# سیستم بازیابی اطلاعات فارسی بر روی دیتاست IR_data_news_12k.json

**میرنیما قاسمیان - ۹۷۱۱۵۲۱۶۴**

---

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' ]

## در ادامه به بررسی کوئری هایی برای ارزیابی قابلیت سرچ کوئری می پردازیم:

**الف) پرسمان ساده**
کوئری : تحریم‌های آمریکا علیه ایران

```
Rank 1:
title:  خبرگزاری فارس ۱۹ ساله شد
url:  https://www.farsnews.ir/news/14001122000809/خبرگزاری-فارس-۱۹-ساله-شد
------------
Rank 2:
title:  اصولی: فدراسیون فوتبال جمهوری اسلامی ایران هستیم نه جزیره مستقل/ با گفتار ساختارشکنانه فدراسیون را به ناکجا آباد می‌برند
url:  https://www.farsnews.ir/news/14001117000518/اصولی-فدراسیون-فوتبال-جمهوری-اسلامی-ایران-هستیم-نه-جزیره-مستقل-با
------------
Rank 3:
title:  احتمال مبادله نازنین زاغری در ازای 530میلیون دلار
url:  https://www.farsnews.ir/news/14001223001080/احتمال-مبادله-نازنین-زاغری-در-ازای-530میلیون-دلار
------------
Rank 4:
title:  متکی: آمریکا با ابزار ناتو به دنبال تجزیه روسیه است
url:  https://www.farsnews.ir/news/14001222000749/متکی-آمریکا-با-ابزار-ناتو-به-دنبال-تجزیه-روسیه-است
------------
Rank 5:
title:  توضیحات یک منبع آگاه درباره وقفه مذاکرات وین
url:  https://www.farsnews.ir/news/14001222000450/توضیحات-یک-منبع-آگاه-درباره-وقفه-مذاکرات-وین
------------
```

باید به گونه‌ای باشد که همه کلمات در وبلاگ‌های مورد نظر باشد و آن که 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.

می‌کند. * تحریم‌های آمریکا علیه ایران آمریکا پس از تسخیر لانه جاسوسی آمریکا در تهران، «کارتر» در تاریخ 8 نوامبر 1979م. با استناد به قانون «کنترل صدور تسلیحات نظامی»، کشتی حامل لوازم یدکی نظامی متعلق به ایران را توقیف کرد. ارزش این لوازم 300 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 (12170)، تمام دارایی‌های ایران در آمریکا را به تصرف خود درآورد. مجموع پس‌اندازها و اوراق بهادار بلوکه شده‌ی ایران بالغ بر 12 میلیارد دلار بود. حدود 1.4میلیارد دلار در بانک «فدرال رزرو» آمریکا، 5.

**ب) یک پرسمان با عملگر NOT**

کوئری : تحریم‌های آمریکا ! ایران

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

```
Rank 1:
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
------------
Rank 2:
title:  انتقاد دانشجویان ایرانی در اروپا به برخورد دوگانه مدعیان حقوق بشر با قضایای اوکراین و جنایت‌های آل سعود
url:  https://www.farsnews.ir/news/14001224000014/انتقاد-دانشجویان-ایرانی-در-اروپا-به-برخورد-دوگانه-مدعیان-حقوق-بشر-با
------------
Rank 3:
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
------------
Rank 4:
title:  تجربه نشان داده به عهد آمریکا در مذاکرات نمی‌شود اعتماد کرد
url:  https://www.farsnews.ir/news/14001203000366/تجربه-نشان-داده-به-عهد-آمریکا-در-مذاکرات-نمی‌شود-اعتماد-کرد
------------
Rank 5:
title:  سود مافیای اسلحه‌سازی آمریکا در ناامن بودن جهان است
url:  https://www.farsnews.ir/news/14001211000898/سود-مافیای-اسلحه‌سازی-آمریکا-در-ناامن-بودن-جهان-است
------------
```

لایحه‌ای به مجلس برده و با تصویب آن واردات نفت از ایران را ممنوع کردند. ریگان که نمی‌خواست کمتر از کنگره ضدتروریست جلوه کند، 3هفته بعد با صدور دستور ویژه‌ی (12613) ورود هرگونه کالا و خدمات از ایران را ممنوع کرد. وی برای صدور این دستور به بند 505 «قانون همکاری‌های بین‌المللی امنیتی و توسعه»، مصوب سال 1985م. 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. * پایان جنگ، تداوم تحریم با پذیرش قطعنامه‌ی (598) توسط

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.

