به نام خدا

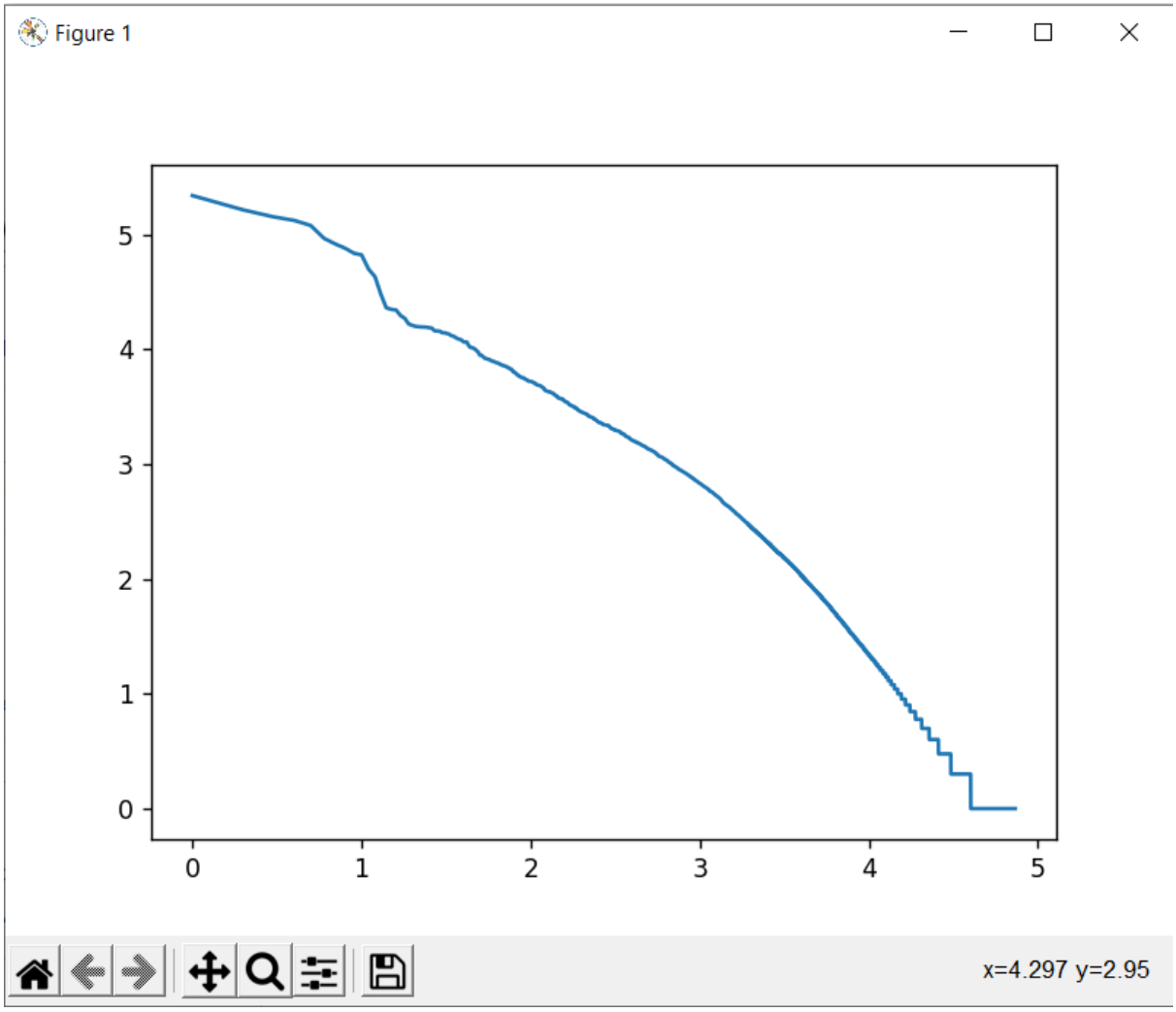
گزارش فاز ۱ پروژه بازیابی اطلاعات

امیر خسروی نژاد

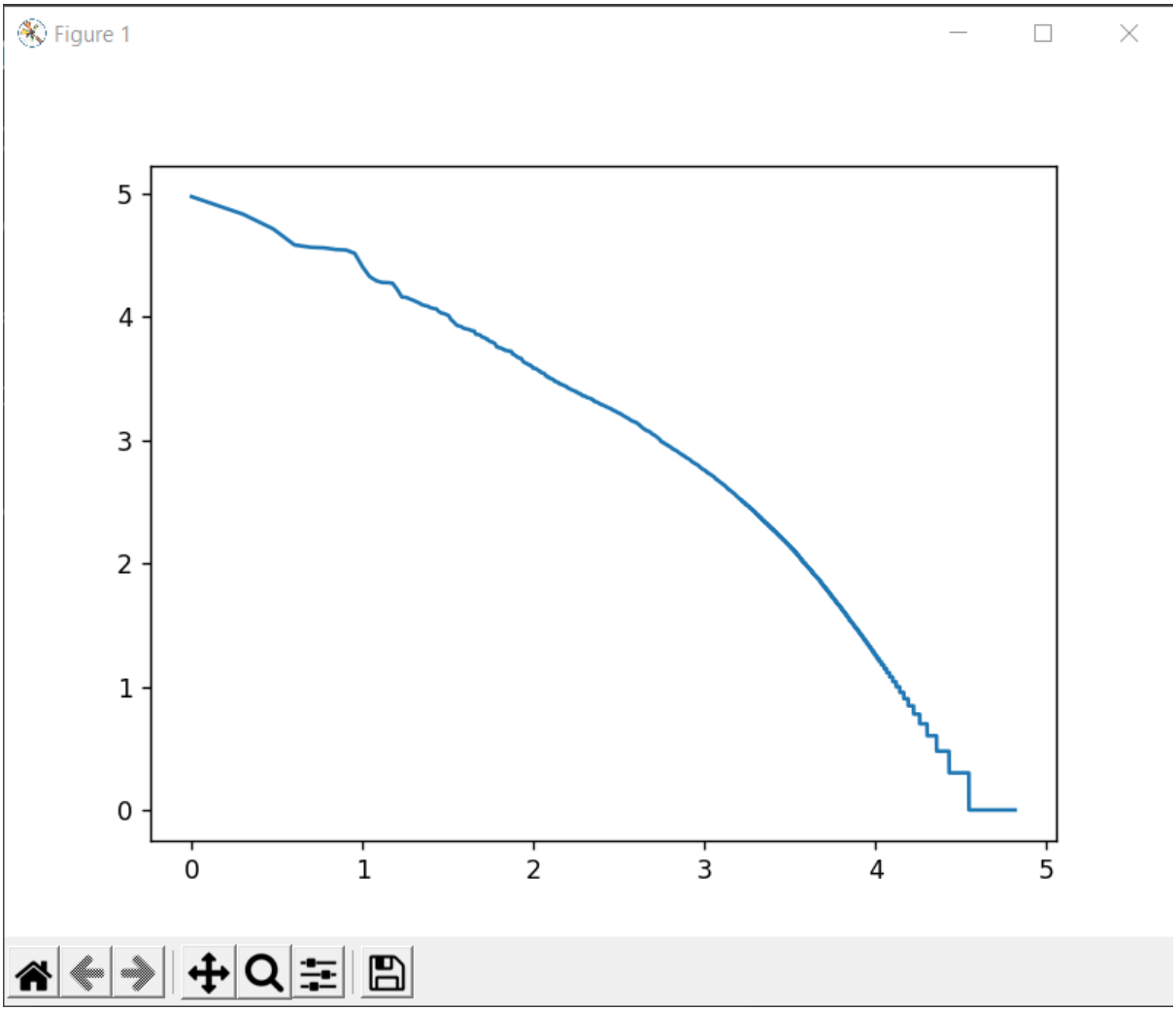
۹۸۳۱۱۱۳

۱. در پیش پردازش روشهای حذف stop word ها را انجام دادم که کلمات پرتکرار را که پیچیدگی محاسباتی را زیاد میکند حذف میکنیم. همچنین علایم نگارشی و علامت سوال، تعجب، نقل قول «» و مشابه آن که پرتکرار هستند هم از لیست واژه‌ها حذف کردیم. اول کار هم عملیات normalization انجام داده‌ایم که شکلهای مختلف کلمات که متفاوت از هم هستند، به شکلی واحد تبدیل شوند.(مثل یونیکدهای مختلف برای حرف ی که همه را به یک یونیکد واحد، تبدیل کردیم.) واضح است که این عملیات هم باعث کاهش حجم index میشود. در آخر هم عملیات stemming انجام شد که مثلا واژه‌های می‌شود و شدند و .. همه به شد که ریشه این کلمات هستند تبدیل میشود. این مورد هم به طور چشمگیری باعث کاهش حجم index میشود.

۲. قبل از حذف stopword نمودار zipf law به صورت زیر بود:



بعد از حذف stopword نمودار zipf law به صورت زیر شد:



میتوان گفت بعد از حذف stop words تا حد خوبی شیب نمودار به منفی ۱ نزدیک شده و قانون به خوبی برقرار است.

