((گزارش فاز دوم پروژه بازیابی اطلاعات))

پویان حسابی ۹۷۳۱۱۲۲

۱)

فرمت کلی نمایش:

```
print("***\nResult " + str(index + 1))
doc_id = similarity_array.index(sorted_array[index])
print("Document ID: " + str(doc_id))
print("Score: " + str(sorted_array[index]))
print("URL: " + str(urls[doc_id]))
print("Title: " + str(url_title[urls[doc_id]]))
print("Number of all token in the document: " + str(len(docs_term_score[doc_id])))
print("Number of given query word: "
```

که در آن شماره سند، امتیاز، url، عنوان خبر، تعداد کل توکن های سند و تعداد تکرار کلمه مربوطه ذکر شده و در تحلیل ها از عوامل بسیار مهم تلقی می‌شود.

الف) یک پرسمان از کلمات ساده و متداول تک کلمه ای: "تحریم"

Result 1

Document ID: 9742

Score: 0.0019854512336915132

URL: https://www.farsnews.ir/news/14000924000773/مرکز-پژوهش‌های-مجلس-مذاکرات-وین-به-توافقی-زودهنگام-منجر-نمی‌شود

Title: مرکز پژوهش‌های مجلس: مذاکرات وین به توافقی زودهنگام منجر نمی‌شود

Number of all token in the document: 594

Number of given query word: 40

***

Result 2

Document ID: 10806

Score: 0.0019358861670254702

URL: https://www.farsnews.ir/news/14000808000646/دولت-خاتمی-۶-میلیارد-دلار-پول-نفت-را-به-خزانه-واریز-نکرد-رئیسی-۱۰۰

Title: دولت خاتمی ۶ میلیارد دلار پول نفت را به خزانه واریز نکرد/ رئیسی ۱۰۰ روز اول، برادری‌اش را به مردم ثابت کرد

Number of all token in the document: 1038

Number of given query word: 16

***

Result 3

Document ID: 6929

Score: 0.0018817633930797903

URL: https://www.farsnews.ir/news/14001222000450/توضیحات-یک-منبع-آگاه-درباره-وقفه-مذاکرات-وین

Title: توضیحات یک منبع آگاه درباره وقفه مذاکرات وین

Number of all token in the document: 587

Number of given query word: 43

***

Result 4

Document ID: 9418

Score: 0.0018213737716246 11

URL: https://www.farsnews.ir/news/14001005000843/بازخوانی-|-بیانات-رهبر-انقلاب-در-دیدار-دست‌اندرکاران-مراسم-سالگرد

Title: بازخوانی | بیانات رهبر انقلاب در دیدار دست‌اندرکاران مراسم سالگرد شهادت سردار سلیمانی

Number of all token in the document: 586

Number of given query word: 10

***

Result 5

Document ID: 7731

Score: 0.0016783757057260258

URL: https://www.farsnews.ir/news/14001123000508/نشست-بررسی-روند-مذاکرات|-آمریکا-استاد-دور-زدن‌های-بین‌المللی-است

Title: نشست بررسی روند مذاکرات| آمریکا استاد دور زدن‌های بین‌المللی است

Number of all token in the document: 510

Number of given query word: 19

تحلیل: نتایج سند اول و سوم را به طور مثال بررسی می‌کنیم:

سند اول:

تعداد تکرار کلمه تحریم: ۴۱، تعداد حدودی کل کلمات: ۳۲۵۳

سند سوم:

تعداد تکرار کلمه تحریم: ۶۰، تعداد حدودی کل کلمات: ۳۹۸۶

بقیه اسناد هم مشابه همین رویکرد و تعداد می‌باشند. و اما چرا در سند سوم کلمه تحریم تعداد بیشتری تکرار شده ولی در رتبه سوم قرار دارد؟!

همانطور که در درس گفته شد در وزن دهی اسناد غیر از tf، idf هم بررسی می‌شود و به نوعی نشان می‌دهد چقدر کلمه rare می‌باشد. از طرفی تعداد کلمات سند سوم به مراتب بیشتر از سند اول می‌باشد و این باعث می‌شود امتیاز کمتری بگیرد. از طرفی بر اساس تعداد توکن های کل سند و تعداد تکرار کلمات آن تا حد زیادی نتیجه درستی نشان می‌دهد ولی به طور مثال در سند سوم باید نتیجه بهتری میداده است که احتمالا بخاطر پیش پردازش تعداد کمتری از کلمات آن حذف شده است.