در سیستم‌های بازیابی اطلاعات، برای بهبود دقت و کارایی جستجو، از **شاخص‌گذاری موقعیتی و وزن‌دهی اسناد** استفاده می‌شود. در این بخش، **پیاده‌سازی شاخص‌گذاری موقعیتی همراه با محاسبه وزن‌های TF-IDF و ایجاد لیست قهرمانان** ارائه شده است. هدف این بخش، پردازش و تحلیل اسناد برای بازیابی سریع و موثر اطلاعات مرتبط با کوئری‌های کاربر است.

## محاسبه وزن‌دهی اسناد و عبارات جستجو

فرآیند وزن‌دهی شامل دو معیار اصلی است:

### TF (Term Frequency):

نشان‌دهنده تعداد دفعات تکرار یک کلمه در سند است. مقدار TF بر اساس فرمول استاندارد $log10(freq) + 1$ محاسبه شده تا تأثیر کلمات پرتکرار کاهش یابد.

### IDF (Inverse Document Frequency):

میزان اهمیت یک کلمه در کل مجموعه اسناد را تعیین می‌کند. این مقدار طبق رابطه $log10(N/n)$ محاسبه می‌شود که در آن N تعداد کل اسناد و n تعداد اسنادی است که شامل آن کلمه هستند.

با استفاده از این دو معیار، مقدار TF-IDF برای هر کلمه در اسناد محاسبه شده و از آن برای وزن‌دهی به عبارات در فرآیند بازیابی استفاده می‌شود.

### ساخت لیست قهرمانان (Champions List)

برای افزایش کارایی بازیابی اطلاعات، لیستی از مهم‌ترین اسناد برای هر کلمه ساخته می‌شود. این لیست شامل r سند با بالاترین وزن TF-IDF برای هر کلمه است که به کاهش تعداد اسنادی که در فرآیند جستجو بررسی می‌شوند، کمک می‌کند.

### نرمال‌سازی اسناد

از آنجایی که طول اسناد متفاوت است، برای مقایسه بهتر وزن‌ها، یک بردار نرمال برای هر سند محاسبه می‌شود. در این مرحله، مجموع مربعات وزن‌های کلمات هر سند محاسبه شده و جذر آن به عنوان مقدار نرمال‌سازی‌شده در نظر گرفته می‌شود.

## نتیجه‌گیری

در این بخش، مکانیزم شاخص‌گذاری موقعیتی و وزن‌دهی به اسناد پیاده‌سازی شده است که به بهبود دقت بازیابی و افزایش سرعت پردازش جستجو کمک می‌کند. مراحل انجام‌شده شامل محاسبه وزن‌های TF-IDF، ایجاد لیست قهرمانان و نرمال‌سازی اسناد است که همگی در راستای بهینه‌سازی سیستم جستجو و کاهش هزینه پردازشی انجام گرفته‌اند.

در زیر چند مثال برای بررسی کارایی سیستم جست و جوی خود را بررسی خواهیم کرد.

### ۱- الف) یک کوئری از کلمات ساده و متداول تک کلمه ای:

```
query = 'فوتبال'
query_tokens = preprocess([query], True, True)[0]
scores = cosine_similarity(query_tokens, positional_index, len(contents), docs_norms)
print_result(scores, k=5)
```

