دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکدهٔ مهندسی کامپیوتر

درس بازیابی اطلاعات

گزارشکار پروژه

موژان میرجلیلی

فرحان کیهان

استاد

دکتر نیک‌آبادی

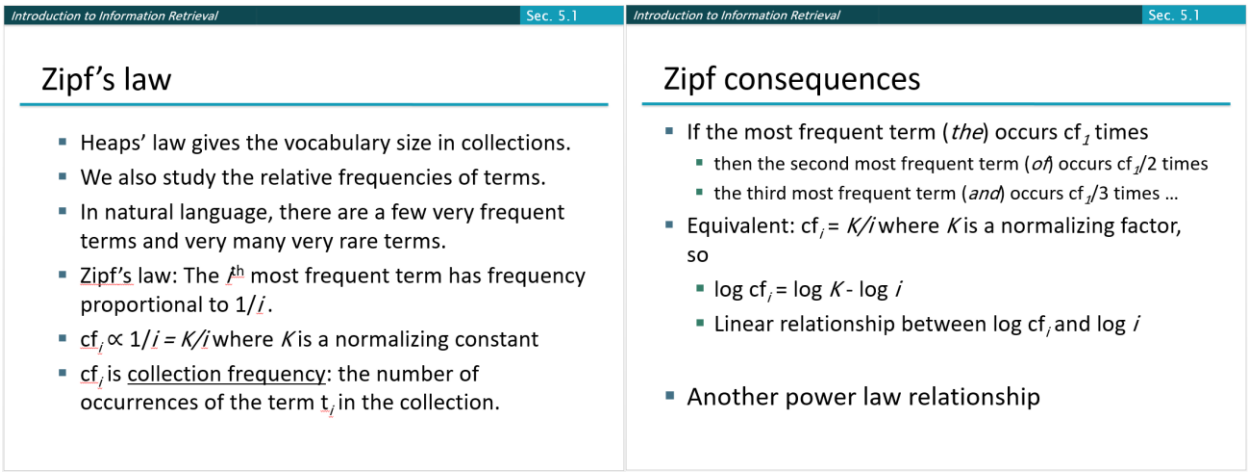
زمستان ۱۴۰۱

فاز ۱

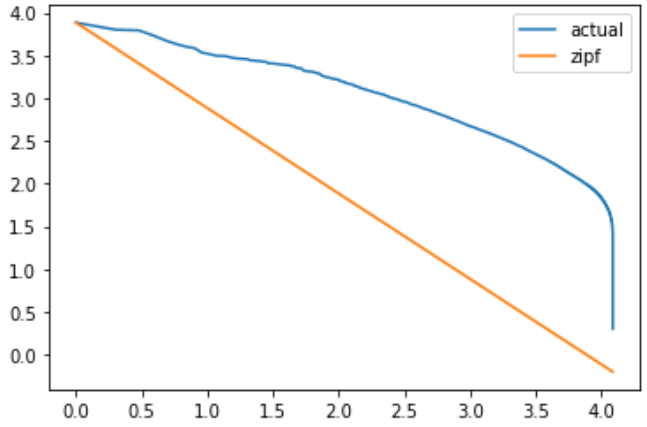
۱) با ذکر مثال شرح دهید که در گام پیش‌پردازش چه عملیاتی انجام داده‌اید. همچنین دلیل انجام هر پردازش را ذکر کنید.

در پیش‌پردازش ابتدا دیتافریم را خواندیم و transpose کردیم تا به فرمت مناسب (هر ستون یک فیچر و هر سطر یک داده) در بیاید. سپس normalize می‌کنیم، در این بخش نیم فاصله‌ها اصلاح می‌شوند و همه کلمات به یک حالت درمی‌آیند. برای مثال "تحریم ها"، "تحریم‌ها" می‌شود. در مرحله بعد یک سری موارد از متن حذف می‌شوند ('"~_*&^٪.$#@?/.<> ,\":;{}[]-()' و '\ ,'\u200d', '\u200d', '\u200d', '\u200f','\u200e', 'xad\n ,'\u200c'). برای مثال هر کدام از این موارد باشند بعد از این مرحله حذف می‌شوند. مرحله بعد tokenization است که کلمات یک متن به صورت کلمه کلمه و یک لیست در می‌آیند برای مثال عبارت تحریم آمریکا می‌شود یک لیست شامل کلمات تحریم و آمریکا. سپس با استفاده از کتابخانه stopwords از hazm، کلماتی نظیر و، در، از، را و غیره را حذف می‌کنیم که ما در اینجا برای عملکرد بهتر، های و ها هم به این لیست اضافه کرده و از توکن‌ها حذف می‌کنیم. مرحله آخر هم stemming است که کلمات هم‌ریشه و مشابه به یک حالت درمی‌آیند برای مثال خبرگزاری، خبرگزار می‌شود. همه موارد ذکر شده جز punctuation list با استفاده از hazm اجرا شدند.

