 'پسند', 'تجربه', '3', 'دوره', 'حضور', 'جام', 'باشگاه', 'فوتسال', 'آسیا', 'دوره', 'فینال', 'مسابقات', 'عنوان', 'قهرمانی', 'مقام', 'دومی', 'بدست', 'آورده', 'اس', 'انتهای', 'پیام']

سوال دوم

تابع نارنجی از رابطه zipfs و تابع آبی مقادیری که از شاخص بدست آمده را نشان می‌دهد.

هدف این تابع آن است که نشان دهد که هرچقدر فرکانس کمتر شود رنک آن بیشتر می‌شود و هر چقدر فرکانس بیشتر شود رنک آن کمتر می‌شود.

ایست واژه‌ها کلمات پرتکرار هستند . پرتکرارترین کلمه در شاخص قبل از حذف ایست واژه‌ها از پرتکرارترین کلمه در شاخص بعد از حذف ایست واژه‌ها، تکرار بیشتري دارد پس همانطور که مشاهده می‌شود قانون zipf's قبل از حذف ایست واژه‌ها عملکرد دقیق‌تري دارد .

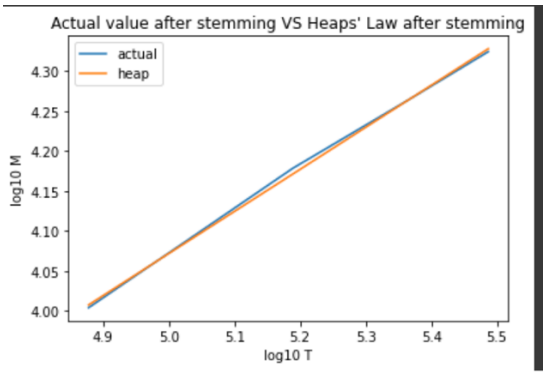


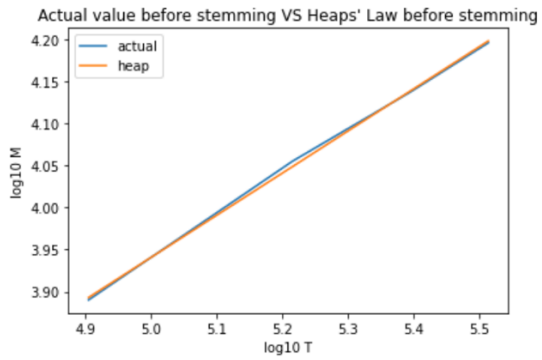
سوال سوم

ابتدا اندازه collectionو vocabulary را براي تعداد مشخصی سند بدست می‌آوریم .

```
After stemming:
First 500 Docs
vocab size :  7767   collection size:  80287
First 1000 Docs
vocab size :  11351   collection size:  164475
First 1500 Docs
vocab size :  13705   collection size:  248561
First 2000 Docs
vocab size :  15679   collection size:  326937
----------------------------------
Before stemming
First 500 Docs
vocab size :  10093   collection size:  75476
First 1000 Docs
vocab size :  15082   collection size:  154565
First 1500 Docs
vocab size :  18421   collection size:  233325
First 2000 Docs
vocab size :  21099   collection size:  306798
```

همانطور که مشاهده می‌شود مقادیر بدست آمده از شاخص و مقادیر رابطه heaps تقریبا روي هم افتاده و قانون heaps براي هر دو حالت عملکرد خوبی دارم ولی بعد از ریشه‌یابی عملکرد دقیق‌تري دارد.





سوال چهارم

ابتدا از کتابخانه هضم استفاده کرده بودم که براي ريشه‌يابي مشکلات زيادي داشت؛ براي مثال بعضی کلمات مفرد را جمع تشخیص میداد (تهران را به تهر تبدیل می‌کرد).

بعضی کلمات که اسم هستند و در آخر آنها حرف م یا ي دارند را فعلی با ضمیر تلقی میکرد (سوم را به سو تبدیل می‌کرد

کتابخانه پارسیوار براي ريشه‌يابي عمکرد بهتري دارد اما نسبت به هضم کندتر است .