نداده است.

ب) یک پرسمان از عبارات ساده و متداول چند کلمه ای: "جمهوری اسلامی"

***

Result 1

Document ID: 9266

Score: 0.003592884428075525

URL: https://www.farsnews.ir/news/14001008000978/بُرشی-از-کتاب-فتنه-تغلب-روایت‌هایی-از-متن-و-فرامتن-آشوب‌های-۸۸

Title: بُرشی از کتاب « فتنه تغلب»/ روایت‌هایی از متن و فرامتن آشوب‌های ۸۸

Number of all token in the document: 2449

Number of given query word: 110

***

Result 2

Document ID: 7789

Score: 0.003541253421152678

URL: https://www.farsnews.ir/news/14001121000223/رئیسی-بدنبال-فصل-جدیدی-از-همکاری‌های-دو-و-چندجانبه-هستیم

Title: رئیسی: بدنبال فصل جدیدی از همکاری‌های دو و چندجانبه هستیم

Number of all token in the document: 750

Number of given query word: 67

***

Result 3

Document ID: 7778

Score: 0.003464339947816835

URL: https://www.farsnews.ir/news/14001121000352/دعوت-احزاب-و-شخصیت‌های-سیاسی-از-مردم-برای-حضور-در-راهپیمایی-۲۲بهمن

Title: دعوت احزاب و شخصیت‌های سیاسی از مردم برای حضور در راهپیمایی ۲۲بهمن

Number of all token in the document: 1266

Number of given query word: 104

***

Result 4

Document ID: 9671

Score: 0.0034454662539566607

URL: https://www.farsnews.ir/news/14000927000679/جزئیات-نامه-لاریجانی-به-شورای-نگهبان-درباره-رد-صلاحیتش

Title: جزئیات نامه لاریجانی به شورای نگهبان درباره رد صلاحیتش

Number of all token in the document: 1999

Number of given query word: 72

***

Result 5

Document ID: 11697

Score: 0.0034054150389602263

URL: https://www.farsnews.ir/news/14000804000796/پرهیز-شدید-آیت‌الله-مهدوی‌کنی-از-رانت-و-سفارش-از-علاقه-خاص-حضرت-امام

Title: پرهیز شدید آیت‌الله مهدوی‌کنی از رانت و سفارش/ از علاقه خاص حضرت امام(ره) تا اصرار بر تربیت دانشجو ـ سرباز برای اسلام

Number of all token in the document: 1999

Number of given query word: 118

تحلیل: بنظر می‌رسد مدل برای داده اول اشتباه عمل کرده است که در نتیجه اول قرار دارد در حالی که سند های دوم و سوم بنظر می‌رسد به مراتب بهتر باشند و تفاوت چندانی دارند. البته در ادامه این روند درست می‌شود و تعداد کلمات و تعداد تکرار کلمات کوئری نزدیک می‌شود. این نشان می‌دهد با افزودن کلمات به کوئری که کلمات متداول و پشت سر همی هم هستند دقت مدل افزایش می‌یابد.

پ) یک پرسمان دشوار و کم تکرار تک کلمه ای: اعتیاد

***

Result 1

Document ID: 11159

Score: 0.0013593361961514929

URL: https://www.farsnews.ir/news/14000821000211/بازدید-رئیس-جمهور-از-کمپ-معتادان-در-شهرری-رئیسی-زیبنده-تهران-و-ری

Title: بازدید رئیس جمهور از کمپ معتادان در شهرری/ رئیسی: زیبنده تهران و ری نیست که افراد سرپناه نداشته باشند

Number of all token in the document: 101

Number of given query word: 4

***

Result 2

Document ID: 4689

Score: 0.0011958284474945446

URL: https://www.farsnews.ir/news/14001023000854/محسن-قاسمی-موطلایی-کشتی-ایران-درگذشت

