



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

درس بازیابی اطلاعات

گزارشکار پروژه

موظان میرجلیلی

فرحان کیهان

استاد

دکتر نیک آبادی

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.

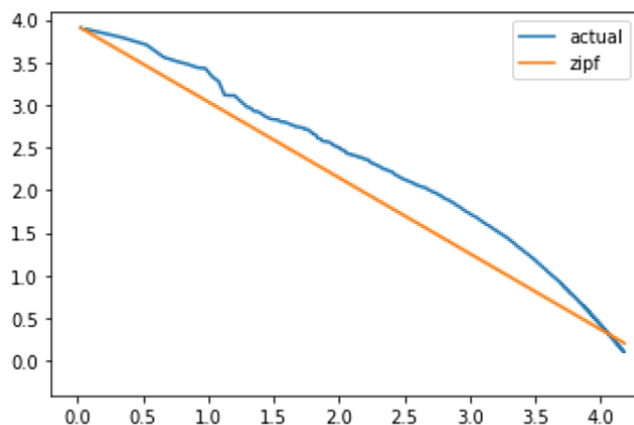
در پیش‌پردازش ابتدا دیتافریم را خواندیم و transpose کردیم تا به فرمت مناسب (هر ستون یک فیچر و هر سطر یک داده) در بیاید. سپس normalize می‌کنیم، در این بخش نیم فاصله‌ها اصلاح می‌شوند و همه کلمات به یک حالت درمی‌آیند. برای مثال "تحریم‌ها"، "تحریم‌ها" می‌شود. در مرحله بعد یک سری موارد از متن حذف می‌شوند ('<@?!\.,:;{}[]()~_*^/%#&'\"' \u200c', '\n').

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، خبر گزار می‌شود. همه موارد ذکر شده جز punctuation list با استفاده از hazm اجرا شدند.

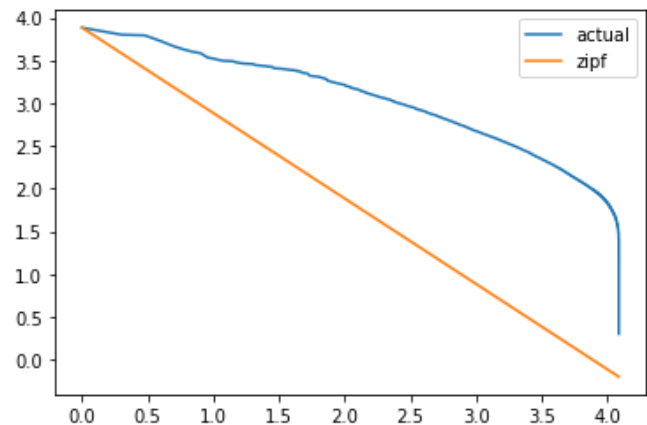
۲) صحت قانون Zipf را در دو حالت قبل و بعد از حذف کلمات پرتکرار از واژه‌نامه بررسی کنید. رسم نمودار برای هر حالت الزامی است. در صورت برقراری / عدم برقراری این قانون در هر حالت، علت را شرح دهید.

Introduction to Information Retrieval Sec. 5.1	Introduction to Information Retrieval Sec. 5.1
<h3>Zipf's law</h3> <ul style="list-style-type: none"> Heaps' law gives the vocabulary size in collections. We also study the relative frequencies of terms. In natural language, there are a few very frequent terms and very many very rare terms. Zipf's law: The i^{th} most frequent term has frequency proportional to $1/i$. $cf_i \propto 1/i = K/i$ where K is a normalizing constant cf_i is collection frequency: the number of occurrences of the term t_i in the collection. 	<h3>Zipf consequences</h3> <ul style="list-style-type: none"> If the most frequent term (<i>the</i>) occurs cf_1 times <ul style="list-style-type: none"> then the second most frequent term (<i>of</i>) occurs $cf_1/2$ times the third most frequent term (<i>and</i>) occurs $cf_1/3$ times ... Equivalent: $cf_i = K/i$ where K is a normalizing factor, so <ul style="list-style-type: none"> $\log cf_i = \log K - \log i$ Linear relationship between $\log cf_i$ and $\log i$ Another power law relationship

بعد:

