به نام خدا

گزارش فاز اول پروژه ی بازیابی اطلاعات

امیرپارسا سلمان خواه

۹۸۳۱۰۳۴

کار های انجام شده در مرحله ی پیش پردازش:

۱-در ابتدا با کمک یک regex آدرس لینک های موجود در خبر را حذف کردم و صرفا عنوان لینک را در خبر نگه داشتم. به عنوان مثال اگر چنین لینکی در خبر داشته باشیم:

[باشگاه پرسپولیس](https://farsnews.ir/special/persepolis)

صرفا عنوان باشگاه پرسپولیس به عنوان نماینده ی لینک در متن خبر باقی می ماند.
این کار به پردازش بهتر متن کمک می کند زیرا اگر آن لینک ها حذف نشوند، خودشان به عنوان یک توکن به index اضافه می شوند و نه تنها استفاده ای از آن ها نمی شود بلکه ممکن است بین دو کلمه ی مرتبط فاصله بیندازند و موجب خراب شدن index شوند.

۲-سپس تمامی punctuation ها را از متن خبر حذف کردم چون هنگام جستجو نیازی به آن ها نداریم و همچنین ممکن است علائمی که به یک کلمه چسبیده اند، کار stemming یا normalization مربوط به آن کلمه را دچار مشکل کنند. همچنین از حجم index نیز با این کار کاسته می شود.

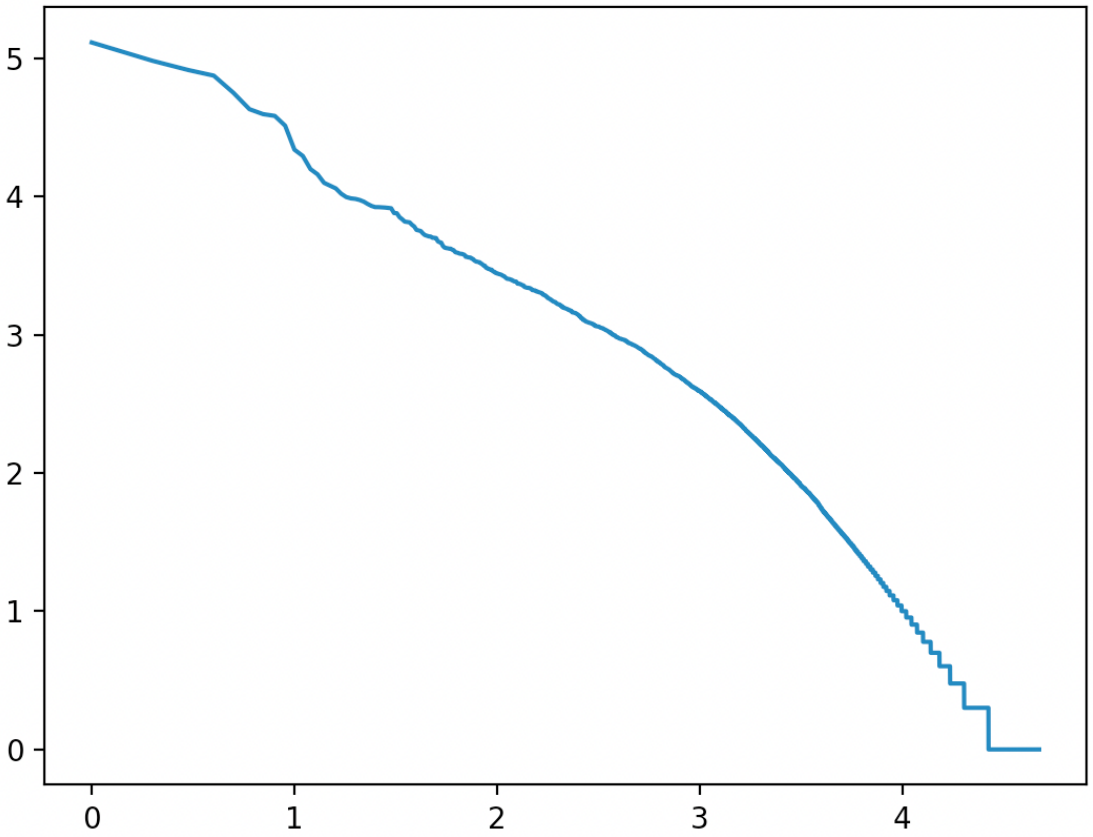
۳- در مرحله ی بعد با کمک توابع آماده ی کتابخانه های معرفی شده، عمل normalization را روی متن خبر انجام دادم. این کار باعث می شود تا کلماتی که ممکن است به چند شکل نوشته شوند یکی شوند. مثلا بودن یا نبودن نیم فاصله بین یک کلمه و نشانه ی جمع یکی از این مسائل است که با نرمال سازی متن، این فواصل به نیم فاصله تبدیل می شوند.