Title: محسن قاسمی» موطلایی کشتی ایران درگذشت«

Number of all token in the document: 89

Number of given query word: 3

***

Result 3

Document ID: 5322

Score: 0.0009645248030530081

URL: https://www.farsnews.ir/news/14001016000156/زمان-ثبت‌نام-از-نامزدهای-ریاست-فدراسیون-تیراندازی-با-کمان-اعلام-شد

Title: زمان ثبت‌نام از نامزدهای ریاست فدراسیون تیراندازی با کمان اعلام شد

Number of all token in the document: 204

Number of given query word: 2

***

Result 4

Document ID: 3982

Score: 0.0005697134099545234

URL: https://www.farsnews.ir/news/14001102000705/حاشیه-بازی-پرسپولیس-و-فولاد-|-خوش-و-بش-حسینی-با-بازیکن-سابق-تیم-ملی

Title: حاشیه بازی پرسپولیس و فولاد | خوش و بش حسینی با بازیکن سابق تیم ملی/ هتریک یک سرخ‌پوش در بیرون ماندن از فهرست

Number of all token in the document: 130

Number of given query word: 1

تحلیل: برخلاف دو کوئری قبل، مدل در این کوئری بسیار بهتر برخورد کرده است. اگرچه نتایج کمی را نشان داده و در هر کدام هم تعداد بار کمی این کلمه تکرار شده است ولی باید این را در نظر داشت که خبرگزاری فارس بیشتر از اخبار سیاسی و ورزشی گزارش می‌دهد و از نظر ارتباط کوئری و لینک خبر هم، سند اول ارتباط بسیار بالایی به موضوع اعتیاد دارد.

ت) یک پرسمان دشوار و کم تکرار چند کلمه ای: دانشگاه صنعتی امیرکبیر

***

Result 1

Document ID: 8125

Score: 0.002241379769227726

URL: https://www.farsnews.ir/news/14001112000587/پیاده‌روی-دانشگاهیان-دانشگاه-امیرکبیر-به-مناسبت-آغاز-دهه-فجر

Title: پیاده‌روی دانشگاهیان دانشگاه امیرکبیر به مناسبت آغاز دهه فجر

Number of all token in the document: 99

Number of given query word: 13

***

Result 2

Document ID: 7579

Score: 0.002058809472002783

URL: https://www.farsnews.ir/news/14001201000323/نامه-۱۳۷-تشکل-دانشجویی|-سامانه-حراج-بازتوزیع-کالای-قاچاق-است

Title: نامه ۱۳۷ تشکل دانشجویی| «سامانه حراج» بازتوزیع کالای قاچاق است

Number of all token in the document: 531

Number of given query word: 95

***

Result 3

Document ID: 11928

Score: 0.0020042819210129422

URL: https://www.farsnews.ir/news/14000801000437/نامه-جمعی-از-اساتید-و-متخصصان-آقای-رئیس‌جمهور-در-گام-دوم-انقلاب-به

Title: نامه جمعی از اساتید و متخصصان/ آقای رئیس‌جمهور در گام دوم انقلاب به داد «مدیریت» در کشور برسید

Number of all token in the document: 1193

Number of given query word: 51

***

Result 4

Document ID: 11694

Score: 0.001970100537307398

URL: https://www.farsnews.ir/news/14000810000142/نامه-بیش-از-۶۰-تشکل-دانشجویی-درخواست-انتصاب-رئیسی-نخبه-و-تحول‌خواه

Title: نامه بیش از ۶۰ تشکل دانشجویی/ درخواست انتصاب رئیسی نخبه و تحول‌خواه برای سازمان فرهنگ و ارتباطات

Number of all token in the document: 230

Number of given query word: 60

***

Result 5

Document ID: 9261

Score: 0.001498309683598833

URL: https://www.farsnews.ir/news/14001009000255/نشست-اساتید-با-رئیس-سازمان-انرژی-اتمی--تاکید-بر-نقش-نخبگان-در-اجرای

Title: نشست اساتید با رئیس سازمان انرژی اتمی / تاکید بر نقش نخبگان در اجرای سند جامع انرژی هسته‌ای کشور

