((گزارش فاز اول پروژه بازیابی اطلاعات))

پویان حسابی ۹۷۳۱۱۲۲

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

در گام پیش پردازش مطابق با دستور پروژه و مباحث درس ابتدا رشته مربوطه را نرمال می‌کنیم برای اینکه یکسری از کاراکتر ها و به طور مثال اعداد فرم یکسانی بگیرند، سپس توکن های آن را استخراج می‌کنیم(هر کلمه به صورت جداگانه) و کلماتی از هر content (محتوای سند) را داریم. از آنجایی که نحوه ریشه یابی(stemming) بر اساس لغات در کتابخانه مربوطه انجام می‌شود سپس این تابع را فراخوانی می‌کنیم، دلیل آن این است که کلماتی مثل رفتیم، رفتند و ... یک کلمه حساب شوند و به نوعی به یک فرم باشند. سپس چک می‌کنیم که کلمه مربوطه stop word نباشد چرا که تاثیری در نتایج ما ندارد و صرفا عملیات را زیاد می‌کند، نمونه ای از استاپ ورد: از، به، تا، و ...

از آنجایی که در پایتون دسترسی به مقادیر دیکشنری(ساختمان داده dict) در O(1) انجام می‌شود دسترسی به محتوای هر سند را بر اساس {url(str) : content(list of tokens)} در نظر گرفته شده است و از طرفی محتوای ما که شامل توکن هایی از کلمات است باید در لیست قرار گیرد تا مکان آنها برای positional index استفاده شود.

```
def pre_processing():
    start_time = time.time()
    tokenized_dict = {}
    stopwords = stopwords_list()
    stemmed_text = []
    for url in url_content:
        tokenized_dict[url] = word_tokenize((Normalizer().normalize(url_content[url])))
        for word in tokenized_dict[url]:
            if word not in stopwords:
                stemmed_text.append(Stemmer().stem(word))
        tokenized_dict[url] = stemmed_text.copy()
        stemmed_text.clear()
    print("Time spent for pre processing: " + str(time.time() - start_time))
    return tokenized_dict
```

از کتابخانه hazm استفاده شده است که نمونه هایی از هر کدام از مراحل در پایین آورده شده:

```
>>> normalizer = Normalizer()
>>> normalizer.normalize('اصلاح نویسه ها و استفاده از نیم‌فاصله پردازش را آسان می کند')
'اصلاح نویسه‌ها و استفاده از نیم‌فاصله پردازش را آسان می‌کند'

>>> word_tokenize('ولی برای پردازش، جدا بهتر نیست؟')
['ولی', 'برای', 'پردازش', '،', 'جدا', 'بهتر', 'نیست', '؟']

>>> stemmer = Stemmer()
>>> stemmer.stem('کتاب‌ها')
'کتاب'
```

```
print(Stemmer().stem("گذاشتم"))
print(Stemmer().stem("گذاشت"))

Test
"C:\Program Files\Python\pyth
گذا
گذا
```

```
print(stopwords_list())

Test
"C:\Program Files\Python\python.exe" "D:/Edu/Term 8/Information Retrieval/Project/Final IR Project/Test.py"
['ابتدا' و 'شش' و 'الزامی' و 'می\u200cگویند' و 'مالیاتی' و 'فروش' و 'نمی‌شد' و 'یک‌است' و 'پر' و 'خاطرنشان' و 'گاه' و 'جمعی' و 'اغلب' و 'دوباره' و 'می\u200cیابد' و 'لذا' و 'زاده' و 'گرفت' و 'اینجا'
```

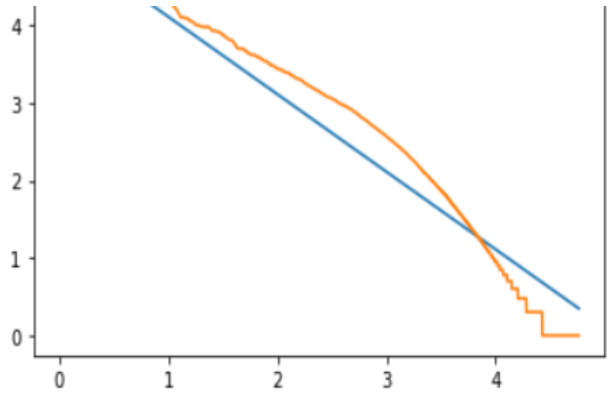
۲) صحت قانون Zipf را در دو حالت قبل و بعد از حذف کلمات پر تکرار از واژه نامه بررسی کنید(رسم نمودار برای هر حالت الزامی است.)در صورت برقراری/ عدم برقراری این قانون در هر حالت، علت را شرح دهید.