۳. قبل از ریشه یابی طبق رابطه $V=kn^b$ برای ۵۰۰ ۱۰۰۰ ۱۵۰۰ و۲۰۰۰ به صورت زیر است:

۵۰۰ سند: v = 10356, n = 102400

۱۰۰۰ سند: v =۱۵۱۵۶

n = ۲۹۲۱۳۱

۱۵۰۰ سند: v =۱۸۲۰۶

n= ۴۴۲۶۵۳

۲۰۰۰ سند: v=۲۰۶۴۰

n=۵۸۱۸۰۵

با جایگذاری ۲ مقدار از v , n های بالا در قانون heaps میتوان k و b را بدست آورد:

B=0.6974, k =0.52

$V = 0.52n^{0.6974}$

بعد از ریشه یابی طبق رابطه $V=kn^b$ برای ۵۰۰ ۱۰۰۰ ۱۵۰۰ و۲۰۰۰ به صورت زیر است:

۵۰۰ سند: v = 9301, n = 102400

۱۰۰۰ سند: v =۱۳۵۳۰

n = ۲۹۲۱۳۱

۱۵۰۰ سند: v =۱۶۲۴۸

n= ۴۴۲۶۵۳

۲۰۰۰ سند: v=۱۸۴۰۱

n=۵۸۱۸۰۵

با جایگذاری ۲ مقدار از v , n های بالا در قانون heaps میتوان k و b را بدست آورد:

B=0.3613, k =143

$V = 0.3613n^{143}$

۴. ۱. بعضی کلمات مثل نشان ممکن است بخاطر ان اخر کلمه به عنوان کلمه جمع در نظر گرفته شود یا حتی به عنوان صفت فاعلی (مثل خندان که خند + ان است) در نظر گرفته شود که باعث رویداد اشتباه میشود.

۲. موارد دیگری هم موجود است که نیم فاصله کلا نرمالسازی و البته ریشه یابی را هم دچار مشکل میکند.

۳. یک مورد دیگری که وجود داشت ترکیب ریشه یابی و حذف کلمات پرتکرار (stop words) بود. مثلا کلمه میشود هم جزأ کلمات پرتکرار است و هم ریشه آن را باید می‌گرفتم. برای بعضی کلمات مثل این مورد، باید اول stemming انجام میدادم و بعد حذف میکردم و گرنه به ارور برخورد میکردم.

۵. بخش ۱: تحریمهای آمریکا علیه ایران

rank: 1

اصولی: فدراسیون فوتبال جمهوری اسلامی ایران هستیم نه جزیره مستقل/ با گفتار ساختارشکنانه فدراسیون را به ناکجا آباد می‌برند

https://www.farsnews.ir/news/14001117000518/اصولی-فدراسیون-فوتبال-جمهوری-اسلامی-ایران-هستیم-نه-جزیره-مستقل-با

rank: 1

۲۰ حقیقت شنیده نشده درباره آزمون؛ از راز و نیاز با خدا پیش از بازی تا صحبت به ۵ زبان +عکس

https://www.farsnews.ir/news/14001009000246/20-حقیقت-شنیده-نشده-درباره-آزمون-از-راز-و-نیاز-با-خدا-پیش-از-بازی-تا

rank: 1

توضیحات یک منبع آگاه درباره وقفه مذاکرات وین

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

rank: 1

بیانیه نمایندگان مجلس/ رفع کاغذی تحریم‌ها تأمین‌کننده حقوق ملت ایران نخواهد بود

https://www.farsnews.ir/news/14001222000366/بیانیه-نمایندگان-مجلس-رفع-کاغذی-تحریم‌ها-تأمین‌کننده-حقوق-ملت-ایران

rank: 1

خطیب‌زاده: تمام تلاش خود را برای اتمام جنگ در اوکراین خواهیم کرد

https://www.farsnews.ir/news/14001221001176/خطیب‌زاده-تمام-تلاش-خود-را-برای-اتمام-جنگ-در-اوکراین-خواهیم-کرد

بخش ۲: تحریمهای آمریکا ! ایران

امام جمعه موقت تهران: مذاکره کنونی با گذشته که نقد دادند و نسیه گرفتند، تفاوت دارد

https://www.farsnews.ir/news/14000919000137/امام-جمعه-موقت-تهران-مذاکره-کنونی-با-گذشته-که-نقد-دادند-و-نسیه-گرفتند

بخش ۴: "کنگره ضد تروریست"

این مورد فقط شامل خبر شماره ۶۹۲۹ بود.

۶۹۲۹

توضیحات یک منبع آگاه درباره وقفه مذاکرات وین

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

در ۲ مورد آخر هم که "تحریم هسته‌ای" آمریکا ! ایران و اورشلیم ! صهیونیست موردی یافت نشد