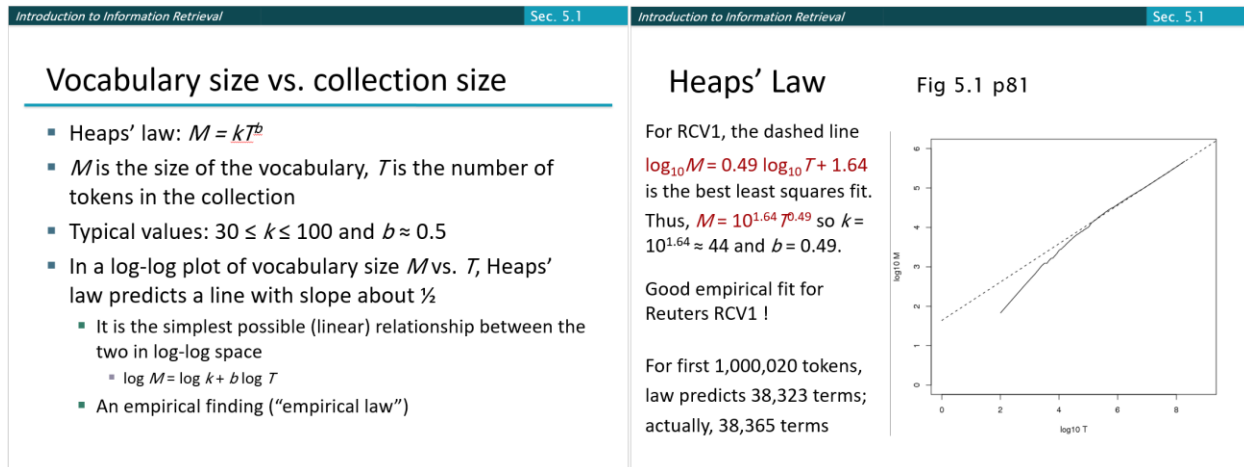


قبل:



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

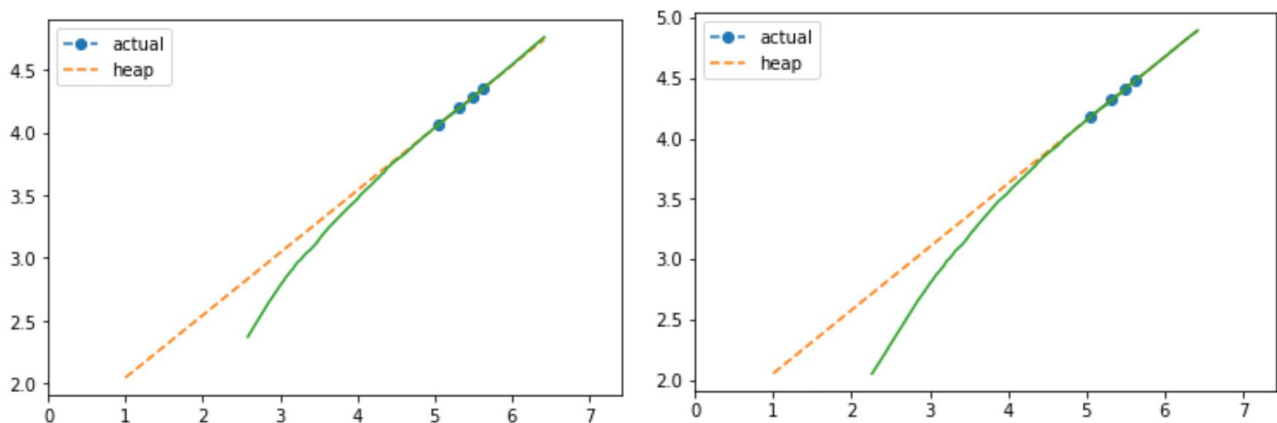
۳) صحت قانون heaps را در دو حالت قبل و بعد از ریشه یابی بررسی کنید. برای بررسی این قانون الزام است با استفاده از اندازه‌ی واژه‌نامه و تعداد توکن‌ها در ۵۰۰، ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ سند اول، اندازه‌ی واژه‌نامه مربوط به کل اسناد تخمین زده شود. در نهایت اندازه‌ی واقعی واژه‌نامه و اندازه‌ی تخمینی در هر دو حالت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

در روند ریشه‌هایی مشکلی نبود صرفاً در برخی موارد که نیاز نبود stemming می‌شد که در روند کار اخلاقی 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.

