به نام خدا

محمدمهدی هجرتی	**۹۷۲۳۱۰۰**
پروژه بازیابی اطلاعات	**فاز اول**
استاد نیک آبادی	**آذر ۱۴۰۰**

۱. **سوال اول**

در گام پیش پردازش ۴ عملیات زیر به ترتیب انجام داده شد.

۱) **نرمال سازی متون:** ابتدا تمام اطلاعات فایل اکسل ورودی خوانده شد و عملیات نرمال سازی روی آن انجام شد و در یک فایل اسکل دیگر با فرمت مشابه ذخیره شد. این عملیات با استفاده از تابع normalize از کتابخانه Hazm انجام شد و در آن تمام متون از منظر رعایت فاصله و نیم فاصله (مثل کلمه می‌رود) و حروف بزرگ و کوچک (مثل کلمه قرآن) و تنوین (مثل کلمه ظاهراً) و اطلاعاتی از این قبیل اصلاح شد تا در زمان سرچ و ایندکس کردن کلمات یکسان را مشابه را تکراری در نظر نگیریم و همچنین کلماتی که در آن فاصله وجود دارد به اشتباه دو کلمه ای در نظر گرفته نشود.

۲) **استخراج توکن:** در این مرحله با استفاده از تابع word_tokenize توکن های هر خبر استخراج و در لیست ذخیره شد. این مرحله برای انجام سرچ راحت تر در ادامه به کار می آید.

۳) **ریشه یابی اول:** در ابتدا با استفاده از تابع stem ریشه یابی کلمات انجام شد. (به طور مثال به جای کلمه ی کتاب ها، کتاب در لیست ذخیره شد.) این کار برای این انجام می شود که کلمات مشابه در زمان جست و جو یکسان در نظر گرفته شوند.

۴) **حذف کلمات پرتکرار:** برای بخش حذف کلمات پرتکرار از لیست stop word های فارسی در این پروژه ی گیت استفاده شد.

۵) **ریشه یابی دوم:** در نهایت با کمک تابع lemmatize ریشه ی افعال بدست آمده و در لیست جایگزین شد. (مثلا فعل می‌روم به ترم رفت#رو تبدیل شد.)

۲. **سوال دوم**

در قانون zipf تعداد تکرار کلمه ی i ام لیست متناسب است با k/i.

در ابتدا و قبل از حذف کلمات پرتکرار این قانون تا حد خوبی برقرار است. اما بعد از حذف با توجه به این که لغات پرتکرار مربوط به کل کلمات فارسی بودند و صرفا از ان مجموعه لغات آورده شده است، اندکی از دقت آن کاسته می شود.

۳. سوال سوم

در قانون heap، بین تعداد ترم های یونیک و کل سایز لغات رابطه وجود دارد.

در ۳ تصویر زیر به ترتیب تعداد کل ترم ها و ترم های یونیک در ۵۰۰ خبر اول، ۱۰۰۰ خبر اول و ۱۵۰۰ خبر اول آورده شده است.

```
all term:  152365
all unique term:  7838
Enter your query:
```

```
all term:  299345
all unique term:  10905
Enter your query:
```

```
all term:  450184
all unique term:  13081
Enter your query:
```

در تصویر زیر تعداد کل ترم ها در تمام اخبار و تعداد ترم های یونیک آورده شده است.

```
all term:  2890346
all unique term:  48310
Enter your query:
```

۴. سوال چهارم

به طور مثال کلمه تیم پس از ریشه یابی از دست رفته و فقط ت باقی می ماند. یا در کلمه ی علی، عل باقی مانده بود.

۵. سوال پنجم

الف) بین الملل

```
Enter your query: بین‌الملل
1 ) docID:  80 | title:  واکنش تند فدراسیون فوتبال به حواشی VAR و ماجرای قرارداد با شرکت اسرائیلی
2 ) docID:  153 | title:  توضیحات مسؤول مسابقات لیگ یک درباره شائعه سکته ناظر بازی
3 ) docID:  248 | title:  داور بین‌المللی هندی مسلمان شد +عکس
4 ) docID:  314 | title:  جوابیه باشگاه آلومینیوم به نسبت محرومیت 2 بازیکن ملپیپوش
5 ) docID:  460 | title:  اعلام زمان انتخابات کمیته ملی المپیک/ صالحی‌امیری: اختلافی با فدراسیون فوتبال ندارم/ دنبال انتقام از فدراسیونی نیستیم
6 ) docID:  556 | title:  دستگیری چند معترض در ونیز/مخالفت با المپیک پکن همچنان ادامه دارد
7 ) docID:  560 | title:  انتقاد عجیب هندبالی‌ها از پروفسور/فرناندز مرد سیاست یا مربی؟
8 ) docID:  617 | title:  آمادگی قوه قضائیه برای حمایت حقوقی و معنوی از ورزشکاران/ اژه‌ای: به دعاوی حقوقی ورزشی به صورت تخصصی رسیدگی خواهد شد
9 ) docID:  624 | title:  اسکی صحرایی کرسی جهانی شد
10 ) docID:  636 | title:  چرا کمیته بین‌المللی المپیک مخالف برگزاری جام جهانی فوتبال هر 2 سال یکبار است؟
Enter your query:
```