فعل مُر َدم و اسم مِردُم هر دو به مردم تبدیل می‌شود .

بعضی کلمات جمع مثل انتخابات مفرد نمی‌شود و همان انتخابات برگردانده می‌شود .

چند مثال دیگر از اشتباهات ريشه‌ياب پارسيوار:

سال‌ها همان سال‌ها می‌ماند

فعل کردي ريشه‌يابي نمی‌شود.

سوال پنجم

الف) پرسمان ساده

کوئری : تحریم‌های امریکا علیه ایران

باید به گونه‌ای باشد که همه کلمات در وبلاگ‌های مورد نظر باشد و انکه تعداد تکرار کلمات بیشتری دارد در رتبه بالاتری قرار گیرد. یکی از خروجی‌های را برای صحت عملکرد چک می‌کنیم. بعضی از اخبار مانند خبر زیر کاملا مرتبط با پرسمان کاربر است ولی در برخی موارد فقط این کلمات را در خبر دارد (ممکن است پراکنده باشد) و از نظر کلیت موضوع مرتبط نیست.

> گروه سیاسی خبرگزاری فارس، احد شیرزاد: به نظر می‌رسد مذاکرات وین به روزهای آخر خود نزدیک شده است و براساس اعلام برخی منابع تا رسیدن به توافق راه زیادی باقی نمانده و تنها نیاز است آمریکا به آنچه در برجام برای برداشتن تحریم‌ها وعده داده عمل کند و تضمین معتبری هم برای انجام این تعهدات داده شود.
>
> براساس اعلام مسئولان سیاست خارجه در کنار دریافت تضمین معتبر، برداشتن همه تحریم‌ها از نکات و مطالبات اصلی ایران در مذاکرات است. اما به نظر می‌رسد آمریکا در مقابل عمل به این تعهد برجامی مقاومت می‌کند.
>
> *** تحریم‌های آمریکا علیه ایران**
>
> آمریکا پس از تسخیر لانه جاسوسی آمریکا در تهران، «کارتر» در تاریخ ۸ نوامبر ۱۹۷۹م. با استناد به قانون «کنترل صدور تسلیحات نظامی»، کشتی حامل لوازم یدکی نظامی متعلق به ایران را توقیف کرد. ارزش این لوازم ۳۰۰ میلیون دلار بود. با اوجگیری کشمکش‌های سیاسی بر سر مسئله تصرف سفارت، دولت موقت، اعلام کرد تمام دارایی‌های خود را از بانک‌های آمریکا خارج خواهد کرد. کارتر با لحاظ این احتمال، در کشور شرایط اضطراری اعلام کرد و با استناد به قانون شرایط اضطراری اقتصاد بین‌الملل (IEEPA) و قانون شرایط اضطراری ملی (NEA)، با صدور دستور ویژه‌ی شماره‌ی (۱۲۱۷۰)، تمام دارایی‌های ایران در آمریکا را به تصرف خود درآورد.

ب) یک پرسمان با عملگر not

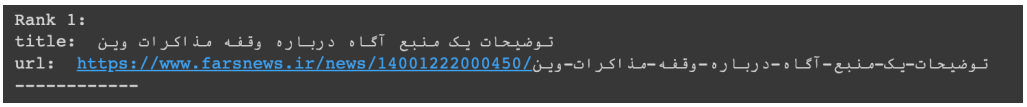
کوئری : تحریم‌های امریکا ! ایران

این کوئری کلمات تحریم و امریکا را دارد و ایران را ندارد که تا حد معقولی با کوئری کاربر برابر است.

تصمیم ژاپن برای این موضوع به دنبال تحریم دیپلماتیک المپیک زمستانی ۲۰۲۲ پکن از سوی آمریکا، استرالیا، انگلیس و کانادا به خاطر نقض حقوق بشر است. چین از ایالات متحده و سایر کشورها به دلیل نقض بی طرفی سیاسی مورد نیاز در روح منشور المپیک انتقاد کرده است.