Output for query: the multi_word_query {ایران فراسیون فوتبال}

[] multi_word_query(query, must_query)

doc number: 5632 -> title: url: https://www.farsnews.ir/news/14001011000035/د هفته سیزدهم لیگ برتر فوتبال/ تداوم شکست ناینبری استقلال در اراک! نبرد قرمزپوشان تهران و تبریز

doc number: 3593 -> title: url: https://www.farsnews.ir/news/14001106000795/ تیم ملی-لیل غیبت هشتمین چیست؟ سکوت فراسیون فوتبال درباره مربی تیم ملی/لیل غیبت هشتمین چیست؟

doc number: 1548 -> title: url: https://www.farsnews.ir/news/140012040000495/ زمان معرفی رسمی میزبان لیگ قهرمانان مشخص شد/خوشحالی سعودی‌ها از ناکامی استقلال برای بازگشت به آسیا

doc number: 6165 -> title: url: https://www.farsnews.ir/news/140010040000429/ فراسیون فوتبال ششام کرد

doc number: 6170 -> title: url: https://www.farsnews.ir/news/14001106000438/ دهنده فراسیون فوتبال ششام کرد

doc number: 1572 -> title: url: https://www.farsnews.ir/news/1400120300104/ دهنده فراسیون فوتبال ششام کرد

doc number: 3620 -> title: url: https://www.farsnews.ir/news/14001106000438/ دهنده فراسیون فوتبال ششام کرد

doc number: 5160 -> title: url: https://www.farsnews.ir/news/14001120000715/ دهنده فراسیون فوتبال ششام کرد

doc number: 6697 -> title: url: https://www.farsnews.ir/news/14001120000429/ دهنده فراسیون فوتبال ششام کرد

doc number: 6698 -> title: url: https://www.farsnews.ir/news/14001120000429/ دهنده فراسیون فوتبال ششام کرد

doc number: 2098 -> title: url: https://www.farsnews.ir/news/14001120000429/ دهنده فراسیون فوتبال ششام کرد

doc number: 3141 -> title: url: https://www.farsnews.ir/news/14001120000429/ دهنده فراسیون فوتبال ششام کرد

doc number: 80 -> title: url: https://www.farsnews.ir/news/14001120000429/ دهنده فراسیون فوتبال ششام کرد

doc number: 84 -> title: url: https://www.farsnews.ir/news/14001120000429/ دهنده فراسیون فوتبال ششام کرد

doc number: 3156 -> title: url: https://www.farsnews.ir/news/14001120000429/ دهنده فراسیون فوتبال ششام کرد

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.

Output for query: using the multi_word_query {تحریم‌های آمریکا!ایران}

In this scenario we must handle the "!" as not for the word after this sign and then aggregate docs

multi_word_query(query, must_query)

doc number: 9984 -> title: url: https://www.farsnews.ir/news/14000812000209/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 11393 -> title: url: https://www.farsnews.ir/news/14000812000209/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 11918 -> title: url: https://www.farsnews.ir/news/14000812000209/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 9361 -> title: url: https://www.farsnews.ir/news/14001007000282/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 7315 -> title: url: https://www.farsnews.ir/news/14001210000581/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 6942 -> title: url: https://www.farsnews.ir/news/14001222000366/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 9377 -> title: url: https://www.farsnews.ir/news/14001007000215/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 7084 -> title: url: https://www.farsnews.ir/news/14001217000557/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 6959 -> title: url: https://www.farsnews.ir/news/14001221001176/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 7346 -> title: url: https://www.farsnews.ir/news/14001209000476/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 7230 -> title: url: https://www.farsnews.ir/news/14001214000432/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 7615 -> title: url: https://www.farsnews.ir/news/14001210000581/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 9280 -> title: url: https://www.farsnews.ir/news/14001008000738/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 11599 -> title: url: https://www.farsnews.ir/news/14000811000788/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

doc number: 7761 -> title: url: https://www.farsnews.ir/news/1400121000783/ امروز-طرف اصلی-فرابخش وحت مردم-خواد بود/ زیرگاک اصاحجلان مقال آمریکا/ دیروز مزاحم بوجام، امروز طرف اصلی مذاکرات