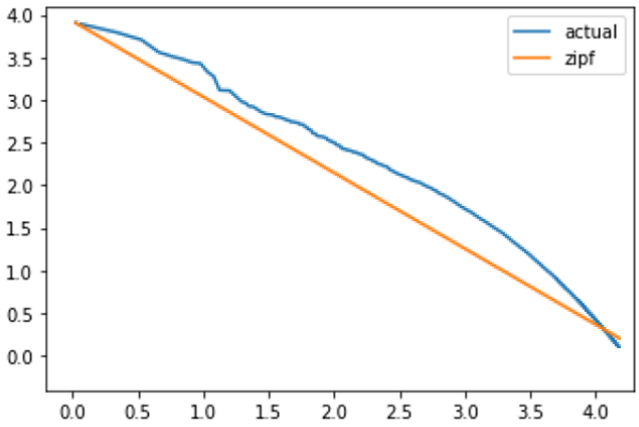
۲) صحت قانون Zipf را در دو حالت قبل و بعد از حذف کلمات پرتکرار از واژه‌نامه بررسی کنید. رسم نمودار برای هر حالت الزامی است. در صورت برقراری/ عدم برقراری این قانون در هر حالت، علت را شرح دهید.



قبل:

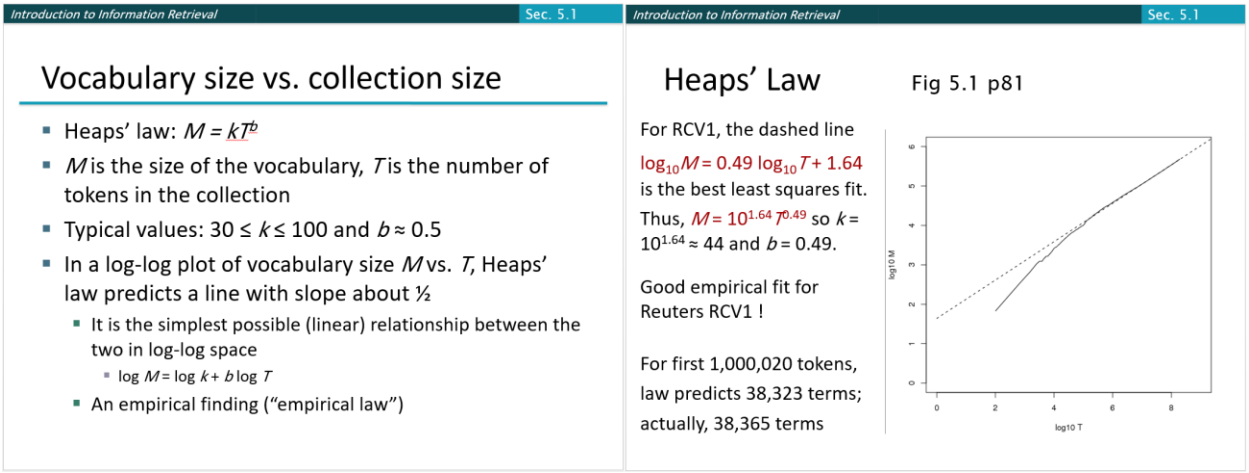


بعد:



مشاهده می‌شود که پس از حذف به zipf بیشتر منطبق می‌شود.

