باسمه تعالی

دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

درس بازیابی اطلاعات

استاد نیک آبادی

گزارش کار

فاز اول پروژه

محمد جواد رجبی

9831025

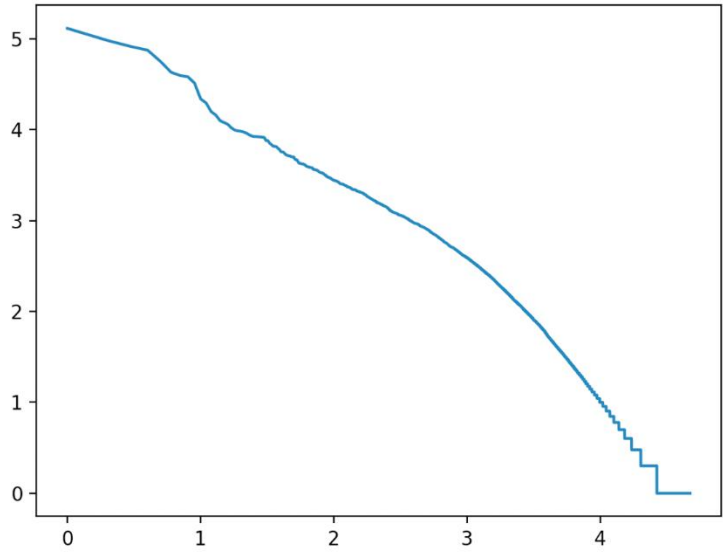
1. با ذکر مثال شرح دهید که در گام پیش پردازش چه عملیاتی انجام دادهاید. همچنین دلیل انجام هر پردازش را ذکر کنید.

در مرحله پیش پردازش ابتدا ما پس از خواندن فایل داده شده در صورت پروژه، متن های مورد نظر آن را نرمال سازی میکنیم تا از پردازش ساختار های غیر ضروری در برنامه مثل وجود whitespace جلوگیری شود و سپس شروع به استخراج توکن های موجود در آن میکنیم تا بوسیله آن دیکشنری موتور جستجوی خود را بسازیم و همزمان کلمه های پر تکرار یه به عبارتی stopwords ها را نیز از توکن های مان حذف میکنیم تا دیکشنری مان سبک تر شود و همچنین به جای پردازش و دخیره خود کلمات از ریشه ی آن ها استفاده میکنیم و در واقع ریشه یابی میکنیم.

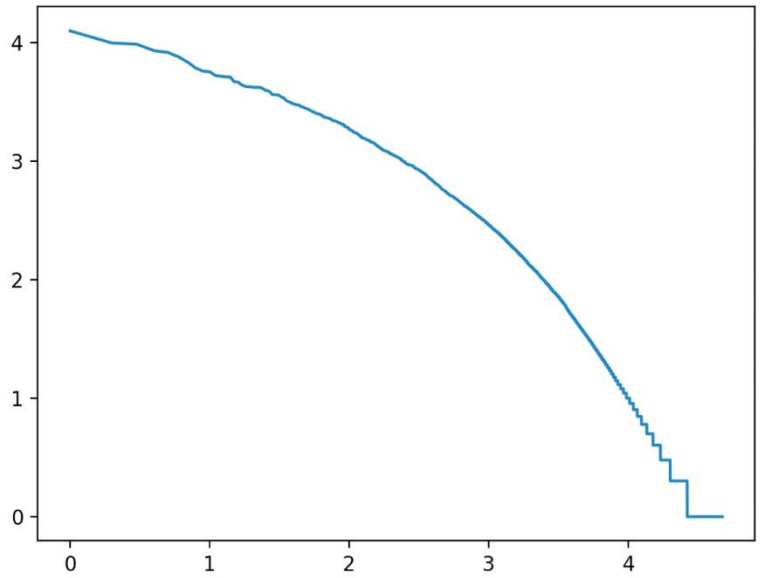
همچنین لازم به ذکر است برای نرمال سازی و حذف کلمات پر تکرار از کتابخانه Hazm و برای ریشه یابی از کتابخانه Parsivar استفاده شده است.

2. صحت قانون Zipf را در دو حالت قبل و بعد از حذف کلمات پرتکرار از واژهنامه بررسی کنید

قبل از حذف stopwords ها :