✓ 0.0s

```
1466 title:  نکونام: نفتی ها بهترین بازی خود را انجام دادند/بازیکنان جدیدمان کیفیت بالای خود را نشان دادند
url:  https://www.farsnews.ir/news/14001204001165/نکونام-نفتی-ها-بهترین-بازی-خود-را-انجام-دادند-بازیکنان-جدیدمان-کیفیت
------------
81 title:  ماجدی: فوتبال کشور به تغییرات نیاز دارد
url:  https://www.farsnews.ir/news/14001223000539/ماجدی-فوتبال-کشور-به-تغییرات-نیاز-دارد
------------
6690 title:  امیدواری ملی پوش سابق فوتبال ساحلی بابت تغییرات در کادرفنی تیم ملی
url:  https://www.farsnews.ir/news/14000927000275/امیدواری-ملی-پوش-سابق-فوتبال-ساحلی-بابت-تغییرات-در-کادرفنی-تیم-ملی
------------
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/تقدیر-مربی-تیم-فوتبال-خلیج-فارس-از-ماجدی-بابت-رسیدگی-به-شائبه-تبانی-در
------------
860 title:  دیدار مدیر تیم ملی ایران با سفیر کره جنوبی/ هیان: امیدوارم مقابل ایران مساوی کنیم
url:  https://www.farsnews.ir/news/14001213000138/دیدار-مدیر-تیم-ملی-ایران-با-سفیر-کره-جنوبی-هیان-امیدوارم-مقابل-ایران
------------
```

همانطور که از تیتر اخبار مشخص است همگی مربوط به فوتبال می باشند. به عنوان مثال در تیتر خبر اول کلمه فوتبال مشاهده نمی شود اما اگر خود خبر را بخوانیم بارها کلمه فوتبال را میبینیم:

سرمربی تیم فوتبال فولاد خوزستان درخصوص تغییراتی که در فدراسیون فوتبال ایجاد شده است، تصریح کرد: به عنوان کسی که در فوتبال هستم فقط می‌توانم نظرم را بگویم کاری به تغییرات و اینکه عزیزی‌خادم چه کاری کرده و چه کاری نکرده ندارم، اما امیدوارم فوتبال ما رو به جلو حرکت کند، چون اگر اینگونه شود برای همه خوب است. نکونام افزود: تجربه نشان داده کسانی که مدیریت فوتبالی داشته‌اند و توانسته‌اند کاری بکنند در بدنه فوتبال بوده‌اند. ماجدی آمده و همه ما خوشحال هستیم، چون او هم مثل دادکان فوتبالی است. امیدوارم فوتبال ما رو به جلو حرکت کند. ماجدی را سال‌های زیادی است که می‌شناسیم؛ او رشد کرده و اگر قرار است در فدراسیون بماند که دوست داریم بماند امیدوارم تمام تصمیمات فوتبالی باشد. فدراسیون همه کاره فوتبال است و دوست داریم که میرشاد ماجدی عزیز موفق باشد، چون از بدنه فوتبال است و درد فوتبال را می‌داند. انتهای پیام/.

## ب) یک کوئری از عبارات ساده و متداول چند کلمه ای:

```
query = 'تیم ملی فوتبال'
query_tokens = preprocess([query], True, True)[0]
scores = cosine_similarity(query_tokens, positional_index, len(contents), docs_norms)
print_result(scores, k=5)
```

✓ 0.0s

```
142 title:  اسکوچیچ: مردم متوجه شده‌اند که می‌توانند هدایت تیم ملی را به من واگذار کنند/هاشمیان سواد اروپایی از فوتبال دارد
url:  https://www.farsnews.ir/news/14001222000329/اسکوچیچ-مردم-متوجه-شده‌اند-که-می‌توانند-هدایت-تیم-ملی-را-به-من-واگذار
------------
2098 title:  مصاحبه فارس با کارشناس فوتبال آسیا| از میراث بزرگ کیروش و قدرت ایران با اسکوچیچ تا انقلاب برانکو در عمان
url:  https://www.farsnews.ir/news/14001124000522/مصاحبه-فارس-با-کارشناس-فوتبال-آسیا|-از-میراث-بزرگ-کیروش-و-قدرت-ایران
------------
1466 title:  نکونام: نفتی ها بهترین بازی خود را انجام دادند/بازیکنان جدیدمان کیفیت بالای خود را نشان دادند
url:  https://www.farsnews.ir/news/14001204001165/نکونام-نفتی-ها-بهترین-بازی-خود-را-انجام-دادند-بازیکنان-جدیدمان-کیفیت
------------
81 title:  ماجدی: فوتبال کشور به تغییرات نیاز دارد
url:  https://www.farsnews.ir/news/14001223000539/ماجدی-فوتبال-کشور-به-تغییرات-نیاز-دارد
------------
6690 title:  امیدواری ملی پوش سابق فوتبال ساحلی بابت تغییرات در کادرفنی تیم ملی
url:  https://www.farsnews.ir/news/14000927000275/امیدواری-ملی-پوش-سابق-فوتبال-ساحلی-بابت-تغییرات-در-کادرفنی-تیم-ملی
------------
```

همانطور که از تیتر اخبار مشخص است همگی درباره ی تیم ملی فوتبال است.