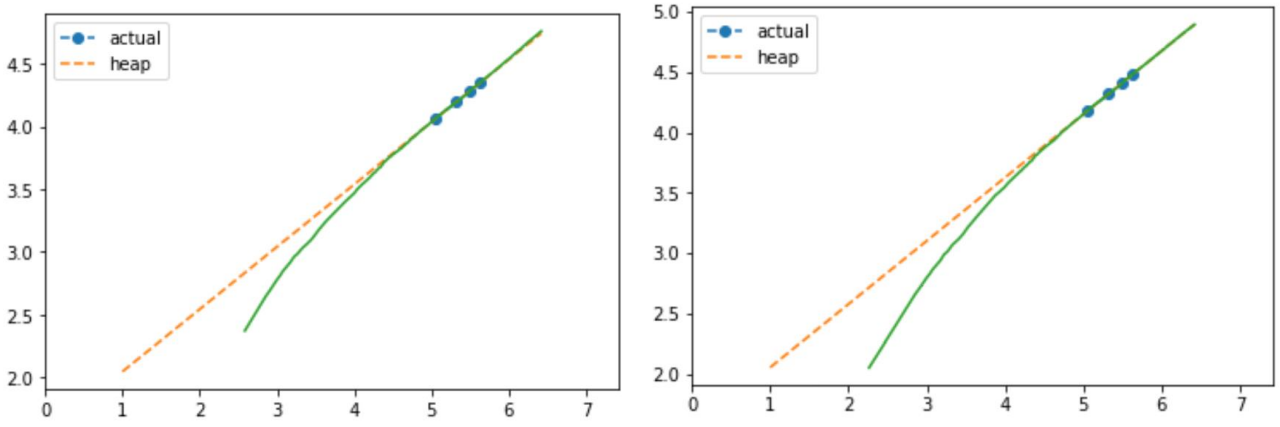
۳) صحت قانون heaps را در دو حالت قبل و بعد از ریشه یابی بررسی کنید. برای بررسی این قانون الزام است با استفاده از اندازه‌ی واژه‌نامه و تعداد توکن‌ها در ۵۰۰، ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ سند اول، اندازه‌ی واژه‌نامه مربوط به کل اسناد تخمین زده شود. در نهایت اندازه‌ی واقعی واژه‌نامه و اندازه‌ی تخمینی در هر دو حالت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



در ابتدا می‌بینیم که دیتافریم به دلیل نرمال نبودن توزیع داده‌ها در heaps صدق نمی‌کند اما بعد از ریشه‌یابی می‌بینیم که در قانون heaps صدق می‌کند.

قبل:

بعد:

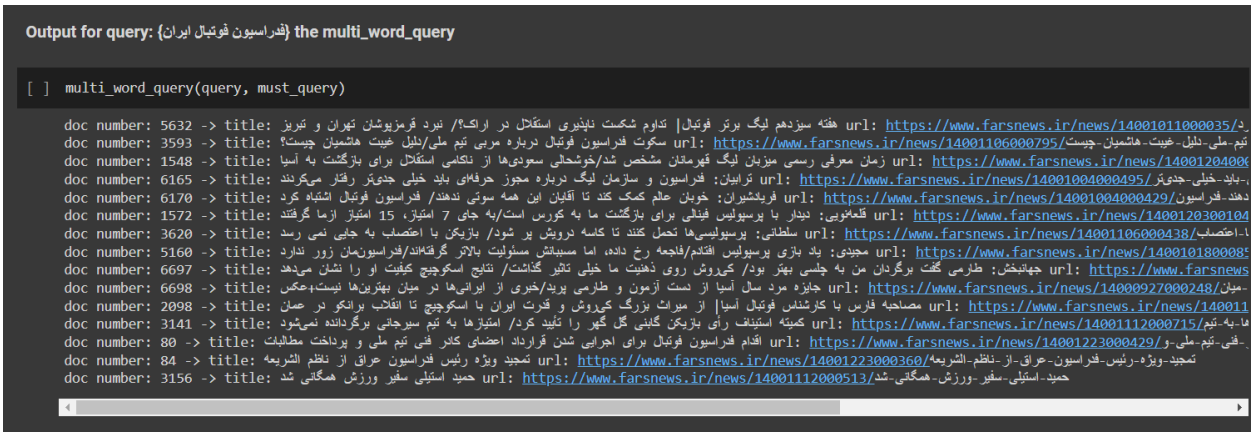


۴) حداقل سه مورد از مواردی که در ریشه‌یابی با چالش روبرو بودید را ذکر کنید. بطور مثال کلماتی که نیازی به ریشه‌یابی ندارند اما طبق روند ریشه‌یابی از دست می روند.

