#### نگین خیرمند ۹۸۳۱۰۲۳ - علی انصاری ۹۸۳۱۱۳۶

### گزارش پروژه بازیابی اطلاعات

۱. دو عمل نرمالسازی متن و استخراج توکنها را از سندها با هم انجام دادیم. به این صورت که با استفاده از Normalizer عمل نرمالسازی و استخراج توکنها را انجام دادیم.

به طور مثال یکی از اختلافات رایج بین متون نیم فاصله و یا فاصله بین کلمات جمع است که با نرمالسازی به یک حالت تبدیل میشوند. سپس هر سند که نرمال شد، به کلمات تقسیم میشود و در نهایت یک آرایه نرمال شده از کلمات به ما برگردانده میشود.

سپس با استفاده از آرایه stop\_words کتابخانه hazm کلمات پرتکرار را حذف کردیم. همچنین علائم نگارشی پرتکراری که وجود داشت را نیز به صورت دستی با یک آرایه تعریف شده حذف کردیم.

به طور مثال تصویری از ده کلمه اول از ۳۸۹ کلمه آرایه hazm را چاپ کردهایم و همچنین لیست علائم نگارشی حذفشده را نیز ذکر کردهام.

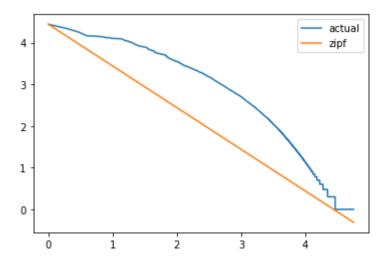
```
print(stop_words[:10])
print(len(stop_words))

['פ', 'د(', 'יִבּא', 'וֹנִ', 'كַּא', 'וֹנַי', '(', 'יִבּוֹ', 'יִבּוֹ')]

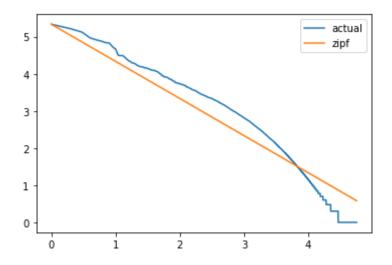
punctuations = ['.', '،', '/', ':', '(', ')', '!', '?', 'ṣ']
```

در نهایت عمل ریشه یابی را انجام دادیم. با استفاده از تابع FindStems از کتابخانه Parsivar این عمل را انجام دادیم. این تابع کلمات جمع را به صورت مفرد تبدیل و همچنین افعال را به صورت اول شخص یا ریشه آن درمی آورد. به طور مثال می شود به صورت شد&شو ذخیره میشود.

# قبل از حذف كلمات پرتكرار:

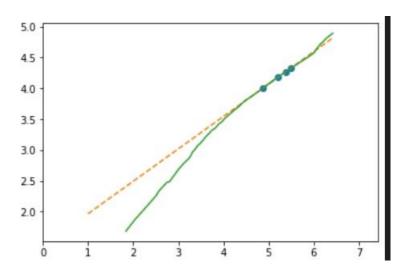


# بعد از حذف كلمات پرتكرار:



كاملا مشهود است كه نمودار ما پس از حذف كلمات پرتكرار بيشتر بر قانون zipf منطبق است.

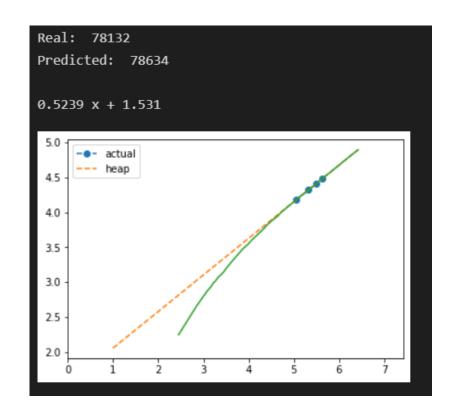
دلیل این مورد این است که چونکه کلماتی مانند از، در، به، و، با و سدر اکثر سندها آمدهاند، در قانون cf/i صدق نمیکنند. اما با حذف این کلمات میبینیم که نتیجه نزدیک به قانون zipf میشود.



سپس دیتاست را شافل کردیم و دوباره دادههارا استخراج کردیم که مشاهده شد در صورت شافلشدن و تقریبا نرمال شدن دیتاست، لگاریتم تعداد کلمات بر حسب تعداد کلمات استخراج شده خطی میشود.