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

Output for query: {"فدراسیون فوتبال"} using the multi_word_query

We must have the given query within the double quotation in our docs

[] multi_word_query(query, must_query)

```
doc number: 0 -> title: اعلام زمان قرعه کشی جام باشگاه های فوتبال آسیا url: https://www.farsnews.ir/news/14001224001005
doc number: 3 -> title: تیم المپیک موفق url: https://www.farsnews.ir/news/14001224000964
doc number: 7 -> title: کمیته انضباطی پیش از دربی url: https://www.farsnews.ir/news/14001224000842
doc number: 16 -> title: اعلام اسامی داوران هفته بیست و دوم لیگ دسته اول url: https://www.farsnews.ir/news/14001224000524
doc number: 33 -> title: درخواست منیر فوتبال کره جنوبی از مردم؛ بیلید با هم ایران را شکست دهیم url: https://www.farsnews.ir/news/14001224000208
doc number: 35 -> title: هزار تماشاگر url: https://www.farsnews.ir/news/14001224000189
doc number: 36 -> title: نخستین جلسه کمیته المپیک با تیم امید فوتبال url: https://www.farsnews.ir/news/14001224000182
doc number: 37 -> title: داور دربی پایتخت معرفی شد url: https://www.farsnews.ir/news/14001224000173
doc number: 45 -> title: اعضای کادر فنی تیم ملی فوتبال نهایه شد url: https://www.farsnews.ir/news/14001223001099
doc number: 53 -> title: مدرعامل صنعت نفت: هیچ یک از وعده ها صلی نشد/منصوریان گزینه اصلی ما برای فصل آینده است url: https://www.farsnews.ir/news/14001223001021
doc number: 79 -> title: 2 گزینه در کنار سیدعلی شانس دارند url: https://www.farsnews.ir/news/14001223000522
doc number: 80 -> title: اقدام فدراسیون فوتبال برای اجرایی شدن قرارداد اعضای کادر فنی تیم ملی و پرداخت مطالبات url: https://www.farsnews.ir/news/14001223000429
doc number: 81 -> title: ماجرای فوتبال کشور به تغییرات نیاز دارد url: https://www.farsnews.ir/news/14001223000539
doc number: 84 -> title: تجدید ویژه رئیس فدراسیون عراق از ناظم الشریعه url: https://www.farsnews.ir/news/14001223000360
doc number: 87 -> title: لیست کره برای دیدار با ایران مشخص شد url: https://www.farsnews.ir/news/14001223000322
```

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

Output for query: {"تحریم هسته ای"آمریکا!ایران"} using the multi_word_query

[] multi_word_query(query, must_query)

```
doc number: 11393 -> title: امروز طرف اصلی url: https://www.farsnews.ir/news/14000812000209
doc number: 7395 -> title: دفتر بسیج دانشجویی 8 دانشگاه تهران: رویکرد "سازش" تجربه تلخ برجام را نصب تاریخ پرافتخار انقلاب کرد url: https://www.farsnews.ir/news/1400120700
doc number: 10862 -> title: دانشگاهی به بهانه آبان، برای پرونده‌سازی حقوق بشری/دانشگاهانی که هیچگاه تشکیل نمی‌شود url: https://www.farsnews.ir/news/14000826000161
doc number: 7346 -> title: تجمع دانشجویان در فروگاه امام| تاکید بر رعایت شروط رهبر انقلاب در مذاکرات url: https://www.farsnews.ir/news/14001209000476
doc number: 7315 -> title: برای لغو تحریم‌ها url: https://www.farsnews.ir/news/14001210000581
doc number: 10162 -> title: روایت خبرنگار فارس از یک همایش دانشجویی/ 2 ساعت و نیم صریح و بدون تعارف با رئیس مجلس url: https://www.farsnews.ir/news/14000915000494
doc number: 9980 -> title: ضلع تفکیک لاینر از مذاکرات url: https://www.farsnews.ir/news/14000920000012
doc number: 7615 -> title: کاهش قدرت چانه‌زنی مسؤولان در مذاکرات بود url: https://www.farsnews.ir/news/14001129000373
```