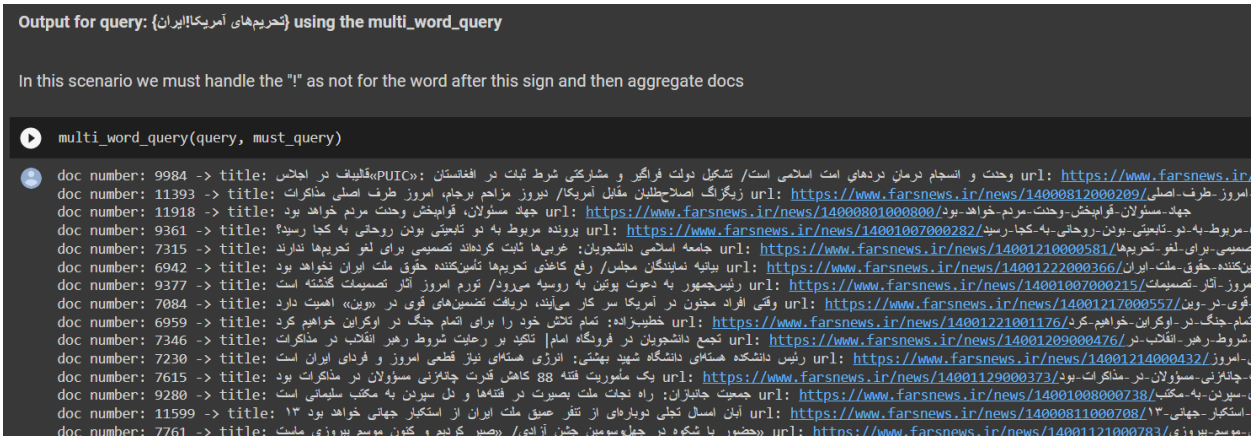
در روند ریشه‌هابی مشکلی نبود صرفا در برخی موارد که نیاز نبود stemming می‌شد که در روند کار اخلالی ایجاد نمی‌کرد. در یک سری موارد نیز ها و های خوب حذف نمی‌شدند که در قسمت stopwords خودمان به لیست اضافه کردیم تا مشکل برطرف شود.

۵) پاسخ به پرسمان در حالت های زیر:

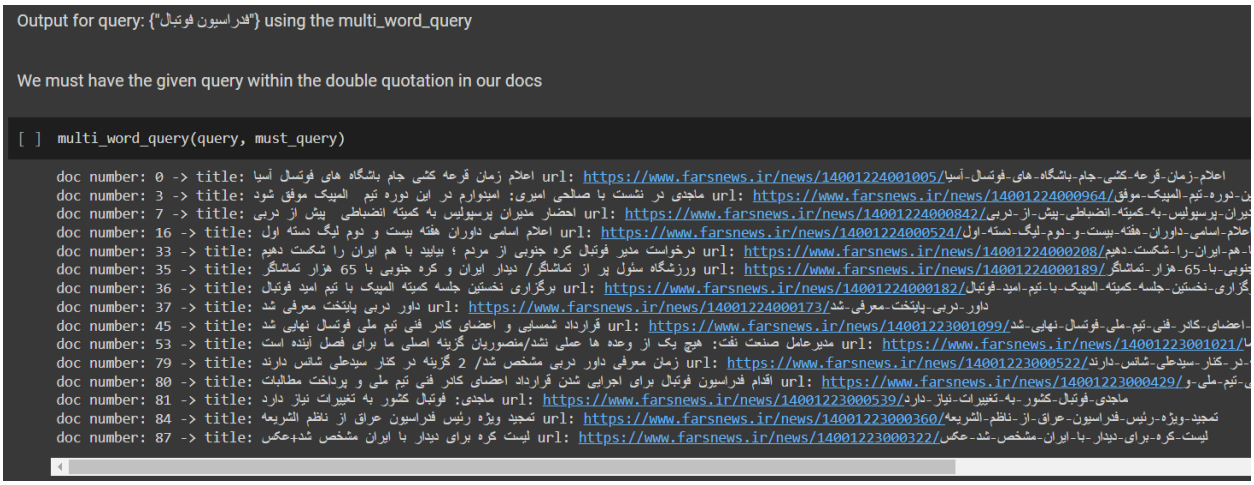
الف) یک پرسمان از کلمات ساده و متداول مانند تحریم‌های آمریکا علیه ایران، در نتایج بازیابی شده انتظار می‌رود اسنادی که کلمات تحریم، آمریکا، علیه و ایران را دارند در بالای لیست و اسنادی که برخی از کلمات را ندارند در رتبه‌های پایینتر لیست قرار داشته باشند.