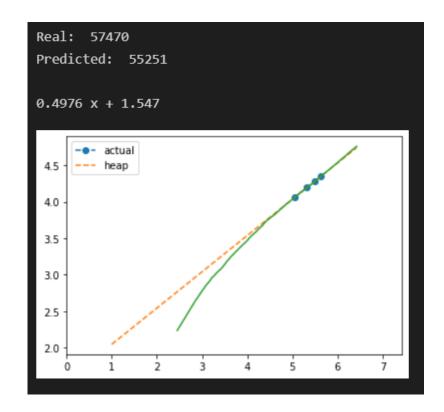
# قبل از ریشهیابی:

500 : 110452 14961 1000 : 210910 20774 1500 : 312548 25612 2000 : 418725 30120 All: 2643556 78132



#### بعد از ریشهیابی:

500 : 110452 11420 1000 : 210910 15616 1500 : 312548 19028 2000 : 418725 22205 All: 2643556 57470



مشاهده میکنیم که در هر دو حالت قانون heap برقرار است.

است k=33.96, b=0.52 ریشه یابی، k=33.96, b=0.52 است k=35.23, b=0.49 است که باز در شرایط صدق میکنند. در حالت پس از ریشه یابی نیز k=35.23, k=35.23, b=0.49 است که باز در شرایط صدق میکنند.

۴. در ریشه یابی به دلیل اینکه از کتابخانه Parsivar استفاده کردیم، مشکلی نخوردیم، اما مشاهده کردیم که این کتابخانه برخی کلمات را که نیاز نیست تغییر میدهد.

به طور مثال کلمه باشگاه را به باشگا تغییر داد یا کلمه آرارات را به آرار تغییر میدهد و کلمه است را به اس تغییر داد.

# الف) کوئری: تحریمهای آمریکا علیه ایران

در تابع نمایش دادن نتایج سعی شده است که حداقل تعداد جملهای که شامل تمامی کلمات کوئری میشود را

نمایش دهیم.

('تحریم مای آمریکا !ایران') ✓ 0.5s Pytho There are 24 results for your search 1 There are 14 matches for your search in this document یک مأموریت فتنه 88 کامش قدرت جانهزنی مسؤولان در مذاکرات بود حوزههای مختلف و ریشههای اصلی وقوع پدیده <mark>ت</mark>حریم اظهار داشت: فلسفه وضع تحریمها، افزایش فشار بر مردم و نظام بوده است تا تابآوری مردم از اسلامی، با اشاره به تاثیر فتنه 88 بر گسترش فشارهای تحریمی تصریح کرد: یکی از مأموریتهای فتنه 88، کاهش قدرت چانهزنی مسؤولان در پای میز مذاکره بود و حمتی اطلاعاتی در دست داریم که برخی آقازادهها در آن سال به <mark>ایریکا</mark> رفت و آمد میکردند و به غربیها و آمریکاییها پیغام میدادند که . ما نغس نظام را در داخل گرفته ایم و اگر شما هم فشارهای تحریمی را افزایش دهید، کار نظام تعام است 2 There are 13 matches for your search in this document سود مافیای اسلحهسازی آمریکا در ناامن بودن جهان است درقضایای جاری اوکراین هم شاهد نقش آفرینی مافیای پشت پرده انگلیسی، فرانسوی و آمریکایی هستیم که با وعده هایی باعث تحریک اوکراین شدند و نماینده ادوار مجلس ضمن تاکید به هماهنگ بودن کشورهای غربی گفت: کشورهای غربی به صورت گروهی عمل م<u>ی کنند و در</u> بزنگاه ها کاصلا هماهنگ . مستند همچنان که شاهد مستیم در جریان نبرد میان روسیه و اوکراین در اوج اتحا﴿ تحریم هایی را به روسیه تحمیل کردند 3 There are 10 matches for your search in this document خطیب جمعه تهران: سازمان همکاری اسلامی در برابر کشتار مردم یمن سکوت کرده است لزوم شكران نعمت انقلاب/ تعام تُحريم ها بايد لغو شود اصام جمعه موقت تهران، همچنين با اشاره به چهل و سومين سالگرد پيروزي انقلاب اسلامي \* اظهار داشت: دهه فجر از مصادیق ایام🏚 است؛ یعنی روزهایی که خداوند نعمت ویژهای به مردم داده است و جز با عنایت خداوند، این انقلاب پیروز ۰ خاتمی افزود: تمام دنیا اعم از آمریکاً) و شوروی و اروپاییها پشت سر شاه بودند، اما مرگ بر شاه گفتن٬های شما مردم باعث شد که شاه برود

