

# گزارش پروژه

9731094

شایان صورتگر

1- در این پروژه از کتابخانه hazm استفاده شده است.

بعد از خوانده شدن ستون content ، در یک حلقه ی for هر سند به تابع نورمالایزر داده می شود .

```
12 # Normalizing
13 normalizer = Normalizer()
14 for doc in docs:
15     doc = normalizer.normalize(doc)
```

این عمل برای یکسان سازی کلمات انجام می شود مثلا تمام قرآن ها که به شکل قرآن هم نوشته می شود به یک شکل واحد در می آیند. همچنین مسائلی مثل نیم فاصله و انواع ی و ک اصلاح می شود.

قرآن می خواند = قرآن می خواند

سپس هر سند با استفاده از تابع word tokenize ، توکنایز شده و در یک حلقه ، کلمات به ترتیب با تابع stemmer ریشه یابی شده و هر کلمه با doc id و پوزیشن آن ، در token stream قرار می گیرد.

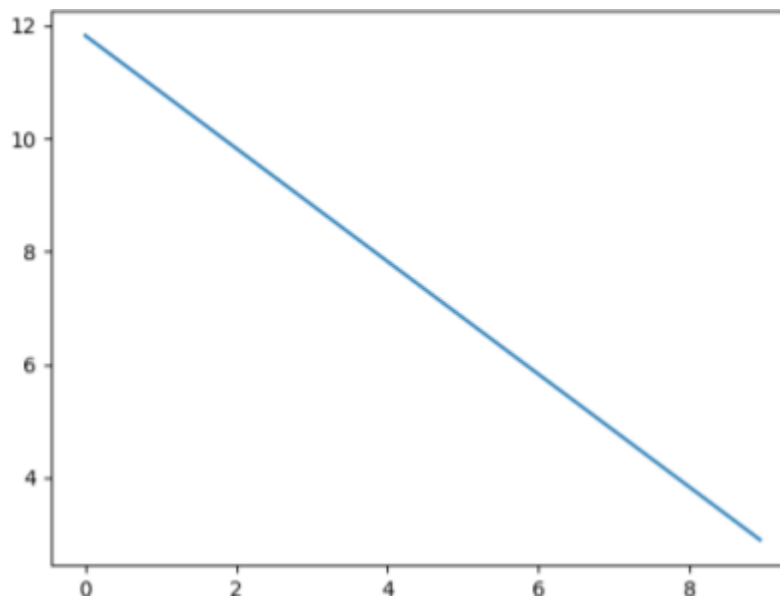
```
20
21 for docid in range(len(docs)):
22     terms = word_tokenize(docs[docid])
23     for pos in range(len(terms)):
24         term = stemmer.stem(terms[pos])
25         term = terms[pos]
26         token_stream.append([term, [docid, pos]])
27 token_stream.sort(key=lambda x: x[0])
28
```

کتاب ها را آورد =

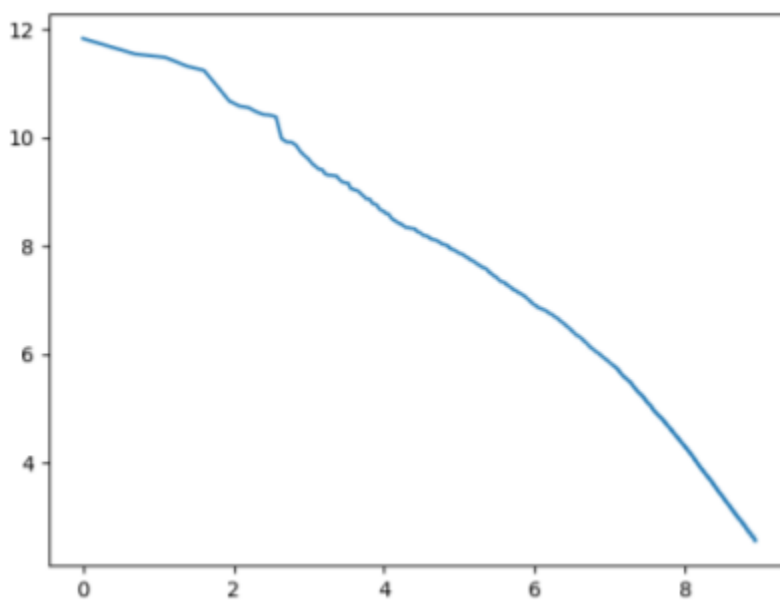
```
[1, 1], کتاب]
[1, 2], را]
[1, 3], آورد]
```