Zipf's law: The i^{th} most frequent term has frequency proportional to $1/i$.

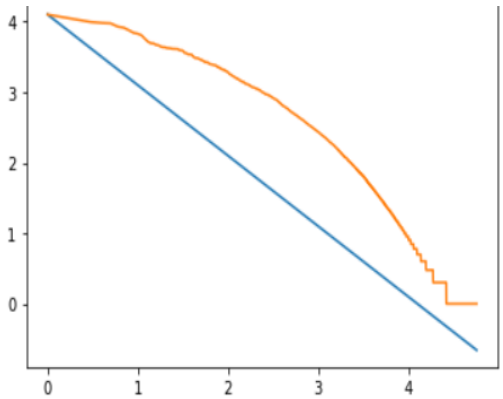
$cf_i \propto 1/i = K/i$ where K is a normalizing constant

cf_i is collection frequency: the number of occurrences of the term t_i in the collection.

حذف نکردن stop words:



حذف stop words:



همانطور که می‌دانیم و گفته شد قانون zipf مبتنی بر تعداد تکرار کلمات است. در واقع رابطه عکس با کلمات پر تکرار دارد. برای stop words از کتابخانه hazm استفاده شده که دقت کمتری نسبت به کتابخانه دیگر دارد ولی سرعت بسیار بیشتری دارد. همانطور که مشاهده می کنید با حذف کلمات پرتکرار فاصله زیادی با تخمین ایده آل(خط آبی) دارد ولی با حذف نکردن کلمات پر تکرار نمودار نارنجی خیلی نزدیکتر به نمودار آبی می باشد پس تخمین ما در حالت با بدون حذف stop words براساس قانون zipf بوده ولی وقتی آنها را حذف کنیم فاصله آن زیاد می‌شود. و در تئوری هم همچنین نتیجه ای انتظار می‌رفت چون با حذف کلمات پرتکرار در واقع آنها را از نمودار حذف می‌کنیم و نمودار نزدیک به داده های دیگر می‌شود و پراکندگی زیاد می شود.

۳) صحت قانون heaps:

بعد از ریشه یابی:

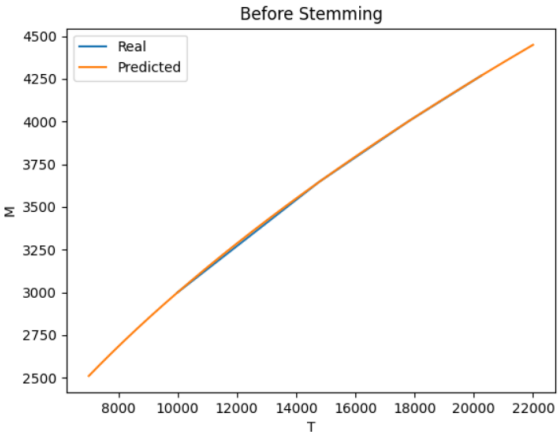
تعداد کل توکن ها	تعداد کلمات مجزا	تعداد اسناد
۴۷۶۲۹	۷۶۶۵	۵۰۰
۹۷۰۴۵	۱۱۰۰۶	۱۰۰۰
۱۴۵۶۱۳	۱۳۰۴۷	۱۵۰۰
۱۹۱۴۷۲	۱۴۷۶۶	۲۰۰۰