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 بازیابی شده وجود داشته باشند.



پ) یک پرسمان با عملگر عبارت مانند "کنگره ضدتروریست"، انتظار میرود اسنادی که شامل عبارت کنگره ضدتروریست در نتایج بازیابی شده وجود داشته باشند؛ به عبارت دیگر موقعیت مکانی کلمات در این حالت مهم است.



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 بازیابی شده وجود داشته باشد.

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.

اگر تمامی سندها اگر باز شوند، مشاهده می‌شوند که مرتبطند. فقط برای مورد ث کلمه اورشلیم کلا در داک‌ها وجود ندارد، در نتیجه از گابن استفاده شد.

فاز ۲

در این پخش قصد داریم تا با به دست آوردن امتیاز و محاسبه معیار شباهت اسناد برای یک کوئری ورودی بهترین اسناد بر برای کاربر استخراج کنیم.

در ابتدا با محاسبه tf.idf برای اسناد و سپس محاسبه tf برای کوئری به معیار امتیازدهی مدنظر خود می‌رسیم، سپس با نرمالایز کردن امتیاز به دست آمده بر نرم۲ بردار اسناد و کوئری به کسینوس بین دو بردار می‌رسیم. به این معیار به دست آمده شباهت کسینوسی گفته می‌شود.

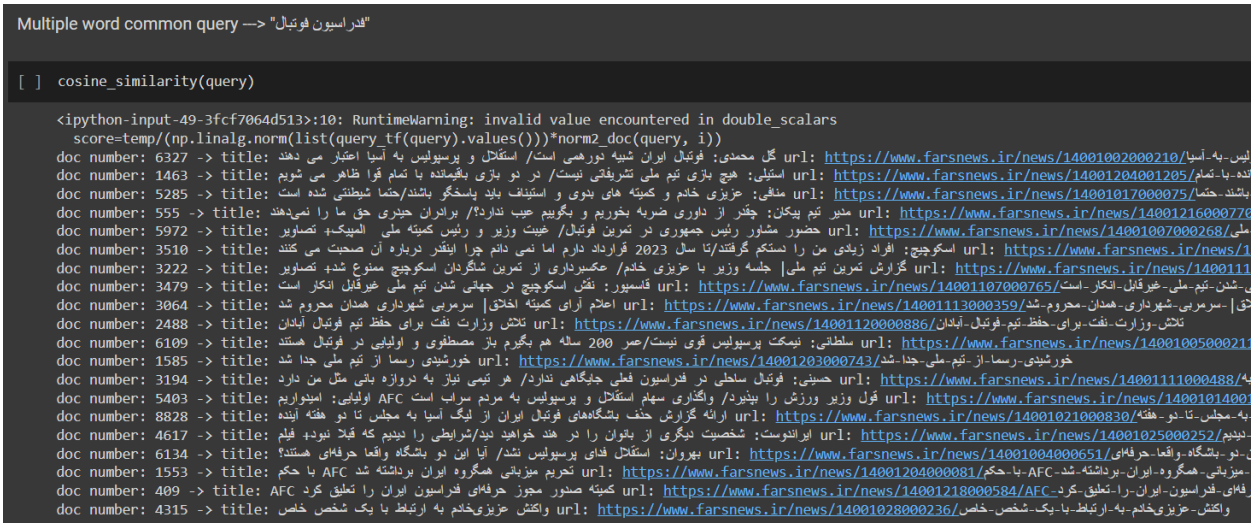
در بخش بعد با در نظر گرفتن معیارهای به‌دست آمده یک champion list برای بهترین اسناد هرکدام از کلمه‌های خود درست می‌کنیم، این کار کمک می‌کند تا بتوانیم در مواقع مواجهه با یک کوئری از بین بهترین اسناد مرتبط با آن کوئری اسناد را استخراج کنیم و همچنین به دلیل اینکه اسناد از پیش آماده هستند با سرعت پردازش سریع‌تری مواجه هستیم ولی از طرفی این یک روش lossy هست و لزوما top-k اصلی را به ما نمی‌دهد ولی سرعت پاسخ‌گویی را چند برابر می‌کند.

