**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.

1. Tokenization

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

1. Stemming

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

1. 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',

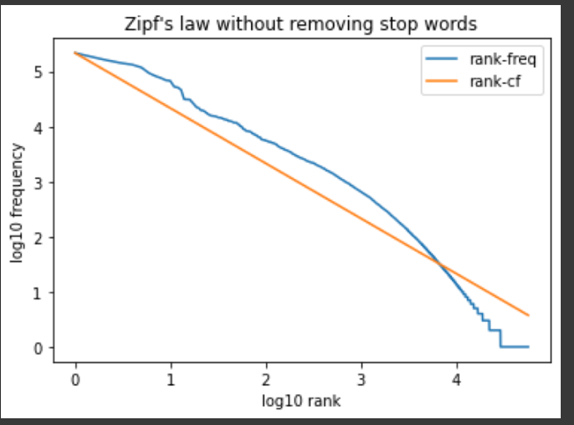
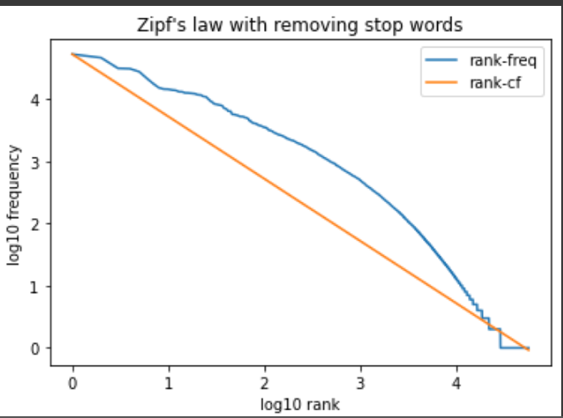
'پسند','تجربه', '3', 'دوره', 'حضور', 'جام', 'باشگاه', 'فوتسال', 'آسیا', 'دوره', 'فینال', 'مسابقات', 'عنوان', 'قهرمانی', 'مقام', 'دومی', 'بدست', 'آورده', 'اس', 'انتهای', 'پیام' [

**سوال دوم**

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

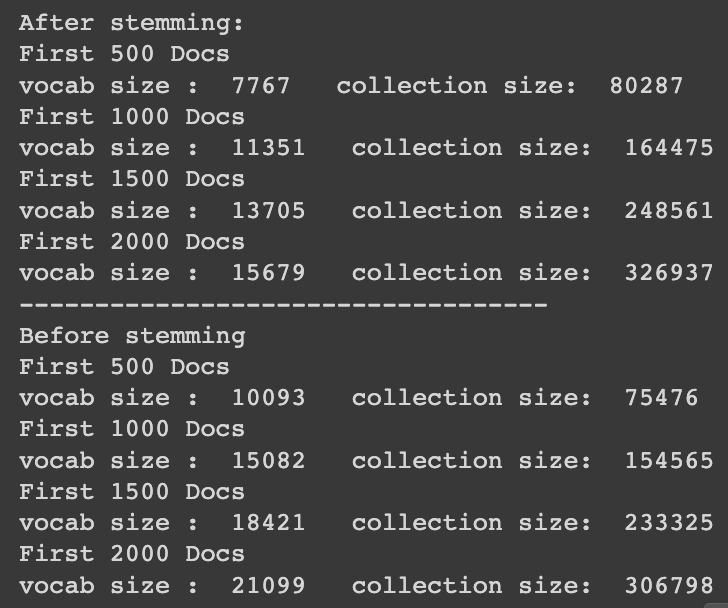
هدف این تابع آن است که نشان دهد که هرچقدر فرکانس کمتر شود رنک آن بیشتر می‌شود و هر چقدر فرکانس بیشتر شود رنک آن کمتر می‌شود.

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.