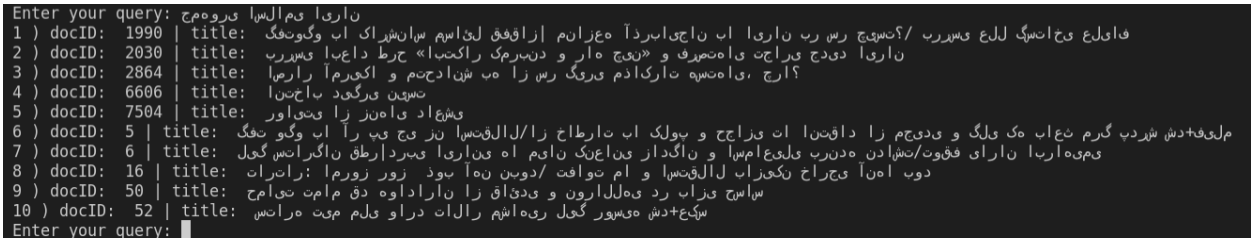
در اسناد بازنمایی شده کلمه ی بین الملل آورده شده است.

ب) دانشگاه امیرکبیر

```
Enter your query: دانشگاه امیرکبیر
1 ) docID:  7236 | title:  برگزاری دوبدای برای محمدسرور رجایی
2 ) docID:  1753 | title:  نامه جمعی از اساتید و متخصصان/ آقای رئیس‌جمهور در گام دوم انتقال به داد «مدیریت» در کشور برسید
3 ) docID:  1959 | title:  نامه ۸ بسیج دانشجویی دانشگاه‌های تهران به معاون اول رئیس‌جمهور
4 ) docID:  2131 | title:  بزرگداشت شهدای مسجد قندوز در مقابل کنسولگری افغانستان/ آمریکا و آل سعود مقصران اصلی جنایت در افغانستان
5 ) docID:  2709 | title:  پیشرفت علمی هوای ما قابل مقایسه با قبل از انقلاب نیست/ شب و روز صهیونیست‌ها در وحشت است
6 ) docID:  2792 | title:  امروز محیط دانشگاه‌های ما عرصه دفاع مقدس است
7 ) docID:  2793 | title:  باید برای ثبت نقش دانشگاهیان در دوران دفاع مقدس کار تحقیقاتی صورت گیرد
8 ) docID:  5021 | title:  دفترچه راهنمای آزمون استخدامی دانشگاه‌ها برای بار چهارم اصلاح شد/تمدید مجدد مهلت ثبت‌نام
9 ) docID:  5022 | title:  دفترچه راهنمای آزمون استخدامی دانشگاه‌ها برای بار چهارم اصلاح شد/تمدید مجدد مهلت ثبت‌نام
10 ) docID:  7520 | title:  اعطای گواهی درجه یک هنری به سعید بیابانکی
Enter your query:
```

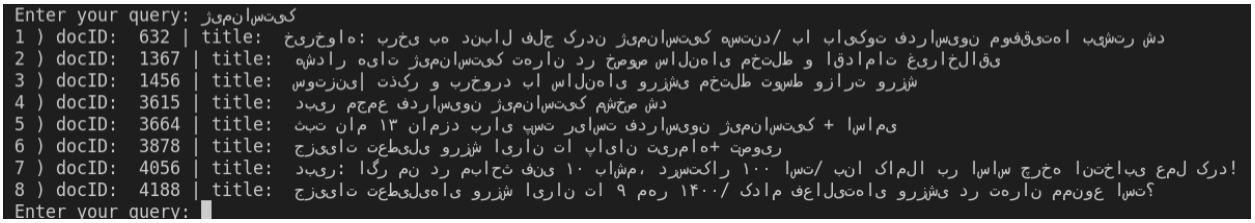
در اولین سند بازنمایی شده خبری در مورد مراسم یادبود می باشد که در دانشگاه امیرکبیر برگزار شده است. در سایر خبر ها صرفا کلمه ی دانشگاه آورده شده است.

پ) جمهوری اسلامی ایران



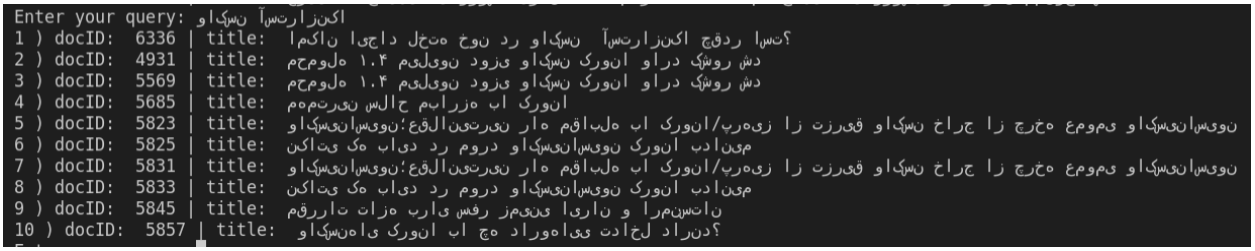
در ۵ سند اول دقیقا عبارت خواسته شده آمده است. با اینکه ممکن است متن خبر در مورد کشور های همسایه ایران باشد. در اسناد دیگر اخبار شامل هر یک از کلمات آورده شده است.

ت) ژیمناستیک



در تمام ۸ سند بازنمایی شده خبری مرتبط با ورزش ژیمناستیک آمده است.

ث) واکسن آسترازنکا



در اولین سند دقیقا کلمه ی واکسن آسترازنکا آورده شده است و در مورد لخته شدن خون در این واکسن توضیح می دهد. سایر اسناد نیز مرتبط با موضوع واکسن هستند.