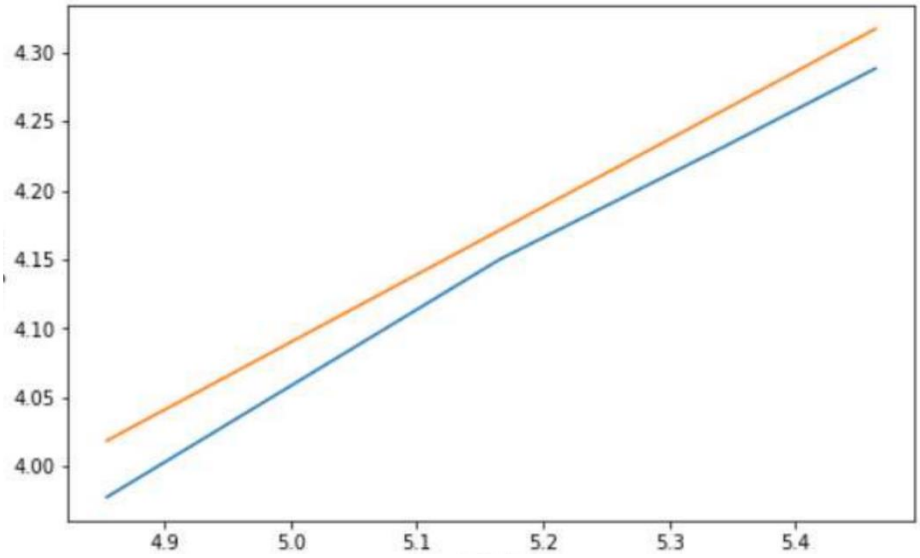
بعد از حذف stopwords ها:



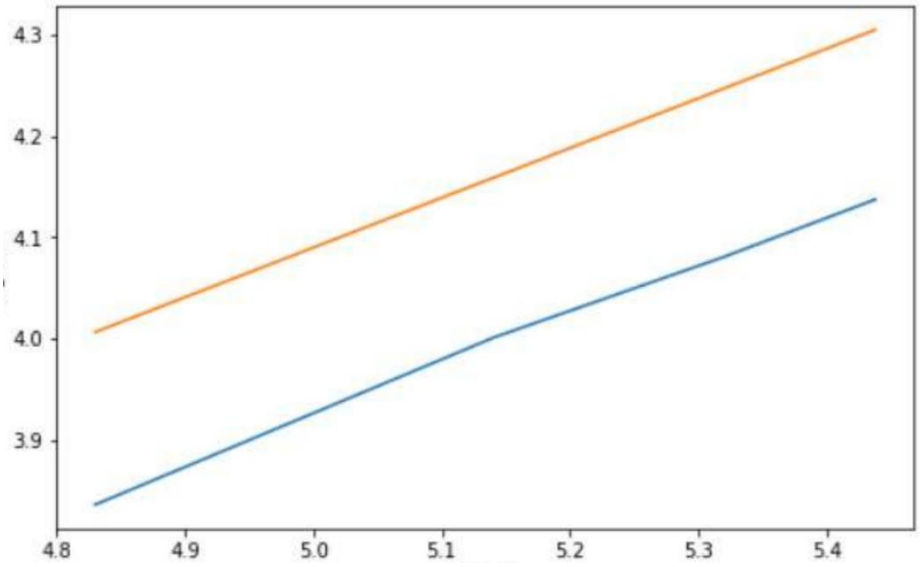
همانطور که در هر دو نمودار مشاهده کردید ، قانون Zipf در هر حالت قبل و بعد از حذف stopwords ها برقرار است .

3. صحت قانون heaps را در دو حالت قبل و بعد از ریشه یابی بررسی کنید

قبل از ریشه یابی:



بعد از ریشه یابی:



همانطور که در این دو نمودار مشاهده می کنید خط نارنجی نشان دهنده تعداد کلمات پیش بینی شده و خط آبی نشان دهنده تعداد واقعی می باشد و ملاحظه می شود که قبل از ریشه یابی این مقادیر بیشتر به مقادیر واقعی نزدیک تر هستند.

4. حداقل سه مورد از مواردی که در ریشه یابی با چالش روبرو بودید را ذکر کنید

1. حذف "ی" آخر بعضی از کلمات مثل تبدیل تختی به تخت
2. تبدیل اشتباه بعضی از اسم ها به ریشه های آن مثل تبدیل عرفان به عرف
3. حذف کامل یا تبدیل بعضی از کلمات به یک حرف مثل تبدیل حیات به ح
4. ریشه یابی نادرست بعضی از کلمات مثل تبدیل راست به راس

5. پاسخ به پرسمان در حالت های زیر:

الف) تحریم های آمریکا علیه ایران

```
PS C:\Users\NOVA\Desktop\Information Retrieval\project_phase1\IR_Project> python .\engine.py
enter what you want to search
تحریم های آمریکا علیه ایران

##############
Doc ID : 11864
Title : اهرم سازی از افغانستان در برجام/ نقطه عزیمت آمریکا در مذاکرات جامع با ایران چیست؟
URL : https://www.farsnews.ir/news/14000803000676/اهرم-سازی-از-افغانستان-در-برجام-نقطه-عزیمت-آمریکا-در-مذاکرات-جامع-با

##############
Doc ID : 9496
Title : گفت وگوی مشروح | ترقی: آمریکا شروط ایران را نپذیرد، پشت در مذاکرات می ماند/ روحانی کشور را به بن بست کشاند
URL : https://www.farsnews.ir/news/14000926000385/گفت-وگوی-مشروح-|-ترقی-آمریکا-شروط-ایران-را-نپذیرد-پشت-در-مذاکرات

##############
Doc ID : 8707
Title : نباید مانند دولت گذشته در مذاکرات افراط کرد/ «توافق موقت» راهبرد اصلی غرب و آمریکا در وین
URL : https://www.farsnews.ir/news/14001024000193/نباید-مانند-دولت-گذشته-در-مذاکرات-افراط-کرد-توافق-موقت-راهبرد-اصلی

##############
Doc ID : 9882
Title : چرا غرب مجبور به تمکین از خواسته  تهران است؟/ توان هسته ای؛ تنها یکی از ظرفیت های قدرت بخش ایران
URL : https://www.farsnews.ir/news/14000919000089/چرا-غرب-مجبور-به-تمکین-از-خواسته -تهران-است-توان-هسته ای-تنها-یکی-از

##############
Doc ID : 8577
Title : رییس جمهور در مسکو/ از تاکید پوتین بر توسعه روابط با ایران تا تشویق رییسی در دومای روسیه
URL : https://www.farsnews.ir/news/14001030000671/رییس-جمهور-در-مسکو-از-تاکید-پوتین-بر-توسعه-روابط-با-ایران-تا-تشویق
PS C:\Users\NOVA\Desktop\Information Retrieval\project_phase1\IR_Project>
```

ب) تحریم های آمریکا ! ایران

```
PS C:\Users\NOVA\Desktop\Information Retrieval\project_phase1\IR_Project> python .\engine.py
enter what you want to search
تحریم های آمریکا ! ایران

##############
Doc ID : 7261
Title : سود مافیای اسلحه سازی آمریکا در ناامن بودن جهان است
URL : https://www.farsnews.ir/news/14001211000898/سود-مافیای-اسلحه سازی-آمریکا-در-ناامن-بودن-جهان-است

##############
Doc ID : 8208
Title : لغو تحریم ها تنها ضرورت احیای برجام است
URL : https://www.farsnews.ir/news/14001110000419/لغو-تحریم ها-تنها-ضرورت-احیای-برجام-است

##############
Doc ID : 8063
Title : تاکتیکی جدید از اصلاح طلبان؛ مقصران دیروز، مدعیان امروز
URL : https://www.farsnews.ir/news/14001112001049/تاکتیکی-جدید-از-اصلاح طلبان-مقصران-دیروز-مدعیان-امروز

##############
Doc ID : 6878
Title : انتقاد دانشجویان ایرانی در اروپا به برخورد دوگانه مدعیان حقوق بشر با قضایای اوکراین و جنایت های آل سعود
URL : https://www.farsnews.ir/news/14001224000014/انتقاد-دانشجویان-ایرانی-در-اروپا-به-برخورد-دوگانه-مدعیان-حقوق-بشر-با

##############
Doc ID : 7627
Title : سران فتنه با ظلم شان در مشکلات اقتصادی امروز مردم شریک اند
URL : https://www.farsnews.ir/news/14001128000542/سران-فتنه-با-ظلم شان-در-مشکلات-اقتصادی-امروز-مردم-شریک اند
PS C:\Users\NOVA\Desktop\Information Retrieval\project_phase1\IR_Project>
```

پ) "کنگره ضدتروریست"

```
enter what you want to search
"کنگره ضدتروریست"

#############
Doc ID : 6929
Title : توضیحات یک منبع آگاه درباره وقفه مذاکرات وین
URL : https://www.farsnews.ir/news/14001222000450/توضیحات-یک-منبع-آگاه-درباره-وقفه-مذاکرات-وین
PS C:\Users\NOVA\Desktop\Information Retrieval\project_phase1\IR_Project>
```

ت) "تحریم های هسته ای" آمریکا ! ایران

```
PS C:\Users\NOVA\Desktop\Information Retrieval\project_phase1\IR_Project> python .\engine.py
enter what you want to search
"تحریم های هسته ای" آمریکا ! ایران

#############
Doc ID : 9694
Title : آمریکا با «راستی آزمایی» لغو تحریم ها مشکل دارد/ اطلاع روحانی از ماجرای افزایش قیمت بنزین
URL : https://www.farsnews.ir/news/14000926000407/آمریکا-با-راستی آزمایی-لغو-تحریم ها-مشکل-دارد-اطلاع-روحانی-از-ماجرای
PS C:\Users\NOVA\Desktop\Information Retrieval\project_phase1\IR_Project>
```

ث) اورشلیم ! صهیونیست

```
PS C:\Users\NOVA\Desktop\Information Retrieval\project_phase1\IR_Project> python .\engine.py
enter what you want to search
اورشلیم ! صهیونیست
There is nothing for you
PS C:\Users\NOVA\Desktop\Information Retrieval\project_phase1\IR_Project>
```