پ) یک پرسمان با عملگر عبارت

کوئری : کنگره ضد تروریست



ریگان که نمی‌خواست کمتر از کنگره ضدتروریست جلوه کند، ۳هفته بعد با صدور دستور ویژه‌ی (۱۲۶۱۳) ورود هرگونه کالا و خدمات از ایران را ممنوع کرد. وی برای صدور این دستور به بند ۵۰۵ «قانون همکاری‌های بین‌المللی امنیتی و توسعه»، مصوب سال ۱۹۸۵م. استناد کرد. در واقع زمانی که ایران درگیر جنگ تحمیلی بود، ایالات متحده به تعریف سیاست‌های تحریمی پرداخت که بتواند بر نتیجه‌ی جنگ ایران و عراق تأثیری شگرف بگذارد و رقیب نوپای اسلامی خود را در جنگی نابرابر از میدان به در کند. اما پایان جنگ، تافته‌ی بافته‌ی آمریکا را ریش ریش کرد.

فقط یک خبر دقیقا عبارت کنگره ضدتروریست را دارد .این خبر به طور خاص در مورد کنگره ضد تروریست نیست ولی این عبارت را دارد .

ت) یک پرسمان پیچیده

کوئری : "تحریم هسته‌ای" امریکا ! ایران

مشاهده می‌کنیم کوئری فوق شامل phrase و کلمات ساده و عملگر not است و به همین علت پیچیده است.هیچ خبری نیست که عبارت تحریم هسته‌اي را داشته باشد ولی ایران را نداشته باشد .براي همين نتايج نامربوط است .

اگر کوئری : "تحریم هسته‌ای" ایران ! آمریکا

اولین خروجی کمی‌مرتبط است ولی بقیه نامرتبط است.

ث) یک پرسمان با کلمات نادر

براي پرسمان اورشلیم ! صهیونیست کلمه اورشلیم در شاخص وجود ندارد براي همین پرسمان غزه ! صهیونیست را جستجو میکنیم .غزه نیز کلمه نادر است .

اخبار بدست آ مده غزه را دارد ولی صهیونیست را ندارد .

```
Rank 1:
title:  سفر تیم فوتبال معلولان فلسطین به تهران
url:  https://www.farsnews.ir/news/14001212000749/سفر-تیم-فوتبال-معلولان-فلسطین-به-تهران
------------
Rank 2:
title:  بیانیه بسیج ورزشکاران در محکومیت اقدام شرم‌آور سرمربی تیم ملی امید
url:  https://www.farsnews.ir/news/14000929000204/بیانیه-بسیج-ورزشکاران-در-محکومیت-اقدام-شرم آور-سرمربی-تیم-ملی-امید
------------
Rank 3:
title:  اقدام تحسین برانگیز سرمربی الجزایر در حمایت از فلسطین/جام قهرمانی به مردم غزه اهدا شد
url:  https://www.farsnews.ir/news/14000928000274/ام-تحسین-برانگیز-سرمربی-الجزایر-در-حمایت-از-فلسطین-جام-قهرمانی-به
------------
Rank 4:
title:  اردوغان علاوه بر آرمان فلسطین بر باورهای مردم ترکیه هم پا گذاشت
url:  https://www.farsnews.ir/news/14001220000842/اردوغان-علاوه-بر-آرمان-فلسطین-بر-باورهای-مردم-ترکیه-هم-پا-گذاشت
------------
Rank 5:
title:  نگاه دوگانه برخی جریان‌ها به یمن و اوکراین نشانه سرسپردگی‌شان به آمریکاست
url:  https://www.farsnews.ir/news/14001209000540/نگاه-دوگانه-برخی-جریان ها-به-یمن-و-اوکراین-نشانه-سرسپردگی‌شان-به
------------
```

تیم ملی فوتبال معلولان قطع عضو فلسطین فوریه گذشته همزمان با روز جهانی معلولان در غزه تشکیل شد. این تیم امیدوار است با درخشش در مسابقات غرب آسیا در ترکیه بتواند جواز حضور در مسابقات جام جهانی فوتبال معلولان که قرار است اکتبر آینده در ترکیه برگزار شود، را به دست بیاورد.تیم فوتبال معلولان قطع عضو فلسطین قرار است در گروه غرب آسیا با عراق، ایران، هند و ازبکستان به رقابت بپردازد.