۴- در مرحله ی بعد در صورت تعیین استفاده از stop_list در ورودی تابع، فرایند حذف کلماتی که در stop_list هستند صورت می گیرد. این کار موجب کاهش حجم index شده و همچنین می تواند موجب افزایش دقت آن شود. به عنوان مثال کلماتی نظیر کلمات ربط یا ضمیر ها از index حذف می شوند.

۵- در مرحله ی آخر کار ریشه یابی صورت می گیرد. مثلا فعل ها به شکل ریشه ی خود در می آیند و کلماتی که جمع بسته شده اند به شکل مفرد خود در می آیند. این کار باعث می شود تا تغییرات جزئی در کلمات موجب پیدا نشدن آن ها نشود و اگر فعلی در query وجود دارد، شخص آن مهم نباشد یا جمع و مفرد بودن کلمات موجب تغییر جواب query نشود. البته این کار در برخی موارد ممکن است ما را دچار مشکل کند.

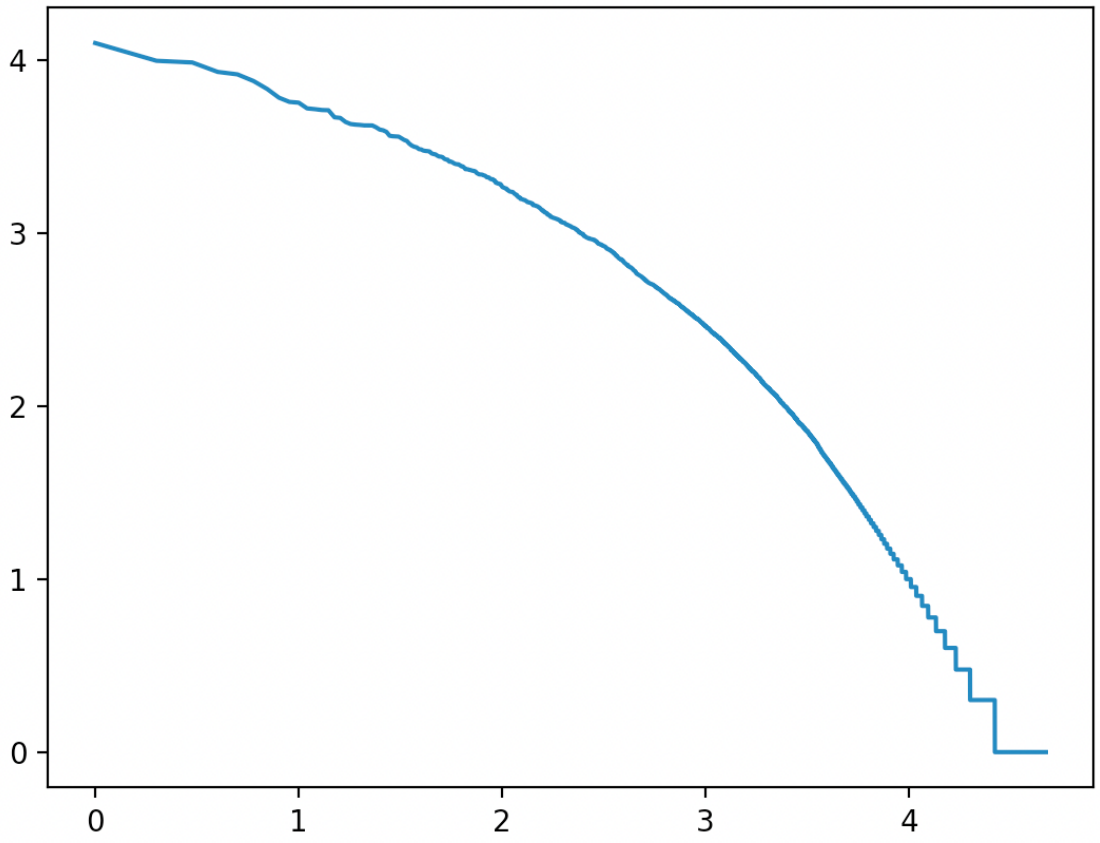
بررسی قانون zipf:

بعد از رسم نمودار لگاریتمی تعداد تکرار هر کلمه نسبت به رنکینگ آن بدون حذف stop word ها شکل زیر به دست آمد:



می بینیم که مقدار شیب نمودار کمی بیش از منفی یک است اما با این حال می توان گفت که قانون zipf با تقریب خوبی برقرار است.

اما با رسم نمودار در حالتی که stop word ها حذف می شوند داریم:



می بینیم که در این حالت شیب نمودار خیلی به منفی یک نزدیک شده و قانون zipf کاملا برقرار است.

بررسی قانون heaps:

در حالت با stemming ابتدا تعداد توکن ها و مجموع طول سند ها را برای مقادیر داده شده به دست آوردم:

۵۰۰ سند: ۷۶۱۲ توکن و طول ۱۵۳۸۳۳
۱۰۰۰ سند: ۱۰۴۶۷ توکن و طول ۳۰۱۹۸۴
۱۵۰۰ سند: ۱۲۵۷۰ توکن و طول ۴۵۳۶۳۳
۲۰۰۰ سند: ۱۷۶۶۰ توکن و طول ۶۷۰۷۵۶

با قانون heaps داریم:

$$V = kn^b \rightarrow \log V = b.\log n + \log k$$

حال مقادیر اول و دوم را جایگذاری می کنیم:

$$\log ۷۶۱۲ = b.\log ۱۵۳۸۳۳ + \log k$$
$$\log ۱۰۴۶۷ = b.\log ۳۰۱۹۸۴ + \log k$$

پس داریم:

$$b = (\log 10467 - \log 7612)/(\log 301984 - \log 153833)$$
پس مقدار آن تقریبا برابر است با ۰.۴۷۲
با جایگذاری در معادله ی اول مقدار k برابر است با:
$$\log k = \log 7612 - 0.47 \log 153833$$
پس مقدار k برابر است با: ۲۷.۱
با جایگذاری مقادیر در دیگر معادله ها متوجه شدم که هر چه جلوتر می رویم خطای آن بیشتر می شود. تا آن جا که در کل اسناد که ۴۷۵۷۳ توکن داریم مقدار ۷۴۱۳۱۰۲ به دست می آید که مقدار خیلی زیادی با ۲۹۱۸۶۵۵ فاصله دارد.

در حالت بدون stemming:
۵۰۰ سند: ۹۷۰۴ توکن و طول ۱۵۳۸۳۳
۱۰۰۰ سند: ۱۳۶۳۳ توکن و طول ۳۰۱۹۸۴
۱۵۰۰ سند: ۱۶۵۵۰ توکن و طول۴۵۳۶۳۳
۲۰۰۰ سند: ۲۳۷۳۳ توکن و طول ۶۷۰۷۵۶
حال مقادیر اول و دوم را جایگذاری می کنیم:
$$\log ۹۷۰۴ = b.\log ۱۵۳۸۳۳ + \log k$$
$$\log ۱۳۶۳۳ = b.\log ۳۰۱۹۸۴ + \log k$$
پس داریم:
$$b = (\log 13633 - \log 9704)/(\log 301984 - \log 153833)$$
پس مقدار آن تقریبا برابر است با ۰.۵۰۴
با جایگذاری در معادله ی اول مقدار k برابر است با:
$$\log k = \log 9704 - 0.504 \log 153833$$
پس مقدار k برابر است با: ۲۳.۵۸

در این حالت هم مقدار ۶۶۲۹۶۵۵ را برای تعداد کل کلمات به دست آوردم که باز هم اختلاف زیادی با مقدار اصلی داشت.