## پ) یک کوئری دشوار و کم تکرار تک کلمه ای:

```
query = 'واترپلو'
query_tokens = preprocess([query], True, True)[0]
scores = cosine_similarity(query_tokens, positional_index, len(contents), docs_norms)
print_result(scores, k=5)
```

✓ 0.0s

```
1388 title:  تجلیل از خانواده شهید حسن نوفلاح با تقدیم مدال قهرمانی
url:  https://www.farsnews.ir/news/14001205000938/تجلیل-از-خانواده-شهید-حسن-نوفلاح-با-تقدیم-مدال-قهرمانی
------------
5690 title:  سرپرست  فدراسیون شنا، شیرجه و واترپلو  منصوب شد
url:  https://www.farsnews.ir/news/14001011000202/سرپرست-فدراسیون-شنا-شیرجه-و-واترپلو-منصوب-شد
------------
5022 title:  رضوانی رئیس فدراسیون شنا ماند/ 3 رای سفید برای تنها کاندیدا
url:  https://www.farsnews.ir/news/14001020000253/رضوانی-رئیس-فدراسیون-شنا-ماند-3-رای-سفید-برای-تنها-کاندیدا
------------
5013 title:  برنامه‌های رئیس فدراسیون شنا برای 4 سال آینده/ رضوانی: تلاش می‌کنیم به اهدافمان در بازی‌های آسیایی برسیم
url:  https://www.farsnews.ir/news/14001020000394/برنامه‌های-رئیس-فدراسیون-شنا-برای-4-سال-آینده-رضوانی-تلاش-می‌کنیم-به
------------
1300 title:  گرفتن ۷۰ نمونه تست دوپینگ در ۸ رشته طی یک هفته
url:  https://www.farsnews.ir/news/14001207000351/گرفتن-۷۰-نمونه-تست-دوپینگ-در-۸-رشته-طی-یک-هفته
```

این کلمه کم تکرار است. همانطور که مشاهده می شود از تیتر اخبار مشخص است که مربوط به واترپلو می باشند.

## ت) یک کوئری دشوار و کم تکرار چند کلمه ای:

```
query = 'واکسن کرونا ایرانی'
query_tokens = preprocess([query], True, True)[0]
scores = cosine_similarity(query_tokens, positional_index, len(contents), docs_norms)
print_result(scores, k=5)
```

✓ 0.0s

```
7937 title:  مرندی: رهبر انقلاب دُز سوم واکسن کرونا را دریافت کرده‌اند
url:  https://www.farsnews.ir/news/14001117000930/مرندی-رهبر-انقلاب-دُز-سوم-واکسن-کرونا-را-دریافت-کرده‌اند
------------
11966 title:  الهیان: وزارت بهداشت درباره علت عدم پیش خرید واکسن فخرا توضیح دهد
url:  https://www.farsnews.ir/news/14000729000462/الهیان-وزارت-بهداشت-درباره-علت-عدم-پیش-خرید-واکسن-فخرا-توضیح-دهد
------------
7228 title:  رئیسی: مسافرت با رعایت اصول بهداشتی بلااشکال است
url:  https://www.farsnews.ir/news/14001214000475/رئیسی-مسافرت-با-رعایت-اصول-بهداشتی-بلااشکال-است
------------
9835 title:  نایب رئیس مجلس: دولت در تزریق واکسن و کاهش نگرانی‌های مردم شاهکار کرده است
url:  https://www.farsnews.ir/news/14000923000300/نایب-رئیس-مجلس-دولت-در-تزریق-واکسن-و-کاهش-نگرانی‌های-مردم-شاهکار
------------
9736 title:  تأکید مخبر بر حمایت از تولیدکنندگان داخلی واکسن کرونا
url:  https://www.farsnews.ir/news/14000924000889/تأکید-مخبر-بر-حمایت-از-تولیدکنندگان-داخلی-واکسن-کرونا
------------
```