۱) پاسخ به پرسمان در حالت های زیر:

الف) یک پرسمان از کلمات ساده و متداول تک کلمه‌ای

با توجه به کلمه ساده انتخاب شده که «فدراسیون» است مشاهده می‌کنیم که در بخش اول اسنادی که وزن بیشتری از آن کلمه را دارند به ما نمایش داده می‌شود ولی در champion list در واقع نتایج بهتری داریم زیرا کلمه فدراسیون می‌تواند استفاده‌هایی خاص منظوره‌ای در اسناد ورزشی داشته باشد و همانطور که مشاهده می‌کنیم روش دوم نتایج مطلوب‌تری را برای ما کسب می‌کند.

ب) یک پرسمان از عبارات ساده و متداول چند کلمه‌ای



```
[ ]  champion_similarity(query)

    <ipython-input-55-97d8424d142a>:10: RuntimeWarning: invalid value encountered in double_scalars
      score=temp/(np.linalg.norm(list(query_tf(query).values()))*norm2_doc(query, i))
    doc number: 6128 -> title: داداش زاده:کسانی که در راس سرخابی ها بودند حالا برای دو باشگاه تصمیم گیر شده اند/آنها با احساسات مردم بازی می کنند url: https://www.farsnews.ir
    doc number: 4918 -> title: عزیزی خادم: فوتبال ایران از وضعیت اسفبار گذشته عبور کرد url: https://www.farsnews.ir/news/14001021000577/-وضعیت-اسفبار-گذشته-عبور-کرد
    doc number: 1950 -> title: فریادشیران:رئیس هزار شعبه با خردجمعی برکنار شد/تا خدا هست،پیوندهای شیطانی پودر می‌شوند url: https://www.farsnews.ir/news/14001128000554/
    doc number: 2268 -> title: عربشاهی: تصمیم هیات رئیسه برای برکناری عزیزی خادم جهادی بود/ فعالیت کمیته اخلاق غیرقانونی است url: https://www.farsnews.ir/news/1400112
    doc number: 6323 -> title: قدم برداشتیم ولی اجازه ندادند VAR عزیزی خادم:وقتی به فدراسیون آمدم یک چای خشک هم در آن نبود/برای url: https://www.farsnews.ir/news/14001
    doc number: 5102 -> title: باشگاه استقلال: رای کمیته بدوی عامل حذف ما از آسیا بود url: https://www.farsnews.ir/news/14001019000346/-بدوی-عامل-حذف-ما-از-آسیا-بود
    doc number: 139 -> title: تقدیر مربی تیم فوتبال خلیج فارس از ماجدی بابت رسیدگی به شائبه تبانی در لیگ جوانان url: https://www.farsnews.ir/news/14001222000297/-تبانی-در
    doc number: 5583 -> title: واکنش نفت آبادان به شکایت پیکان از ساکی با کنایه به مدیرعامل باشگاه تهرانی و فدراسیون فوتبال url: https://www.farsnews.ir/news/1400101200066
    doc number: 4603 -> title: هشدار عزیزی خادم به بی اخلاقی در فوتبال؛ ستاره هم باشید با بی اخلاقی حذف می شوید url: https://www.farsnews.ir/news/14001025000363/-اخلاقی
    doc number: 3057 -> title: باشگاه گل‌گهر :با افتضاح کمیته بین‌الملل تاریخی‌ترین اشتباه رقم خورد/ماجرای پاتوسی را فراموش کردید؟ url: https://www.farsnews.ir/news/1400111300
    doc number: 1824 -> title: ساکت: فراز و نشیب‌های سپاهان برای ما هم تلخ بود/ نبودیکا مورد حمایت باشگاه است url: https://www.farsnews.ir/news/14001130000599/-حمایت
    doc number: 3457 -> title: عزیزی‌خادم: ما را به جادوگری متهم کردند ولی جادوگری شرک است/هنوز برای کیروش مالیات می‌دهیم/مگر اسکوچیچ رفته است؟ url: https://www.farsnews
    doc number: 3141 -> title: کمیته استیناف رأی بازیکن گابنی گل گهر را تأیید کرد/ امتیازها به تیم سیرجانی برگردانده نمی‌شود url: https://www.farsnews.ir/news/1400111200071
    doc number: 6764 -> title: اعلام رای کمیته انضباطی در خصوص بازیکن گابنی/ گل گهر در بازی با پیکان، استقلال و سپاهان بازنده شد url: https://www.farsnews.ir/news/14000
    doc number: 5252 -> title: عزیزی‌خادم: حذف سرخابی ها درسی برای آینده فوتبال! / حضور هواداران مقابل عراق و امارات قطعی نیست url: https://www.farsnews.ir/news/14001
    doc number: 2626 -> title: ماجدی:فدراسیون فوتبال بدون حمایت مجلس و دولت کاری از پیش نمی برد/ به اراجیفی مثل جادوگری اعتقاد ندارم url: https://www.farsnews.ir/news/14
    doc number: 7305 -> title: شعارهای زیبا اما درون‌تهی/ وقتی غرب ورزش را سیاسی می‌کند url: https://www.farsnews.ir/news/14001210000621/-غرب-ورزش-را-سیاسی-می‌کند
    doc number: 5221 -> title: حباب اهمال و کم کاری فدراسیون فوتبال ترکید/ قربانی به نام استقلال و پرسپولیس url: https://www.farsnews.ir/news/14001018000261/-نام-استقلال-و
    doc number: 5005 -> title: ریخت و پاش میلیاردی از جیب مردم در تور تفریحی عزیزی‌خادم و یاران/ چرا فدراسیون سکوت می‌کند؟ url: https://www.farsnews.ir/news/140010200
    doc number: 1974 -> title: اخبار مهم جلسه 9 ساعته مدیران فدراسیون| اصولی: پیام اقدام هیات رئیسه خرد جمعی بود/ مجمع 24 اردیبهشت برگزار می شود url: https://www.farsnews.
```