چالش های ریشه یابی:
۱-ان در آخر برخی کلمات مانند شبان(به معنی چوپان) ممکن است توسط ریشه یاب به غلط به عنوان نشانه ی جمع برداشت شوند و از آن حذف شوند و در این صورت معنای کلمه به طور کامل تغییر می کند.
۲- تر در آخر برخی کلمات مانند کفتر ممکن است به عنوان وابسته ی پسین در نظر گرفته شود در حالی که جزوی از کلمه هستند و در صورت حذف معنای آن تغییر می کند.
۳- یک سری کلمات مانند مات ممکن است توسط ریشه یاب به درستی ریشه یابی نشوند. دلیل آن را نمی دانم اما ریشه یاب ماژول hazm این کلمه را به یک استرینگ خالی تبدیل می کند.

پاسخ به پرسمان ها:

الف) بین‌الملل

لیست doc_id های برگردانده شده(۱۰ تای اول) و عناوین آن ها به شرح زیر است.

۱۵۴ :توضیحات مسؤول مسابقات لیگ یک درباره شایعه سکته ناظر بازی

۷۳۴ :گزارش تمرین پرسپولیس| روحیه شاد قبل از مصاف با الهلال/ پا به توپ شدن گل‌محمدی و باقری

۱۰۰۹ :معاون بین‌الملل و مدیر کمیته حرفه ای سازی باشگاه استقلال منصوب شدند

۱۰۷۴ :باشگاه پرسپولیس خواستار ارسال تاییدیه AFC برای هتل محل اقامت

۱۵۷۸ :نامه پرسپولیس به کنفدراسیون فوتبال آسیا برای اسکان شاگردان گل محمدی در ریاض

۱۷۳۰ :واکنش عضو کمیسیون اقتصادی مجلس به گرانی لوازم خانگی داخلی/ تولیدکنندگان بدانند کسی به آنها چک سفید امضاء نداده است

۱۷۴۴ :جهاد مسئولان، قوام‌بخش وحدت مردم خواهد بود

۱۷۷۳ :بیرون از این خانه هیچ خبری نیست

۱۷۸۸ :نخست وزیر پیشین عراق: آمریکا باید از عراق خارج شود و این خواسته همه ملت ماست

۱۸۳۱ :کرمی: به بعضی از همسایگان ایران تذکر می دهم که حرمت بزرگ‌تر از خود را نگاه دارند

۱۸۴۸: گزارش ناصحیح وزارت نفت و سازمان محیط زیست از بنزین تولیدی پتروشیمی‌ها/ زنگنه و ابتکار عامل تشویش اذهان عمومی

به عنوان مثال در سند اول جمله ی زیر وجود دارد:
با نفوذ و تجربه‌ای که در حوزه بین‌الملل داشت.

ب) دانشگاه امیرکبیر

۱۰ سند اول به شرح زیر است:

۱۹۶۰ :نامه ۸ بسیج دانشجویی دانشگاه‌های تهران به معاون اول رئیس جمهور

۲۱۳۲ :بزرگداشت شهدای مسجد قندوز در مقابل کنسولگری افغانستان/ آمریکا و آل سعود مقصران اصلی جنایت در افغانستان

۲۷۹۳ :امروز محیط دانشگاه‌های ما عرصه دفاع مقدس است

۲۷۹۴ :باید برای ثبت نقش دانشگاهیان در دوران دفاع مقدس کار تحقیقاتی صورت گیرد

۷۲۳۷ :برگزاری یادبودی برای محمدسرور رجایی

۳۶۳ :دایی: می‌خواهم مردم مرا به عنوان انسان به یاد بیاورند نه دایی

۳۹۱ :تجلیل دانشگاه آزاد از قهرمانان المپیک و کشتی‌گیران مدال‌آور+ تصاویر

۳۹۴ :یزدانی: فکر می‌کردم امتیاز فینال جهانی را به من می‌دهند/سبک حریف روسی تغییر کرده بود

۵۵۳ :میزبانان لیگ کشتی معرفی شدند/ زمان شروع مسابقات مشخص شد

۵۶۱: انتقاد عجیب هندبالی‌ها از پروفسور/فرناندز مدرس است یا مربی؟

در ۵ سند اول کلمه ی دانشگاه امیرکبیر به شکل دقیق آمده است و در سند های بعدی یکی از کلمات دانشگاه یا امیرکبیر وجود دارد.