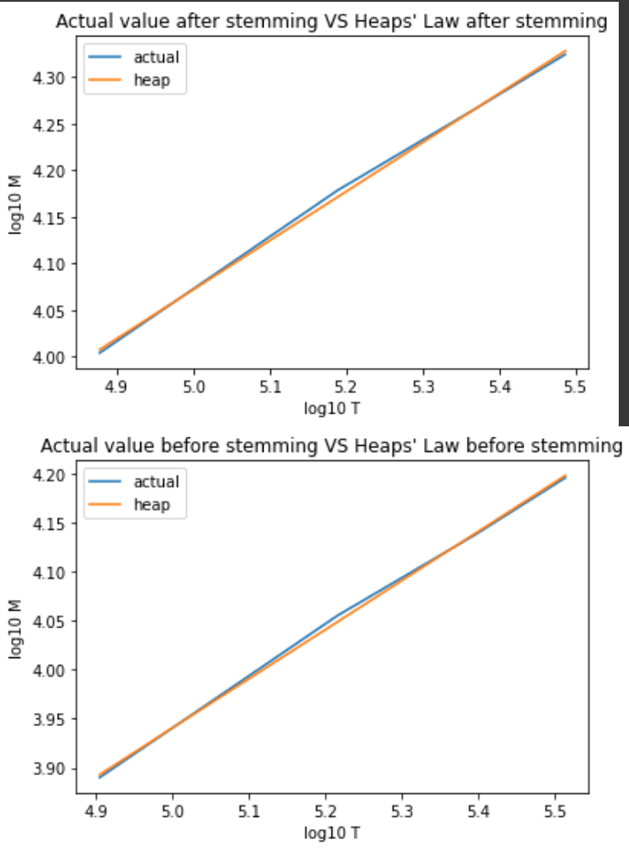


**سوال سوم**

ابتدا اندازه collection و vocabulary را براي تعداد مشخصی سند بدست می‌آوریم.



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 ریشه‌یابی مشکلات زیادي 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 ریشه‌یابی عمکرد بهتري دارد اما نسبت به هضم کندتر است.

فعل ُمر َدم و اسم َمردُم هر دو به مردم تبدیل می‌شود.

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

چند مثال دیگر از اشتباهات ریشه‌یاب پارسیوار:

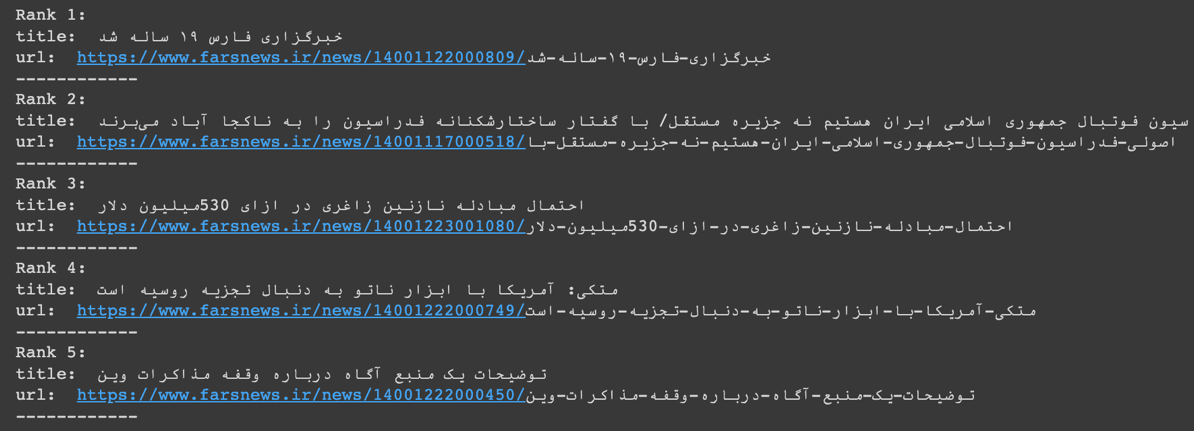
سال‌ها همان سال‌ها می‌ماند

فعل کردي ریشه‌یابی نمی‌شود.

**سوال پنجم**

**الف) پرسمان ساده**

کوئری : تحریم‌های امریکا علیه ایران



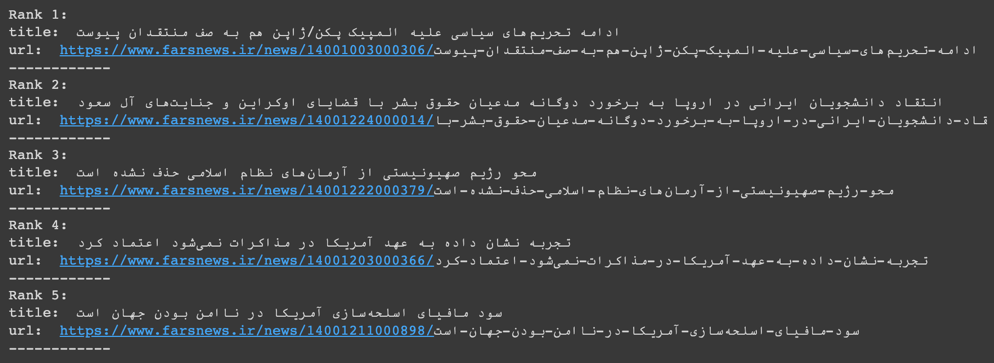
باید به گونه‌ای باشد که همه کلمات در وبلاگ‌های مورد نظر باشد و انکه تعداد تکرار کلمات بیشتری دارد در رتبه بالاتری قرار گیرد. یکی از خروجی‌های را برای صحت عملکرد چک می‌کنیم. بعضی از اخبار مانند خبر زیر کاملا مرتبط با پرسمان کاربر است ولی در برخی موارد فقط این کلمات را در خبر دارد ( ممکن است پراکنده باشد)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