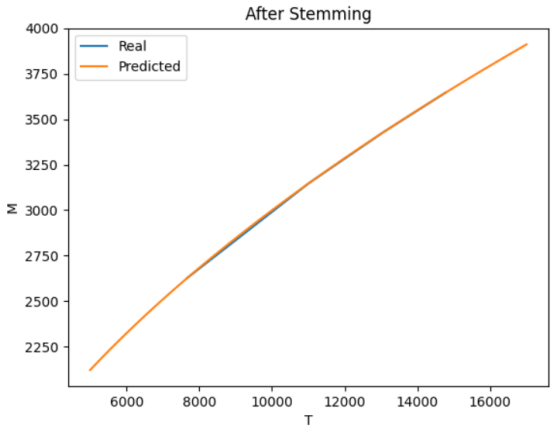
قبل از ریشه یابی:

تعداد کل توکن ها	تعداد کلمات مجزا	تعداد اسناد
۷۵۵۷۸	۹۹۸۸	۵۰۰
۱۵۴۳۵۵	۱۴۷۶۸	۱۰۰۰
۲۳۲۹۷۹	۱۷۸۲۱	۱۵۰۰
۳۰۶۰۳۲	۲۰۲۵۰	۲۰۰۰

M = k * power (T, b)

. b = 0.5 و k = 30 و T: number of tokens و M: size of vocabs





بر اساس نمودار ها مشاهده می‌شود تقریبا هر دو از قانون heaps تبعیت می‌کنند. در حالتی که بعد از ریشه یابی است مقدار نزدیک تری به مقداری که پیشبینی شده است دارد و به این معناست که ریشه یابی باعث می‌شود که قانون heaps پر رنگ تر شود و عملا بعد از آن شبیه تر است، در نتیجه می توان مشاهده کرد که با حذف کلمات هم ریشه علاوه بر اینکه حجم لغات کاهش می یابد دقت مدل ما نیز افزایش می یابد و مدل به حالت واقعی نزدیک می شود. و احتمالا هر چه تعداد اسناد بیشتر شود، بیشتر به این مقدار نزدیک می‌شود که به این معناست ریشه یابی لغات بسیار سودمند می‌باشد.

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

```
array = ["نوشتار" ,"است" ,"رفتند" ,"رفتیم" ,"رفتم" ,"انسان" ,"کتاب"]
for word in array:
    print(Stemmer().stem(word), end=" ")
```

نوشتار است رفتند رفت رف انس کتاب

مشاهده می‌شود کلمه (است) به (اس) تبدیل شده و به طور کلی حروف بعد از آن را حذف کرده است. کلمه (رفتند) همان (رفتند) باقی مانده و بدون تغییر است، کلمه (رفتیم) به درستی بوده ولی کلمه (رفتم) تبدیل به (رف) شده است و کلمه (انسان) به اشتباه ریشه یابی شده است.

۵) پاسخ به پرسمان در حالتهای زیر:

فرم کلی پاسخ به کوئری به این شکل می‌باشد که ابتدا شماره سند، url، عنوان چاپ می‌شود.

```
def show_ranked_documents(ranked_docs: list):
    for counter, tuple_ in enumerate(ranked_docs):
        if counter > 4:
            break
        doc_id = tuple_[0]
        print("***\nResult " + str(counter + 1))
        print("Document ID: " + str(doc_id))
        print("URL: " + str(urls[doc_id]))
        print("Title: " + str(url_title[urls[doc_id]]))
```

الف) تحریم های آمریکا علیه ایران

***

Result 1

Document ID: 9418

URL: https://www.farsnews.ir/news/14001005000843/بازخوانی-|-بیانات-رهبر-انقلاب-در-دیدار-دست‌اندرکاران-مراسم-سالگرد

Title: بازخوانی | بیانات رهبر انقلاب در دیدار دست‌اندرکاران مراسم سالگرد شهادت سردار سلیمانی

رفع تحریم دست دشمن است. ژنرال‌های جنگ نرم آمریکا و استکبار هستند. با ملّت ایران بدی کرد. بظاهر مذهبیِ متمایل به یکی از فرقه‌ها و علیه مقاومت را ایشان مدّتها قبل پیش‌بینی کرد

تحلیل: مشاهده می‌شود تنها یک سند بازگردانی شد چرا که اسنادی که هم آمریکا و هم ایران و هم علیه داشته باشند بسیار کم می‌باشند. این سند تا حدی مرتبط به کوئری کاربر است از این جهت که علیه را به عنوان یک استاپ ورد در نظر نگرفته و تاثیر بسزایی در کوئری گذاشته است. از طرفی لغات تحریم و آمریکا و ایران همگی در سند مربوطه موجود است.