Number of all token in the document: 170

Number of given query word: 18

تحلیل: از آنجایی که نباید انتظار داشته باشیم مطلب علمی یا مطلبی به طور خاص راجع به دانشگاه صنعتی امیرکبیر برگرداند این نتایج منطقی است. دلیل آوردن آنها نامه هایی است که دانشجویان امیرکبیر در آنها مشارکت کردند. مقدار امتیاز ها اعداد بالایی هستند به طور مثال سند اول دارای امتیاز ۰.۰۰۲۲ می‌باشد که این بدان معنی است از نظر تکرار کلمه در این سند و اختصاصی بودن آن نتیجه قابل توجهی است. البته همچنان کمی مشکل رتبه بندی و خطا وجود دارد بین سند اول و دوم که احتمالا بخاطر ریشه یابی نا دقیق این کتابخانه می‌باشد.

۲) تکرار ب در فاز اول: یک پرسمان از عبارات ساده و متداول چند کلمه ای: "جمهوری اسلامی"

فاز اول:

فاز دوم:

Result 1

Document ID: 7778

Title: دعوت احزاب و شخصیت‌های سیاسی از مردم برای حضور در راهپیمایی ۲۲بهمن

Result 2

Document ID: 7651

Title: انقلابی مانند انقلاب اسلامی مورد حمایت جریان‌های فرامذهبی دنیا قرار نگرفته است

Result 3

Document ID: 9266

Title: بُرشی از کتاب « فتنه تغلب»/ روایت‌هایی از متن و فرامتن آشوب‌های ۸۸

Result 4

Document ID: 10174

Title: تحقق فرآیند مکتب انقلاب اسلامی وظیفه همه مسؤولان است/ گرفتاری‌های اقتصادی به دلیل اجرای دستورات لیبرالی است

Result 5

Document ID: 11697

Title: پرهیز شدید آیت‌الله مهدوی‌کنی از رانت و سفارش/ از علاقه خاص حضرت امام(ره) تا اصرار بر تربیت دانشجو ـ سرباز برای اسلام

Result 1

Document ID: 9266

Title: بُرشی از کتاب « فتنه تغلب»/ روایت‌هایی از متن و فرامتن آشوب‌های ۸۸

Number of all token in the document: 2449

Number of given query word: 110

Result 2

Document ID: 7789

Title: رئیسی: بدنبال فصل جدیدی از همکاری‌های دو و چندجانبه هستیم

Number of all token in the document: 750

Number of given query word: 67

Result 3

Document ID: 7778

Title: دعوت احزاب و شخصیت‌های سیاسی از مردم برای حضور در راهپیمایی ۲۲بهمن

Number of all token in the document: 1266

Number of given query word: 104

Result 4

Document ID: 9671

Title: جزئیات نامه لاریجانی به شورای نگهبان درباره رد صلاحیتش

Number of all token in the document: 1999

Number of given query word: 72

Result 5

Document ID: 11697

Title: پرهیز شدید آیت‌الله مهدوی‌کنی از رانت و سفارش/ از علاقه خاص حضرت امام(ره) تا اصرار بر تربیت دانشجو ـ سرباز برای اسلام

Number of all token in the document: 1999

تحلیل: اسنادی که با هم تفاوت دارند:

Document ID: 7789

Title: رئیسی: بدنبال فصل جدیدی از همکاری‌های دو و چندجانبه هستیم

Document ID: 9671

Title: جزئیات نامه لاریجانی به شورای نگهبان درباره رد صلاحیتش

دو سند بالا فقط در فاز دوم آمده اند و دو سند پایین فقط در فاز اول:

Document ID: 10174

Title:تحقق فرآیند مکتب انقلاب اسلامی وظیفه همه مسؤولان است/ گرفتاری‌های اقتصادی به دلیل اجرای دستورات لیبرالی است

Document ID: 7651

Title: انقلابی مانند انقلاب اسلامی مورد حمایت جریان‌های فرامذهبی دنیا قرار نگرفته است

به مقایسه آنها می‌پردازیم سند ۷۶۵۱ بسیار مرتبط با کوئری می‌باشد که در فاز دوم نیامده است ولی از نظر عملکرد بنظر می‌آید شباهت زیادی دارد. البته بر اساس تعداد کلمات و امتیاز ها متوجه می‌شویم در حالت اول فقط تعداد تکرار کلمه در سند بررسی شده است در حالی که در حالت دوم علاوه بر تکرار طول کوئری، طول سند و میزان rare بودن آن هم در نظر گرفته شده است.