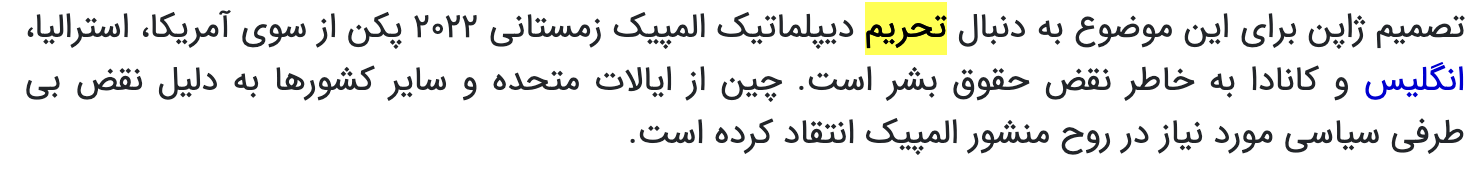


**ب) یک پرسمان با عملگر not**

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.

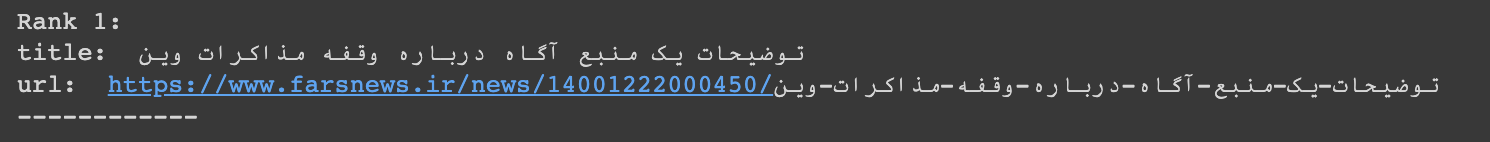


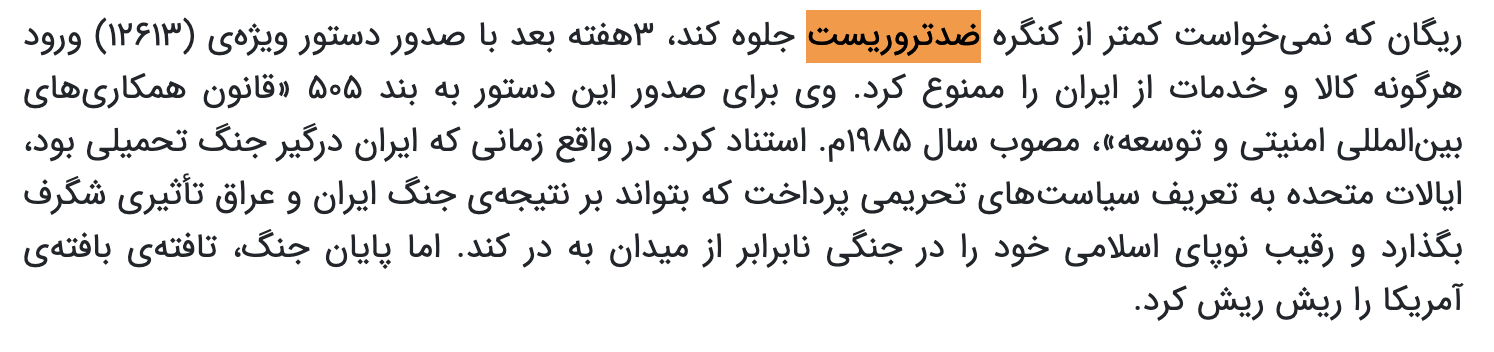




**پ) یک پرسمان با عملگر عبارت**

کوئری : کنگره ضد تروریست





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 همین نتایج 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.