این عبارت کم تکرار است. همانطور که میبینیم نتایج به دست آمده تا حد زیادی مرتبط است. مثلا خبر اول که بالاترین امتیاز را دارد درباره این است که رهبر واکسن ایرانی زده است و خبر دوم درباره ی واکسن فخرا است که یک واکسن ایرانی است

## مقایسه با نتایج قبل از پیاده سازی مکانیزم شاخص‌گذاری موقعیتی و وزن‌دهی به اسناد:

```
query = 'واکسن کرونا ایرانی'
res = search_query(query)
print_output(res)
```

```
Rank 1:
title:  شروط حضور هواداران در تمام ورزشگاه‌های کشور/ مهدی: تغییر زمان دربی نهایی نشده است
url:  https://www.farsnews.ir/news/14001217000675/شروط-حضور-هواداران-در-تمام-ورزشگاه‌های-کشور-مهدی-تغییر-زمان-دربی
------------
Rank 2:
title:  نحوه توزیع سکوهای آزادی برای تماشاگران در دیدار ایران-عراق از زبان خبرنگار عراقی
url:  https://www.farsnews.ir/news/14001106000376/نحوه-توزیع-سکوهای-آزادی-برای-تماشاگران-در-دیدار-ایران-عراق-از-زبان
------------
Rank 3:
title:  وزیر کشور: ممکن است دوز سوم واکسن برای سفرهای خارجی اجباری شود
url:  https://www.farsnews.ir/news/14001217000520/وزیر-کشور-ممکن-است-دوز-سوم-واکسن-برای-سفرهای-خارجی-اجباری-شود
------------
Rank 4:
title:  سخنگوی دولت: ۹۹ درصد سواحلی که در اختیار نهادهای دولتی بود آزاد شد
url:  https://www.farsnews.ir/news/14001212000835/سخنگوی-دولت-۹۹-درصد-سواحلی-که-در-اختیار-نهادهای-دولتی-بود-آزاد-شد
------------
Rank 5:
title:  همایش علمی دانشگاهیان ایرانی خارج کشور| حمایت فارغ‌التحصیلان ایرانی دانشگاه‌های برتر دنیا برای بازگشت به کشور
url:  https://www.farsnews.ir/news/14001212000242/همایش-علمی-دانشگاهیان-ایرانی-خارج-کشور|-حمایت-فارغ‌التحصیلان-ایرانی
------------
```

همانطور که مشاهده میکنیم نتایج تقریبا نامربوط اند.

نتایج بدست آمده برای کوئری دشوار و کم تکرار، تقریبا نامرتبط است. در ابتدا ما اسنادی را که حاوی کلمات کوئری هستند را بازیابی میکنیم و اسناد را بر اساس اینکه در کدام سند تعداد بیشتری از کلمات کوئری آمده است مرتب میکنیم. برای مثال در کوئری «واکسن کرونا ایرانی» اسنادی را که کلمات واکسن»، «کرونا» و «ایرانی را دارند در رتبه های بالاتر قرار میگیرند و اسنادی که بعضی از این کلمات را ندارند در رتبه های بندی قرار میگیرند.

در پس از اینکه ما پیاده سازی مکانیزم شاخص‌گذاری موقعیتی و وزن‌دهی به اسناد انجام دادیم تعداد تکرار یک کلمه در یک سند و اهمیت آن کلمه در کل اسناد را با استفاده از tf-idf پیدا و اسناد می کنیم و با استفاده از شباهت کسینوسی اسناد را از نظر شباهت مقایسه می کنیم. پس هر کلمه یک وزن دارد و اسناد بر اساس وزن و اهمیت کلمات موجود در کوئری بازیابی و رتبه بندی می شوند.

برای همین است که ما در فاز دوم نتایج دقیق تری گرفتیم و توانستیم به خوبی نتایج را رتبه دهی کنیم.