۲) تکرار ت در فاز اول: یک پرسمان دشوار و کم تکرار چند کلمه ای: "دانشگاه صنعتی امیرکبیر"

فاز اول:

Result 1

Document ID: 7579

Title: نامه ۱۳۷ تشکل دانشجویی| «سامانه حراج» بازتوزیع کالای قاچاق است

Result 2

Document ID: 11694

Title: نامه بیش از ۶۰ تشکل دانشجویی/ درخواست انتصاب رئیسی نخبه و تحول‌خواه برای سازمان فرهنگ و ارتباطات

Result 3

Document ID: 11928

Title: نامه جمعی از اساتید و متخصصان/ آقای رئیس‌جمهور در گام دوم انقلاب به داد «مدیریت» در کشور برسید

Result 4

Document ID: 9286

Title: فتنه ۸۸؛ گناه نابخشودنی| روایت «آذر آمریکایی» اصلاح‌طلبان تندرو / از درخواست تحریم تا اهانت به امام (ره)

Result 5

Document ID: 11603

Title: مسلمانان و دولت‌های اسلامی موظف به مقابله با ظلم و استکبار هستند

فاز دوم:

Result 1

Document ID: 8125

Title: پیاده‌روی دانشگاهیان دانشگاه امیرکبیر به مناسبت آغاز دهه فجر

Number of all token in the document: 99

Number of given query word: 13

Result 2

Document ID: 7579

Title: نامه ۱۳۷ تشکل دانشجویی| «سامانه حراج» بازتوزیع کالای قاچاق است

Number of all token in the document: 531

Number of given query word: 95

Result 3

Document ID: 11928

Title: نامه جمعی از اساتید و متخصصان/ آقای رئیس‌جمهور در گام دوم انقلاب به داد «مدیریت» در کشور برسید

Number of all token in the document: 1193

Number of given query word: 51

Result 4

Document ID: 11694

Title: نامه بیش از ۶۰ تشکل دانشجویی/ درخواست انتصاب رئیسی نخبه و تحول‌خواه برای سازمان فرهنگ و ارتباطات

Number of all token in the document: 230

Number of given query word: 60

Result 5

Document ID: 9261

Title: نشست اساتید با رئیس سازمان انرژی اتمی / تاکید بر نقش نخبگان در اجرای سند جامع انرژی هسته‌ای کشور

Number of all token in the document: 170

Number of given query word: 18

تحلیل: اسنادی که با هم تفاوت دارند:

Document ID: 8125

Title: پیاده‌روی دانشگاهیان دانشگاه امیرکبیر به مناسبت آغاز دهه فجر

Document ID: 9261

Title: نشست اساتید با رئیس سازمان انرژی اتمی / تاکید بر نقش نخبگان در اجرای سند جامع انرژی هسته‌ای کشور

دو سند بالا فقط در فاز دوم آمده اند و دو سند پایین فقط در فاز اول:

Document ID: 9286

Title: فتنه ۸۸؛ گناه نابخشودنی| روایت «آذر آمریکایی» اصلاح‌طلبان تندرو / از درخواست تحریم تا اهانت به امام (ره)

Document ID: 11603

Title: مسلمانان و دولت‌های اسلامی موظف به مقابله با ظلم و استکبار هستند

با مشاهده دو سند سبز که در حالت قبل فقط آمده اند پی می‌بریم که اگر صرفا بر اساس تعداد کلمات این بررسی را انجام دهیم چقدر اشتباه است چرا که ربط چندانی ندارد. سند مربوطه اول کاملا مرتبط و "مختص" به دانشگاه امیرکبیر می‌باشد که نشان می‌دهد tf-idf ملاک خیلی می‌باشد.

از طرفی سرعت کوئری ها هم با وجود index elimination بسیار پایین می‌آید.