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

```
Output for a rare word query: {گابن اسیوی نیست}

[ ] multi_word_query(query, must_query)

doc number: 6764 -> title: url: https://www.farsnews.ir/news/140009250
doc number: 4419 -> title: url: https://www.farsnews.ir/news/14001027000456/
doc number: 3141 -> title: url: https://www.farsnews.ir/news/14001112000715/
doc number: 3079 -> title: url: https://www.farsnews.ir/news/14001113000180/
doc number: 6662 -> title: url: https://www.farsnews.ir/news/
doc number: 6375 -> title: url: https://www.farsnews.ir/news/14001001000373/
doc number: 820 -> title: url: https://www.farsnews.ir/news/1400
doc number: 5612 -> title: url: https://www.farsnews.ir/news/14001012000223/
doc number: 6799 -> title: url: https://www.f
doc number: 5657 -> title: url: https://www.farsnews.ir/news/1400101100
doc number: 5789 -> title: url: https://www.farsnews.ir/news/14001010000506/
doc number: 3103 -> title: url: https://www.farsnews.ir/news/14001112001065/
doc number: 3224 -> title: url: https://www.farsnews.ir/news/14001111000661/
doc number: 3232 -> title: url: https://www.farsnews.ir/news/14001111000603/
doc number: 4900 -> title: url: https://www.farsnews.ir/news/14001021000865/
doc number: 5632 -> title: url: https://www.farsnews.ir/news/14001011000003/
```