4 There are 9 matches for your search in this document

مسائل اقتصادی کشور را معطل نتیجه مذاکرات نعیکنیم

مسائل اقتصادی کشور را معطل نتیجه مذاکرات نعیکنیم

جبار کوچکینژاد نعاینده مردم رشت در مجلس شورای اسلامی در گفتوگو با خبرنگار پارلمانی خبرگزاری فارس، گفت: تجربه مذاکرات در سالهای قبل با

، اروپا و آمریکا قابل اعتماد نیستند

عشو کمیسیون برنامه، بودجه و محاسبات مجلس اظهار داشت: ما باید با تکیه بر توان داخلی و ایجاد ثبات و رشد اقتصادی در کشور از این طریق در

مسیر خنثیسازی تحریم ما گام برداریم و با قوی شدن خودمان زمینه را برای بیاثر کردن تحریم ما فراهم کنیم

5 There are 7 matches for your search in this document

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

امریکا، اسرائیل و منافقین در خباثت و رذانتشان ما را حساس کردند و به همین خاطر، کشور ما برای مقابله با آنها به صنایع نظامی در حد

محیرالعقولی دست یافت که امروز مقتدرترین کشور در منطقه است و در جهان نیز موقعیتی دارد که هیچ کشوری جرأن حمله به خودش نمیدهد

محیرالعقولی دست یافت که امروز مقتدرترین کشور در منطقه است و در جهان نیز موقعیتی دارد که هیچ کشوری جرأن حمله به خودش نمیدهد

محیر العقولی دست یافت که امروز مقتدرترین کشور در منطقه است و در جهان نیز موقعیتی دارد که هیچ کشوری جرأن حمله به خودش نمیدهد

محیر العقولی دست یافت که امروز مقدر در منطقه است و در جهان نیز موقعیتی دارد که هیچ کشوری جرأن حمله به خودش نمیدهد

محبد الاملام صدیقی تصریح کرد: امروز تیم مذاکره کننده ما محکم و با مطالعه هستند و بسته هایی برای رفع همه تحریر عربی المقاله هستند و بسته هایی برای رفع همه تحریر عربی المقود کنده به تحریر المقاله همین خودش نمیدهد

با توجه به جملات چاپشده، کلمه ایران در خروجیها مشاهده نشد.

پ) کوئری "کنگره ضدتروریست"

```
query("کنگره شدتروریست") مدتروریست") به مدتروریست") مدتروریست") به وسود و 0.4s

There are 1 results for your search

1 There are 5 matches for your search in this document

و مدتروریست جلوه کند، 3 هفته بعد با صدور دستور ویژه ی (12613) ورود مرگونه کالا و خدمات از ایران را معنوع و دریاد دریاری در
```

## ت) كوئرى: "تحريم هستهاى" آمريكا اليران

## نتیجه ای به همراه نداشت. به همین دلیل عبارت ( "تحریم هستهای" آمریکا) را سرچ کردم.

```
و query("تحریم هسته ای" آمریکا") Python

There are 4 results for your search

1 There are 31 matches for your search in this document

1 There are 31 matches for your search in this document

1 There are 31 matches for your search in this document

1 There are 31 matches for your search in this document

1 There are 31 matches for your search in this document

1 There are 31 matches for your search alone of the part of t
```

#### 3 There are 5 matches for your search in this document

آمریکا با «راستیآزمایی» لغو تحریمها مشکل دارد/ اطلاع روحانی از ماجرای افزایش قیمت بنزین