در بخش بعدی هم بعد از محاسبه فرکانس کلمات ، 15 توکن پرتکرار (10 کلمه) حذف شد.دلیل این امر کم اهمیت بودن این کلمات در پرسمان هستند مصل از در به برای و ...

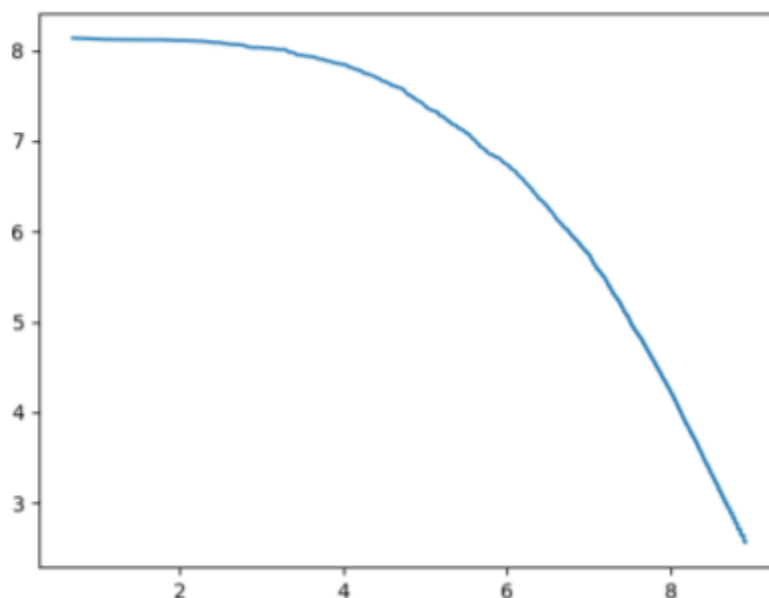
2-طبق قانون زیپف نمودار باید مشابه نمودار زیر باشد:



قبل از حذف stop words



بعد از حذف



می بینیم که در دو حالت قانون برقرار است. برای رسم نمودار ها باید آن ها را لگاریتمی رسم کنیم. تعداد کلمات با رتبه آنها رابطه توانی دارند. کلماتی که رتبه بالا دارند تعداد دفعات تکرار خیلی زیادی دارند. هر چقدر که جلو تر میرویم تعداد کلمات به شکل نمایی کاهش پیدا میکنند تا جایی که بعضی کلمات به ندرت استفاده میشوند. از آنجایی که این رابطه نمایی است پس با گرفتن لگاریتم به یک رابطه خطی تبدیل میشود و نمودارهای بالا به دست می آید.

تعداد قبل از ریشه یابی (اعداد تقریبی است):

	word	token
500	150000	8000
1000	300000	11000
1500	450000	13000
2000	670000	20000

$$\text{Log } m = 1.73$$

$$\text{Log } t = 1.56$$

$$\text{Log } 52500 * 1.73 - 1.56 = 4200000$$

بعد از ریشه یابی:

	word	token
500	150000	7500
1000	300000	10500
1500	450000	12500
2000	670000	17500

$$\text{Log } T - 1.55 \quad \text{Log}(M) = 1.74$$

$$\text{Log}(M) = 1.74 * \text{Log}(47000) - 1.55 = 6.620000 \Rightarrow M = 4,191,000$$

در هیچ کدام از حالات اعداد به دست آمده به مقادیر واقعی نزدیک نبودند. یک دلیل آن می تواند صدق نکردن این قانون در فارسی باشد زیرا بسیاری از کلمات از ترکیب ساخته می شوند که در شمارش با زبان انگلیسی فرق می کند و بنابراین در تعداد کل اختلال ایجاد می کنند.

