

به نام خدا

محمد مهدی 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. (مثلاً 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 صرفاً 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: للملانیب
1 ) docID: 80 | title: لیئارسا نکرش اب دادرارق یارحام و VAR یسراوح هب لابتوف نویساردف دنت شنگاو
2 ) docID: 153 | title: یراب رطان هتکس هعاش هرابرد کی گیل تاوایسم لوؤسم تاوایسموت
3 ) docID: 248 | title: سگخ+ دش ناملسم یدنه یللملانیب رواد
4 ) docID: 314 | title: شوپیللم یگیزاب 2 تیمورحم هب تبسن موینیمولآ هاکشراپ هسباوح
5 ) docID: 460 | title: میسین ینویساردف اب یفالنخا: یریم ایجلالاس/کیپملا یلم هتیمک تاباجننا نامز مالغا
6 ) docID: 556 | title: دراد همادا نانچمه یگب کیپملا اب بفللام/نانوی رد مرتعم دنج یرگتسپد
7 ) docID: 560 | title: گیرم ای تنسا سیردم زدنارک/روسیفورپ را اهیلادننه بیج داقننا
8 ) docID: 617 | title: دش دهاوخ یگدیسر یسپخت تروص هب یسزرو یقوق یواعد هب: یاهزا/اناراکشزرو را یونغم و یقوق تمام یارب هئامق هوک یگدامآ
9 ) docID: 624 | title: دش یناهج یسک بحاس یکسا
10 ) docID: 636 | title: تنسا رایکی لاس 2 ره لابتوف یناهج ماح یرارگرب فلجام کیپملا یللملانیب هتیمک ارج
Enter your query: █
```

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

ب) دانشگاه امیرکبیر

```
Enter your query: هاکشناد ریگریما
1 ) docID: 7236 | title: ییاجر رورسپدمحم یارب یدوبدای یرارگرب
2 ) docID: 1753 | title: دیسرب روشک رد «تیریدم» داد هب بالقنا مود ماگ رد روهمسپئر یاقا/اناسپختم و دیتاسا را یغج همان
3 ) docID: 1959 | title: روهمس سپئر لوا نواعم هب نارهت یاههاکشناد ییوجشناد چیسب ۸ همان
4 ) docID: 2131 | title: ناتسناغفا رد تیانج یلما نارمقم دوعس لآ و اکیرما/ناتسناغفا یرگلوسیک لباوقم رد زودنق دجسم یادهش تشادگرزب
5 ) docID: 2709 | title: تنسا تیزجو رد اهتسینویس زور و بش/تسین بالقنا را لبق اب هسپاقم لباوقم ییاوه عیانیس تفریشپ
6 ) docID: 2792 | title: تنسا سیدقم عافد هصرع ام یاههاکشناد طقم زورما
7 ) docID: 2793 | title: درگی تروص یثاقیقحت راک سیدقم عافد نارود رد نایهاکشناد شقن تنب یارب دیاب
8 ) docID: 5021 | title: مانیتب تلهم ددجم دیدم/دش حالما مراجه راب یارب اههاکشناد یمادجتسا نومزآ یامنه ار هجرتفد
9 ) docID: 5022 | title: مانیتب تلهم ددجم دیدم/دش حالما مراجه راب یارب اههاکشناد یمادجتسا نومزآ یامنه ار هجرتفد
10 ) docID: 7520 | title: یکنابابی دیعس هب یرنه کی هجرد یهاوگ یاطغا
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

```
Enter your query: ناریا یمالسا یروهمج
1 ) docID: 1990 | title: فایله یخاتسگ لیلع یسررب /؟تسپج رس رب ناریا اب ناجیابردآ هعرانم ارافق لئاسم ساینشراک اب وگوتفگ
2 ) docID: 2030 | title: ناریا دودج یراجت یاهتصرف و «نیج هار و دنبرمک راکتبا» حرط داعبا یسررب
3 ) docID: 2864 | title: ؟ارج ،یاهتسبه تارکادم یریگ رس زا هب شنادحتم و اکیرم رارصا
4 ) docID: 6606 | title: تسپین یریگید باختنا
5 ) docID: 7504 | title: شوعاد یاهنر زا ییاور
6 ) docID: 5 | title: ملیف+دش شرپ گرم نخاب هک یلگ و یدیجم را داقتنا ات یراجح و پولک اب تارطاخ را/لالقتسا نر یج یب را اب وگو تفگ
7 ) docID: 6 | title: یمیهاربا ناریا ففوت/تشانن هدنرب یلیغامسا و ناگداز یناعک نایم اه یناریا یبردارطق ناگراتس گول
8 ) docID: 16 | title: دوب اهنا یجراج نکیزاب لالقتسا و ام توافت /دوبن نهآ بود زور زورما راترات
9 ) docID: 50 | title: ساسج یزاب رد یهللارون و یدئاق زا ناراداهه دق مامت تیامح
10 ) docID: 52 | title: سگ+دش هیسور گول ریهاشم رالات دراو یلم میت هراتس
Enter your query:
```