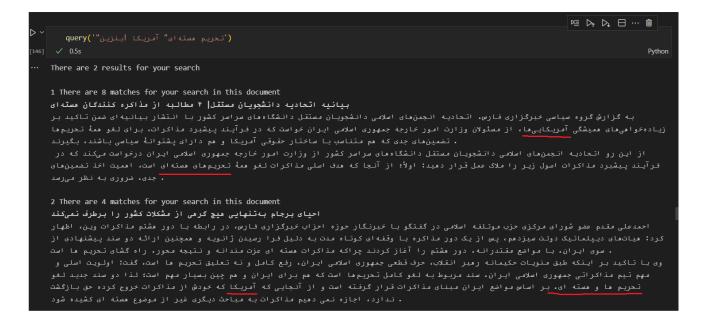
یک کلاهبرداری سیاسی کرد میدانست دولت بعدی اگر دموکرات باشد، میخواهد برگردد به برجام یک تعداد از تعریم های مسته ای ما را لیبلش را پرداشت کرد، تحریم های حقوق بشری بانک مرکزی را از هسته ای درآورد برد در تروریست ما در برجام تحریم های برجامی را قرار شده طرف مقابل پردارد آمریکای جدید دولت جدید می گوید این بانک مرکزی الان دیگر در هسته ای نیست، در تروریسم است، پس اگر می خواهید بانک مرکزی را بردارید، شما باید در تروریسم هم وارد مذاکره شوید، ما هر چه قدر به اینها می گوییم دوستان قبل هم همین را گفتند که آقا شما که می دانید این خود ترامپ این کار را کرد و اصلا این ربطی به تروریسم و حقوق بشر ندارد، ولی آنها تصویب کردند و از این نمیگذرند لذا اینها مواردی است که بعضی وقت ها می گفتند که ما 90 درصد کار را تعام کردیم، حجمی نگاه کنیم وزنی نگاه کنیم این 5 درصد کدام ها بوده است که مانده است ؟

#### 4 There are 4 matches for your search in this document

احمیای برجام بهتنهایی میچ گرمی از مشکلات کشور را برطرف نمیکند

احمدعلی مقدم عضو شورای مرکزی حزب موتلفه اسلامی در گفتگو با خبرنگار حوزه احزاب خبرگزاری فارس، در رابطه با دور هشتم مذاکرات وین، اظهار نرد: هیاتهای دیپلماتیک دولت سیزدهم، پس از یک دور مذاکره با وقفهای کوتاه مدت به دلیل فرا رسیدن ژانویه و معچنین ارائه دو سند پیشنهادی از . سوی ایران، با مواضع مقتدرانه، دور هشتم را آغاز کردند چراکه مذاکرات هسته ای عزت مندانه و نتیجه محور، راه گشای تحریم ها است وی با تاکید بر اینکه طبق منویات حکیمانه رهبر انقلاب، حرف قطعی جمهوری اسلامی ایران، رفع کامل و نه تعلیق تحریم ها است، گفت: اولویت اصلی و مهم تیم مناکراتی جمهوری اسلامی ایران، سند مربوط به لغو کامل تحریم ها است که هم برای ایران و مچین بسیار مهم است؛ لذا دو سند جدید لغو تحریم ها و هسته آی بر اساس مواضع ایران مبنای مذاکرات قرار گرفته است و از آنجایی که آمریکا که خودش از مذاکرات خروج کرده حق بازگشت

در نهایت برای اینکه مثالی برای قسمت ت داشته باشم عبارت ( "تحریم هستهای" آمریکا !بنزین) را جستجو کردم که با توجه به عدم وجود بنزین دو تا از نتایج کوئری قبل باقی ماندند.



#### ث) کوئری اور شلیم اصهیونیست

```
query('اورشليم أصهيونيست') ( علي المهيونيست') ( علي المهيونيست') ( علي المهيونيست') ( Python there are 0 results for your search
```

از آنجا که نتیجهای نداشت عبارت تنهای اورشلیم را جستجو کردم که باز هم نتیجهای نداشت یعنی کلا کلمه اورشلیم در اسناد ما وجود ندارد.

سیس برای انجام قسمت ث عبارت ( کامرون اتیم ) را جستجو کردم که تنها یک سند را آورد.

```
query('کامرون !نیم')

√ 0.2s

Python

There are 1 results for your search

1 There are 3 matches for your search in this document

1 There are 3 matches for your search in this document

جام صلت مای فوتبال آفریفا | شیرمای رام نشدنی با برد استارت زدند

به گزارش خبرگزاری فارس، رقابت مای فوتبال جام صلت مای آفریفا در کشور کامرون امروز از ساعت 30:10 به وقت ایران آغاز شد و شیرمای رام نشدنی

م قاره سیاه در مصاف با بورکینافاسو با نتیجه 2 بر یک به پیروزی رسیدند
```