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

فاز ۲

در این بخش 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 نرم ۲ 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 به دست آمده یک champion list برای بهترین اسناد هر کدام 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 لزوماً 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

```
One word common query -> "فدراسیون"

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: 12118 -> title: روسیه: اقتصادی ایران و روسیه url: https://www.farsnews.ir/news/140007260005
doc number: 2482 -> title: ثبت درخواست توسط هیئت رئیسه url: https://www.farsnews.ir/news/14001120000902
doc number: 2437 -> title: معرفی نفرت برتر مرحله پنجم مسابقات آزاد تیراندازی url: https://www.farsnews.ir/news/14001121000653
doc number: 2449 -> title: اسامی بازیکنان دعوت شده به اردوی تیم ملی بسکتبال اعلام شد url: https://www.farsnews.ir/news/14001121000484
doc number: 2453 -> title: تصمیمات سازمان لیگ فوتبال راهگشا یا بلائی جان؟ url: https://www.farsnews.ir/news/14001119001101
doc number: 2458 -> title: درخواست برای برگزاری عزیزوخام مطابق اسانده است/ فیفا واکنشی نشان نمی‌دهد url: https://www.farsnews.ir/news/14001121000288
doc number: 2460 -> title: واکنش آلبا به ادعاهای دروغین در خصوص بوکس ایران url: https://www.farsnews.ir/news/14001120000886
doc number: 2464 -> title: بحث استیضاح نیست بلکه عزل موقت است url: https://www.farsnews.ir/news/14001120000968
doc number: 2466 -> title: پرسپولیس قسط اول بپردازد و پدیده را پرداخت کرد url: https://www.farsnews.ir/news/14001121000206
doc number: 2475 -> title: سوشمشکی B المپیک زمستانی یکن درخواست برای آزمایش نمونه url: https://www.farsnews.ir/news/14001121000077
doc number: 2478 -> title: بیانیه مس سونگون در خصوص تصمیم سازمان لیگ فوتبال url: https://www.farsnews.ir/news/14001120001064
doc number: 2483 -> title: مشاور مدیرعامل و سرپرست تیم پیکان استعفاء کرد url: https://www.farsnews.ir/news/14001120000968
doc number: 2429 -> title: فعالیت کمیته اخلاق فدراسیون فوتبال قانونی است؟! بلاخر شدن احتمالی احکام و بیانیه هابوند url: https://www.farsnews.ir/news/140011210000906
doc number: 2486 -> title: شمسابی: از سرمربیگری تیم ملی فوتبال استعفا ندادم/ منتظر امضای قرارداد هستم url: https://www.farsnews.ir/news/14001120000886
doc number: 2488 -> title: تلاش وزارت نفت برای حفظ تیم فوتبال آبادان url: https://www.farsnews.ir/news/14001120000841
doc number: 2491 -> title: وکیل بیرانوند با پرسپولیس به توافق رسید url: https://www.farsnews.ir/news/14001120000780
doc number: 2495 -> title: مصوبت نشست هیئت رئیسه فدراسیون والیبال چه بود؟ url: https://www.farsnews.ir/news/14001120000720
doc number: 2500 -> title: هشدار فدراسیون فوتبال در خصوص نماینده قلابی باشگاه بوونتوس url: https://www.farsnews.ir/news/14001120000710
doc number: 2501 -> title: منع کلیه فعالیت‌های دو بازیکن در فوتبال بانوان و مردان توسط کمیته اخلاق url: https://www.farsnews.ir/news/14001120000684
doc number: 2503 -> title: رئیس دپارتمان حقوقی فدراسیون فوتبال: ارکان قضایی در صدور رای مشکل هستند
```



```
[ ] 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: url: https://www.farsnews.ir
-وضعیت-اسفیر-گشته-خور-کرد-14001021000577: https://www.farsnews.ir/news/14001021000577
doc number: 4918 -> title: url: https://www.farsnews.ir/news/14001128000554/
-فرانشیون:رئیس هزار شیهه با خردجمعی برکنار شد/تا خدا هست:پیوندهای شیطانی پودر می‌شوند
doc number: 1950 -> title: url: https://www.farsnews.ir/news/14001128000554/
-عرب‌ها: تصمیم هیأت رئیسه برای برکناری عزیزی خدیم جهلی بود/ فعالیت کمیته اخلاق غیرقانونی است
doc number: 2268 -> title: url: https://www.farsnews.ir/news/1400112
-فدراسیون فوتبال: VAR عزیزی خدیم: وقتی به فدراسیون آمدیم یک جای خشک هم در آن نبود/ برای
doc number: 6323 -> title: url: https://www.farsnews.ir/news/14001019000346/
-باشگاه استقلال: رای کمیته بدوی عمل حنف ما از آسیا بود
doc number: 5102 -> title: url: https://www.farsnews.ir/news/14001222000297/
-تئانی در: 14001222000297
doc number: 139 -> title: url: https://www.farsnews.ir/news/1400101200066
-واکنش نفت آبادان به شکایت پیکان از سکتی با کنایه به منیرعامل باشگاه تهرانی و فدراسیون فوتبال
doc number: 5583 -> title: url: https://www.farsnews.ir/news/14001025000363/
-خداقی: هشدار عزیزی خدیم به بی اخلاقی در فوتبال؛ ستاره هم باشید با بی اخلاقی حنف می شوید
doc number: 4603 -> title: url: https://www.farsnews.ir/news/14001113000
-باشگاه گل‌گهر: با افشاح کمیته بین‌الملل تاریخی‌ترین اشتباه رقم خورد/ماجرا یاتوسی را فراموش کردید؟
doc number: 3057 -> title: url: https://www.farsnews.ir/news/14001130000599/
-سکت: فزاد و نشیب‌های سیاهان برای ما هم تلخ بود/ نوبتکیا مورد حمایت باشگاه است
doc number: 1824 -> title: url: https://www.farsnews.ir/news/14001130000599/
-عزیزی‌خدیم: ما را به جادوگری متهم کردند ولی جادوگری شرک است/ هنوز برای کی‌روش مایلت می‌دهیم/ مگر اسکوچیچ رفته است؟
doc number: 3457 -> title: url: https://www.farsnews.ir/news/14001112000714
-کمیته استیناف رای بازیکن گیلانی گل گهر را تأیید کرد/ امتیازها به تیم سورجانی برگردانده نمی‌شود
doc number: 3141 -> title: url: https://www.farsnews.ir/news/14001112000714
-اعلام رای کمیته انضباطی در خصوص بازیکن گیلانی/ گل گهر در بازی با پیکان، استقلال و سیاهان بازنده شد
doc number: 6764 -> title: url: https://www.farsnews.ir/news/14001112000714
-حنف سرخابی ها درسی برای آینده فوتبال! حضور هواداران مقابل عراق و امارات قطعی نیست
doc number: 5252 -> title: url: https://www.farsnews.ir/news/14001112000714
-ماجی: فدراسیون فوتبال بنون حمایت مجلس و نولت کاری از پیش نمی برد/ به اراجیفی مثل جادوگری اعتقاد ندارم
doc number: 2626 -> title: url: https://www.farsnews.ir/news/14001112000714
-ورزش را-سلیسی می‌کند: 14001210000621
doc number: 7305 -> title: url: https://www.farsnews.ir/news/14001210000621
-حبیب اهل و کم کاری فدراسیون فوتبال ترکیه/ قربانی به نام استقلال و پرسپولیس
doc number: 5221 -> title: url: https://www.farsnews.ir/news/14001018000261/
-ریخت و پاش میلیارڈی از جیب مردم در تور تفریحی عزیزی‌خدیم و باران/ چرا فدراسیون سکوت می‌کند؟
doc number: 5005 -> title: url: https://www.farsnews.ir/news/140010200
-پایم اقدام هیأت رئیسه خرد جمعی بود/ مجمع 24 اردیبهشت برگزار می شود
doc number: 1974 -> title: url: https://www.farsnews.ir/news/140010200
```