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

```
Enter your query: کیتسانمیژ
1 ) docID: 632 | title: دش رتشیب اهتقیقوم نویساردف توکیاب اب /دنتسبه کیتسانمیژ ندرک جلف لابند هب یخرب :هواخریخ
2 ) docID: 1367 | title: یقالخاریغ تامادقا و طلنختم یاهنلاس صومخ رد نارهت کیتسانمیژ تاهه رادشه
3 ) docID: 1456 | title: شزرو ترازو طلنختم شزرو یاهنلاس اب دروخرب و رگدنت اینزتوس
4 ) docID: 3615 | title: دش صخشم کیتسانمیژ نویساردف عمجم ریبد
5 ) docID: 3664 | title: یماسا + کیتسانمیژ نویساردف تساری تسپ یارب دزمان ۱۳ مان تبث
6 ) docID: 3878 | title: ریوصت +تامریت نایاب ات ناریا شزرو یلیطعت تاییزج
7 ) docID: 4056 | title: !درک لمع یباختنا هخرج ساسا رب الماک انب /تسا ۱۰۰ راکتسرد ،مشاب ۱۰ ینف نخابم رد نم رگا :ریبد
8 ) docID: 4188 | title: ؟تسا عونمم نارهت رد شزرو یاهتیلعاف مادک /۱۴۰۰ رهم ۹ ات ناریا شزرو یاهتلیطعت تاییزج
Enter your query:
```

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

ث) واکسن آسترازنکا

```
Enter your query: اکنزارتسا نسکاو
1 ) docID: 6336 | title: ؟تسا ردقچ اکنزارتسا نسکاو رد نوخ هتخل داچیا ناکما
2 ) docID: 4931 | title: دش روشک دراو انورک نسکاو یزود نویلیم ۱۰۴ هلومجم
3 ) docID: 5569 | title: دش روشک دراو انورک نسکاو یزود نویلیم ۱۰۴ هلومجم
4 ) docID: 5685 | title: انورک اب هرابم خالسا یرتمهم
5 ) docID: 5823 | title: نویسانیسکاو یمومع هخرج زا جراخ نسکاو قیزرت زا زیهرپ/انورک اب هلباقم هار نیرتینالقع:نویسانیسکاو
6 ) docID: 5825 | title: مینادب انورک نویسانیسکاو دروم رد دیاب هک یتاکن
7 ) docID: 5831 | title: نویسانیسکاو یمومع هخرج زا جراخ نسکاو قیزرت زا زیهرپ/انورک اب هلباقم هار نیرتینالقع:نویسانیسکاو
8 ) docID: 5833 | title: مینادب انورک نویسانیسکاو دروم رد دیاب هک یتاکن
9 ) docID: 5845 | title: ناتسنمرا و ناریا ینیمز رفس یارب هرات تاررقم
10 ) docID: 5857 | title: ؟دنراد لخادت ییاهوراد هج اب انورک یاهنسکاو
Enter your query:
```

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.