در بخش بعدی کلمه انتخاب شده «فدراسیون فوتبال» است و همانطور که مشاهده می‌کنیم در اینجا هر دو روش نتایج قابل قبول مرتبط با کوئری وارد شده کسب می‌کنند ولی در زمان اجرا champion list با سرعت بالاتری اطلاعات را برای ما نمایش می‌دهد زیرا اطلاعات مرتبط به آن از پیش آماده هستند.

پ)یک پرسمان دشوار و کم تکرار تک کلمه‌ای

```
One rare word --> "گابن"

[ ]  cosine_similarity(query)

    <ipython-input-49-3fcf7064d513>:10: RuntimeWarning: invalid value encountered in double_scalars
      score=temp/(np.linalg.norm(list(query_tf(query).values()))*norm2_doc(query, i))
    doc number: 6799 -> title: 8 نکته از هفته دهم لیگ برتر| اعتراف سرمربی استقلال،بازی پرکنایه پرسپولیس و سوپرگل لیست مازاد یحیی/تکرار حرکت مجیدی توسط قلعه نویی url: https://www.fa
    doc number: 3717 -> title: 8 نکته از هفته شانزدهم لیگ برتر| از مشکلات شاعرانه تا کنایه ستاره میلیاردی پرسپولیس/ در انتظار شوک محرم، کمیته مسابقات مردود شد url: https://www.farsne
    doc number: 3224 -> title: زمان صدور رای پرونده بازیکن گابنی گل‌گهر مشخص شد url: https://www.farsnews.ir/news/14001111000661/-رای-پرونده-بازیکن-گابنی-گل‌گهر-مشخص-شد
    doc number: 3232 -> title: خرید جدید ژاوی به شهر بارسلونا رسید +عکس url: https://www.farsnews.ir/news/14001111000603/-خرید-جدید-ژاوی-به-شهر-بارسلونا-رسید-عکس
    doc number: 3236 -> title: بارسلونا در نقاشی فروش دمبله url: https://www.farsnews.ir/news/14001111000540/-فروش-دمبله--ستاره-فرانسوی-ماندنی-شد / ستاره فرانسوی ماندنی شد؟
    doc number: 3257 -> title: پرونده بازیکن گل گهر به روزهای آخر رسید/دادرس: من گزارش 52صفحه‌ای از پلیس فتا ندیدم url: https://www.farsnews.ir/news/14001111000302/-ای-از
    doc number: 3445 -> title: خرید جدید بارسلونا کیست؟ url: https://www.farsnews.ir/news/14001108000467/-خرید-جدید-بارسلونا-کیست
    doc number: 3457 -> title: عزیزی‌خادم: ما را به جادوگری متهم کردند ولی جادوگری شرک است/هنوز برای کیروش مالیات می‌دهیم/مگر اسکوچیچ رفته است؟ url: https://www.farsnews.ir/n
    doc number: 3807 -> title: عضو کمیته استیناف: احتمال شکستن رای کمیته انضباطی در مورد بازیکن گابنی 50-50 است url: https://www.farsnews.ir/news/14001104000123/-گابنی
    doc number: 3141 -> title: کمیته استیناف رأی بازیکن گابنی گل گهر را تأیید کرد/ امتیازها به تیم سیرجانی برگردانده نمی‌شود url: https://www.farsnews.ir/news/14001112000715/-تیم
    doc number: 3894 -> title: پلیس فتا: از سکوت فدراسیون فوتبال متعجبیم/ نمی‌توانیم پاسخ هواداران استقلال و گل گهر را بدهیم url: https://www.farsnews.ir/news/14001103000320/-ران