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

گاین" → "One rare word

```
[ ] 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: url: https://www.farsnews.ir/news/14001110000661/
-رای پرونده بازیکن گیلانی-گل‌گهر- مشخص شد/ 14001111000661
doc number: 3717 -> title: url: https://www.farsnews.ir/news/14001110000661/
-رای پرونده بازیکن گیلانی-گل‌گهر- مشخص شد/ 14001111000661
doc number: 3224 -> title: url: https://www.farsnews.ir/news/14001110000661/
-خرید جدید- ژاوی- به شهر- بارسلونا- رسید- عکس/ 14001111000663
doc number: 3232 -> title: url: https://www.farsnews.ir/news/14001110000663/
-خرید جدید- ژاوی- به شهر- بارسلونا- رسید- عکس/ 14001111000663
doc number: 3236 -> title: url: https://www.farsnews.ir/news/14001110000540/
-فروش- نمبله- ستاره- فرانسوی- ماندنی- شد/ 14001111000540
doc number: 3257 -> title: url: https://www.farsnews.ir/news/14001111000302/
-ای- از- 14001111000302
doc number: 3445 -> title: url: https://www.farsnews.ir/news/14001108000267/
-خرید- جدید- بارسلونا- کیست/ 14001108000267
doc number: 3457 -> title: url: https://www.farsnews.ir/news/14001108000267/
-خرید- جدید- بارسلونا- کیست/ 14001108000267
doc number: 3807 -> title: url: https://www.farsnews.ir/news/14001104000123/
-ن- گیلانی/ 14001104000123
doc number: 3141 -> title: url: https://www.farsnews.ir/news/14001112000715/
-به- تیم/ 14001112000715
doc number: 3894 -> title: url: https://www.farsnews.ir/news/14001103000320/
-ران/ 14001103000320
doc number: 4040 -> title: url: https://www.farsnews.ir/news/14001030000070/
-ی- و- بازیگشت- فکری- به- لندن/ 14001030000070
doc number: 4138 -> title: url: https://www.farsnews.ir/news/14001030000273/
-فروش- پیشنهاد- باشگاه- سعودی- به- ستاره- ارسال/ 14001030000273
doc number: 4419 -> title: url: https://www.farsnews.ir/news/14001027000456/
-ت- ها- را- از- دست- داد/ 14001027000456
doc number: 4459 -> title: url: https://www.farsnews.ir/news/14001026000849/
-تکرار- پیروزی- استقلال- مقابل- پیکان/ 14001026000849
doc number: 4564 -> title: url: https://www.farsnews.ir/news/14001025000784/
-رساله- ای/ 14001025000784
doc number: 3203 -> title: url: https://www.farsnews.ir/news/14001110000887/
-منتفی- شد- بازیگشت- ستاره- گیلانی- به- لندن/ 14001110000887
doc number: 3103 -> title: url: https://www.farsnews.ir/news/14001112001065/
-حضور- ستاره/ 14001112001065
doc number: 6764 -> title: url: https://www.farsnews.ir/news/1400092500
-یک- اوبامیدگ- برای- بارسلونا- با- تصمیم- داور- فیلم/ 14001201001230
doc number: 1709 -> title: url: https://www.farsnews.ir/news/14001201001230
```

```
[ ] 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: 6799 -> title: url: https://www.farsnews.ir/news/1400111000661
doc number: 3717 -> title: url: https://www.farsnews.ir/news/1400111000603
doc number: 3224 -> title: url: https://www.farsnews.ir/news/1400111000661
doc number: 3232 -> title: url: https://www.farsnews.ir/news/1400111000603
doc number: 3236 -> title: url: https://www.farsnews.ir/news/1400111000540
doc number: 3257 -> title: url: https://www.farsnews.ir/news/1400111000302
doc number: 3445 -> title: url: https://www.farsnews.ir/news/14001108000467
doc number: 3457 -> title: url: https://www.farsnews.ir/news/14001104000123
doc number: 3807 -> title: url: https://www.farsnews.ir/news/1400112000715
doc number: 3141 -> title: url: https://www.farsnews.ir/news/1400112000715
doc number: 3894 -> title: url: https://www.farsnews.ir/news/14001020000070
doc number: 4044 -> title: url: https://www.farsnews.ir/news/14001030000273
doc number: 4138 -> title: url: https://www.farsnews.ir/news/14001027000456
doc number: 4419 -> title: url: https://www.farsnews.ir/news/14001026000849
doc number: 4459 -> title: url: https://www.farsnews.ir/news/14001025000784
doc number: 4564 -> title: url: https://www.farsnews.ir/news/1400111000887
doc number: 3203 -> title: url: https://www.farsnews.ir/news/1400112001065
doc number: 3103 -> title: url: https://www.farsnews.ir/news/14000925000
doc number: 6764 -> title: url: https://www.farsnews.ir/news/14001201001230
doc number: 1709 -> title: url: https://www.farsnews.ir/news/14001201001230
```