پ)دانشگاه صنعتی امیرکبیر

۱۰ سند اول به شرح زیر است:
۵۰۲۲ :دفترچه راهنمای آزمون استخدامی دانشگاه‌ها برای بار چهارم اصلاح شد/تمدید مجدد مهلت ثبت‌نام
۵۰۲۳ :دفترچه راهنمای آزمون استخدامی دانشگاه‌ها برای بار چهارم اصلاح شد/تمدید مجدد مهلت ثبت‌نام
۳۶۳ :دایی: می‌خواهم مردم مرا به عنوان انسان به یاد بیاورند نه دایی
۱۷۲۶ :سیدمحسن دهنوی عضو هیئت امنای صندوق نوآوری و شکوفایی شد
۱۷۵۴ :نامه جمعی از اساتید و متخصصان/ آقای رئیس‌جمهور در گام دوم انقلاب به داد «مدیریت» در کشور برسید
۱۷۸۰ :وزیر علوم: علم و عقل دو بال دانایی است/ علم باید برای جامعه ثروت‌آفرین باشد
۱۹۶۰ :نامه ۸ بسیج دانشجویی دانشگاه‌های تهران به معاون اول رئیس جمهور
۲۰۹۲ :حجت‌الاسلام رستمی فقدان فعال دانشجویی دانشگاه شریف را تسلیت گفت
۲۱۳۲ :بزرگداشت شهدای مسجد قندوز در مقابل کنسولگری افغانستان/ آمریکا و آل سعود مقصران اصلی جنایت در افغانستان
۲۴۴۸: سیدرضا مرتضوی و مهدی دوستی استانداران اصفهان و هرمزگان شدند

در ۲ خبر اول عبارت دانشگاه صنعتی امیرکبیر به شکل دقیق آمده است اما از خبر سوم به بعد حداقل یکی از کلمه ها وجود ندارد. مثلا در خبر سوم عبارت دانشگاه صنعتی شریف آمده است.

ت)ژیمناستیک
تنها ۷ خبر بازگردانده شد:
۶۳۳ :خیرخواه: برخی به دنبال فلج کردن ژیمناستیک هستند/ با بایکوت فدراسیون موفقیت‌ها بیشتر شد
۱۳۶۸ :هشدار هیات ژیمناستیک تهران در خصوص سالن‌های مختلط و اقدامات غیراخلاقی
۳۶۱۶ :دبیر مجمع فدراسیون ژیمناستیک مشخص شد
۳۶۶۵ :ثبت نام ۱۳ نامزد برای پست ریاست فدراسیون ژیمناستیک + اسامی
۳۸۷۹ :جزییات تعطیلی ورزش ایران تا پایان تیرماه+ تصویر
۴۰۵۷ :دبیر: اگر من در مباحث فنی ۱۰ باشم، درستکار ۱۰۰ است/ بنا کاملا بر اساس چرخه انتخابی عمل کرد!
۴۱۸۹: جزییات تعطیلی‌های ورزش ایران تا ۹ مهر ۱۴۰۰/ کدام فعالیت‌های ورزشی در تهران ممنوع است؟
در همه ی اخبار کلمه ی ژیمناستیک وجود دارد.

ث)واکسن آسترازنکا
تعداد زیادی خبر بازگردانده شد که ۱۰ تای اول به شکل زیر است:
۴۹۳۲ :محموله ۱.۴ میلیون دوزی واکسن کرونا وارد کشور شد
۵۵۷۰ :محموله ۱.۴ میلیون دوزی واکسن کرونا وارد کشور شد
۵۶۸۶ :مهم‌ترین سلاح مبارزه با کرونا
۵۸۲۶ :نکاتی که باید در مورد واکسیناسیون کرونا بدانیم
۵۸۳۴ :نکاتی که باید در مورد واکسیناسیون کرونا بدانیم

۵۸۴۶ :مقررات تازه برای سفر زمینی ایران و ارمنستان
۵۸۵۸ :واکسن‌های کرونا با چه داروهایی تداخل دارند؟
۶۳۳۷ :امکان ایجاد لخته خون در واکسن آسترازنکا چقدر است؟
۳۲۸ :تارتار: ۳ امتیاز باارزش در آبادان بدست آوردیم/نفت این فصل از سال گذشته بهتر است
۳۷۵: مهدی:استقلال درخواستی برای تعویق بازی هایش نداشته است/ پدیده تا ۳ ساعت قبل از بازی لیستش را ثبت نکرده بود

۸ خبر اول عبارت واکسن آسترازنکا را به شکل دقیق در خود دارند و ۲ خبر آخر تنها شامل کلمه ی واکسن هستند.