همچنین سائز مجموعه ی ما به نسبت مجموعه هایی دیگر کم است که باعث کاهش دقت می شود.

-4

مشکل در ریشه یابی فامیلی ها که نباید ریشه یابی می شدند: مسلمی => مسلم

بسیاری از کلمات به خودی خود معنا داشتند و با اصلشان تفاوت داشتند: نارنجی => نارنج

صفات نسبی نباید ریشه یابی می شدند : کاغذی => کاغذ

کلمه ی برترین به بر تبدیل شد که اصلا نباید اتفاق می افتاد

-5

گفت وگو با آر پی جی زن استقلال/از خاطرات با کلوپ و حجازی تا انتقاد از مجیدی و گلی که باعث مرگ پدرش شد+فیلم تارتار: امروز روز ذوب آهن نبود/ تفاوت ما و استقلال بازیکن خارجی آنها بود رودنیل: مدافع استقلال مرتکب خطای پنالتی نشد/ داور قضاوت خوبی انجام داد حاشیه بازی ذوب آهن و استقلال|خروج مربی سیاهان از ورزشگاه با درخواست ناظر/بازگشت متفرقه ها به جایگاه خبرنگاران حاشیه بازی ذوب آهن و استقلال|اعتراض ناظر مسابقه به حراست/تشویق ایسلندی در فولادشهر حسینی: لیگ برتر مانند ماراتن است/می توانستیم گل های بیشتری به هوادار بزنیم حاشیه بازی ذوب آهن و استقلال|صورت خونی هافبک آبی ها /بازیکن ذوب آهن با آمبولانس از ورزشگاه خارج شد گلایه های عجیب سرمربی فجر از شرایطش در این تیم؛ سخته کردم/ از من خوششان نمی آید بگویید هفته دوم لیگ برتر! گل گهر و پیکان پیروز شدند/ دومین باخت فجرسپاسی و هوادار

پرسمان داور یک کلمه ای و متداول است و چون اطلاعات زیادی به ما نمی دهد عملا معلوم نیست که نتایج چقدر خوب بوده اند.

واکنش بازیکن نفت مسجدسلیمان به صحبت های پیروانی: با شرافت بازی کردیم/ این حرف ها زشت است کریمی: ما بیشتر از استقلال و پرسپولیس به خاطر نبود هوادار هورر کردیم/تمام فوتبال به جان یک طرفدار نمی ارزد کادر فنی نفت مقرر شکست مقابل تراکتور/خرمگاه: انتظار چنین بازی خوبی از شاگردانم را نداشتم، اما اسیر اشتباهات شدیم تارتار: امروز روز ذوب آهن نبود/ تفاوت ما و استقلال بازیکن خارجی آنها بود مجیدی: بازیکنانم باید در بازی بعد با کیفیت تر بازی کنند/یامگا را جذب کردیم چون در شرایط مسابقه بود استقلال طلسم 5 ساله ذوب آهن و 6 ساله در لیگ برتر را شکست گلایه های عجیب سرمربی فجر از شرایطش در این تیم؛ سخته کردم/ از من خوششان نمی آید بگویید بازیکنان پدیده باز هم تعویض نکردند/مهاجری استعفا کرد؛ میثاقیان در آستانه بازگشت به لیگ برتر واکنش مهدوی کیا به پیروزی پرگل ایران مقابل نبال 10 باشگاه ارزشمند فوتبال جهان را بشناسید/سیتیزن ها بالاتر از پاریسی ها در صدر\*عکس

پرسمان داور دو کلمه ای و متداول است و مانند بخش قبل نمی توان ارزیابی درستی از میزان رضایت کاربر داشت.