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

```
Multiple rare words -> "فراسیون فوتبال 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: 6327 -> title: url: https://www.farsnews.ir/news/14001002000210
doc number: 1797 -> title: url: https://www.farsnews.ir/news/14001002000436
doc number: 3048 -> title: url: https://www.farsnews.ir/news/14001024000169
doc number: 3057 -> title: url: https://www.farsnews.ir/news/14001015000545
doc number: 4661 -> title: url: https://www.farsnews.ir/news/14001103000320
doc number: 5377 -> title: url: https://www.farsnews.ir/news/14000927000272
doc number: 5373 -> title: url: https://www.farsnews.ir/news/14001010000506
doc number: 3894 -> title: url: https://www.farsnews.ir/news/1400100500002
doc number: 3457 -> title: url: https://www.farsnews.ir/news/140010100005
doc number: 6700 -> title: url: https://www.farsnews.ir/news/1400112000715
doc number: 6566 -> title: url: https://www.farsnews.ir/news/1400120300102
doc number: 3101 -> title: url: https://www.farsnews.ir/news/14001010000506
doc number: 1572 -> title: url: https://www.farsnews.ir/news/1400100500002
doc number: 5709 -> title: url: https://www.farsnews.ir/news/14001010000506
doc number: 6108 -> title: url: https://www.farsnews.ir/news/140010100005
doc number: 5657 -> title: url: https://www.farsnews.ir/news/1400112000715
doc number: 3141 -> title: url: https://www.farsnews.ir/news/140010100005
doc number: 5294 -> title: url: https://www.farsnews.ir/news/14001025000
doc number: 4587 -> title: url: https://www.farsnews.ir/news/14000926000180
doc number: 6740 -> title: url: https://www.farsnews.ir/news/14000926000180
```

```
 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: 3057 -> title: url: https://www.farsnews.ir/news/1400111300043
doc number: 3457 -> title: url: https://www.farsnews.ir/news/14001112000715/
doc number: 3141 -> title: url: https://www.farsnews.ir/news/14001112000715/
doc number: 6764 -> title: url: https://www.farsnews.ir/news/14000925/
doc number: 6128 -> title: url: https://www.farsnews.ir/news/14001021000577/
doc number: 4918 -> title: url: https://www.farsnews.ir/news/14001021000577/
doc number: 1950 -> title: url: https://www.farsnews.ir/news/14001128000554/
doc number: 2268 -> title: url: https://www.farsnews.ir/news/1400112300/
doc number: 6323 -> title: url: https://www.farsnews.ir/news/14001002/
doc number: 5102 -> title: url: https://www.farsnews.ir/news/14001019000346/
doc number: 139 -> title: url: https://www.farsnews.ir/news/14001222000297/
doc number: 5583 -> title: url: https://www.farsnews.ir/news/14001012000661/
doc number: 4603 -> title: url: https://www.farsnews.ir/news/14001025000363/
doc number: 1824 -> title: url: https://www.farsnews.ir/news/14001130000599/
doc number: 2626 -> title: url: https://www.farsnews.ir/news/14001/
doc number: 5252 -> title: url: https://www.farsnews.ir/news/14001017/
doc number: 7305 -> title: url: https://www.farsnews.ir/news/14001210000621/
doc number: 5221 -> title: url: https://www.farsnews.ir/news/14001018000261/
doc number: 5005 -> title: url: https://www.farsnews.ir/news/140010200000/
doc number: 1974 -> title: url: https://www.farsnews.ir/news/140010200000/
```

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.