    doc number: 4040 -> title: هفته شانزدهم لیگ برتر| اولتیماتومی نرم به قلعه نویی و بازگشت فکری به لیگ url: https://www.farsnews.ir/news/14001030000070/-بازگشت-فکری-به
    doc number: 4138 -> title: از عرش به فرش ؛ پیشنهاد باشگاه سعودی به ستاره آرسنال url: https://www.farsnews.ir/news/14001030000273/-پیشنهاد-باشگاه-سعودی-به-ستاره-آرسنال
    doc number: 4419 -> title: شوک بزرگ به ستاره اخراجی آرسنال / اوباميانگ جام ملت ها را از دست داد؟ url: https://www.farsnews.ir/news/14001027000456/-ها-را-از-دست-داد
    doc number: 4459 -> title: مرحله یک هشتم نهایی جام حذفی|تکرار پیروزی استقلال مقابل پیکان؟ url: https://www.farsnews.ir/news/14001026000849/-تکرار-پیروزی-استقلال-مقابل-پیکان
    doc number: 4564 -> title: پاسخ باشگاه پیکان به اظهارات وکیل گل‌گهر: با راه انداختن بحث رسانه ای جو سازی می‌کنند url: https://www.farsnews.ir/news/14001025000784/-رسانه-ای
    doc number: 3203 -> title: انتقال اوباميانگ به بارسلونا منتفی شد/بازگشت ستاره گابنی به لندن url: https://www.farsnews.ir/news/14001111000887/-منتفی-شد-بازگشت-ستاره-گابنی-به-لندن
    doc number: 3103 -> title: اوباميانگ رسما قراردادش با آرسنال را فسخ کرد/در انتظار حضور ستاره گابنی در بارسا url: https://www.farsnews.ir/news/14001112001065/-حضور-ستاره
    doc number: 6764 -> title: اعلام رای کمیته انضباطی در خصوص بازیکن گابنی/ گل گهر در بازی با پیکان، استقلال و سپاهان بازنده شد url: https://www.farsnews.ir/news/1400092500
    doc number: 1709 -> title: هت تریک اوباميانگ برای بارسلونا با تصمیم داور!+فیلم url: https://www.farsnews.ir/news/14001201001230/-فیلم-داور-تصمیم-با-بارسلونا-برای-اوباميانگ-یک
```

کلمه سخت انتخاب شده ما «گابن» است. در سال گذشته بازی کردن یک بازیکن گابنی غیرمجاز در لیگ فوتبال ایران تا مدت‌ها سوژه خبرگزاری‌های داخلی بود به همین دلیل انتظار می‌رود که با وارد کردن این عبارت به اسناد مرتبط از دسته ورزشی برخورد کنیم که مشابه آنچه در تصویر می‌بینید تمامی اسناد استخراج شده مرتبط با همین موضوع هستند.

ت) یک پرسمان دشوار و کم تکرار چند کلمه‌ای

در این بخش با ترکیب کلمات «فدراسیون فوتبال گابن» می‌خواهیم اسناد بازگشتی را مشاهده کنیم. ما در بین اسناد، اسنادی را موجود داریم که عینا ترکیب این ۳ کلمه را دارند ولی در اینجا به دلیل اینکه تنها تکرار کلمات مختلف از اهمیت بالایی برخوردار است نمی‌توانیم در واقع مرتبط‌ترین سند را که شامل این کلمه است را مشاهده کنیم در صورتی که با استفاده از positional index می‌توان سند مرتبط را دریافت کرد.

۲) موارد ب و ت را با روش مکانی فاز یک نیز تکرار کنید و نتایج دو حالت را با هم مقایسه و تحلیل کنید.

در فاز ۱، داک‌هایی که هر دو کلمه را دارند در الویتند ولی در فاز ۲ این مورد ۱۰۰٪ اینگونه نیست ولی از لحاظ زمانی عملکرد فاز ۲ بهتر است.