#### ایوان مسجد سلیمان

کمالوند نیامده قهر کرد/نایبامانی تیم لیگ برتری در غفلت مدیران نفتی کشور واکنش مدیر سپاهان به اعتراض پیروانی: فراموش کردند در جام حذفی حق ماهشهر را به آبادان منتقل کردند دعوت خطیبی و نویدکیا به کمیته انضباطی هفته دوم لیگ برتر| سپاهان مدعی در شهر اولین‌ها، جدال مس رفسنجان با پدیده بحران زده بازی با زینت شاید وقتی دیگر/ سپاهان به دنبال همکاری با باشگاه طارمی امیری: دکتر پیکان دستم را جا انداخت/وجود عقرب در خوابگاه طبیعی است عجیب و باورنکردنی؛ موش و عقرب در خوابگاه تیم لیگ برتری فوتبال+ فیلم پیروزی ذخیره‌های پیکان در دیدار تدارکاتی دعوت از 25 بازیکن به اردوی تیم فوتبال امید با حضور 4 بازیکن استقلال بازیکن تیم نفت مسجد سلیمان از بیمارستان مرخص شد

با وجود اینکه تیم نفت مسجد سلیمان متداول است ولی نسبتاً دقت خوبی دارد و حداقل می‌دانیم تمام خبرها ارتباط خوبی با پرسیمان دارند.

#### ژیمناستیک

خیرخواه: برخی به دنبال فلج کردن ژیمناستیک هستند/ با بایکوت فدراسیون موفقیت‌ها بیشتر شد هشدار هیات ژیمناستیک تهران در خصوص سالن‌های مختلط و اقدامات غیراخلاقی دبیر مجمع فدراسیون ژیمناستیک مشخص شد ثبت نام ۱۳ نامزد برای پست ریاست فدراسیون ژیمناستیک + اسامی جزئیات تعطیلی ورزش ایران تا پایان تیرماه+ تصویر دبیر: اگر من در مباحث فنی ۱۰ باشم، درستکار ۱۰۰ است/ بنا کاملاً بر اساس چرخه انتخابی عمل کرد! جزئیات تعطیلی‌های ورزش ایران تا ۹ مهر ۱۴۰۰/ کدام فعالیت‌های ورزشی در تهران ممنوع است؟

چون کلمه ی ژیمناستیک دشوار بوده تمام نتایج برگردانده شده و می‌توان گفت نهایت تلاشمان را کرده ایم.

#### اتحاد جماهیر شوروی

گفتوگو با کارشناس مسائل قفقاز| منازعه آذربایجان با ایران بر سر چیست؟/ بررسی علل گستاخی علی‌اف نماینده ارمنه در مجلس: آذربایجان به‌زودی هزینه سنگین اقداماتش در قبال ایران را می‌پردازد بیانیه دانشجویان شمال‌غرب کشور در محکومیت اظهارات مقامات جمهوری آذربایجان سردار شکارچی: امروز ارتش آمریکا و رژیم جعلی اسرائیل از قدرت ایران می‌لرزند بیانیه سپاه پاسداران به مناسبت هفته دفاع مقدس/تحقق اخراج آمریکا از منطقه حتمی است بررسی عکسی از یک ترور علیرضاییگی: دولت آذربایجان بداند امنیت با تکیه بر بیگانگان شکننده است رویای والیبالیست‌های جوان محقق نشد/ موتور که دیر روشن شد؛ بدشمنی و دیگروهیج! آیت‌الله رئیسی: شهید سلیمانی در جهان اسلام اقدامات عملی کرد/ امت اسلامی باید با همدیگر متحد باشند بازیگوشی نظامی!

اتحاد جماهیر شوروی یک پرسیمان دشوار است و می‌توان گفت که تمام اخبار مربوطه برگردانده شده است.ولی با توجه اینکه کلمه ی اصلی در این پرسیمان شوروی است اگر بر روی این کلمه تاکید بیشتری می‌شد نتایج نامربوط کمتر بود