ب)تحریم های آمریکا ! ایران

Result 1

Document ID: 6929

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

Title: توضیحات یک منبع آگاه درباره وقفه مذاکرات وین

برداشتن تحریم‌ها وعده داده عمل کند و تضمین معتبری هم برای انجام این تعهدات داده شود. آمریکا در مقابل عمل به این تعهد برجامی مقاومت می‌کند

***

Result 2

Document ID: 7209

URL: https://www.farsnews.ir/news/14001214001141/بحران-اوکراین-مصداق-ماهیت-مافیایی-آمریکاست

Title: بحران اوکراین، مصداق ماهیت مافیایی آمریکاست

***

Result 3

Document ID: 7183

URL: https://www.farsnews.ir/news/14001215000516/میزگرد-تلاطم-در-اوکراین|-ائتلاف-اقتصادی-حول-محور-دلار-فرو-می‌پاشد

Title: میزگرد «تلاطم در اوکراین»| ائتلاف اقتصادی حول محور دلار فرو می‌پاشد

***

Result 4

Document ID: 9742

URL: https://www.farsnews.ir/news/14000924000773/مرکز-پژوهش‌های-مجلس-مذاکرات-وین-به-توافقی-زودهنگام-منجر-نمی‌شود

Title: مرکز پژوهش‌های مجلس: مذاکرات وین به توافقی زودهنگام منجر نمی‌شود

***

Result 5

Document ID: 11864

URL: https://www.farsnews.ir/news/14000803000676/اهرم‌سازی-از-افغانستان-در-برجام-نقطه-عزیمت-آمریکا-در-مذاکرات-جامع-با

Title: اهرم‌سازی از افغانستان در برجام/ نقطه عزیمت آمریکا در مذاکرات جامع با ایران چیست؟

تحلیل: نتایج تعداد زیادی از اسناد می‌باشد که صرفا ۵ تا از آنها نمایش داده شده است. همانطور که دیده می‌شود هر ۵ سند کاملا مرتبط به کوئری می‌باشد و همگی مربوط به تحریم های آمریکا می‌باشد ولی یک باگ اساسی دارد این کوئری آن هم لغت ایران می‌باشد که وقتی ریشه یابی می‌شود کلمه ایرا می‌شود که به این شکل کوئری و نتایج را خراب می‌کند.

پ) کنگره ضدتروریست

Result 1

Document ID: 6929

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

Title: توضیحات یک منبع آگاه درباره وقفه مذاکرات وین

ریگان که نمی‌خواست کمتر از کنگره ضدتروریست جلوه کند

تحلیل: کلمه «ضدتروریست» تنها در یک سند آمده است. در اسناد دیگر کلمه ضد و تروریست جداگانه آمده که باعث می‌شود آنها در نتایج نیاورد. اینجاست که مفهوم spell correction اهمیت پیدا می‌کند چرا که کاربر کلمه را به اشتباه وارد کرده است.

ت) "تحریم هسته ای" آمریکا !ایران

Result 1

Document ID: 9496

URL: https://www.farsnews.ir/news/14000926000385/گفت‌وگوی-مشروح-|-ترقی-آمریکا-شروط-ایران-را-نپذیرد-پشت-در-مذاکرات

Title: گفت‌وگوی مشروح | ترقی: آمریکا شروط ایران را نپذیرد، پشت در مذاکرات می‌ماند/ روحانی کشور را به بن‌بست کشاند

محدود کردن مذاکره به موضوع رفع یکباره تحریم‌ها است. که موضوعی در مورد هسته‌ای باقی نمانده که بخواهد در مورد آن مذاکره کند. سه شرط را برای بازگشت آمریکا به برجام مطرح کرده است

***

Result 2

Document ID: 7230

URL: https://www.farsnews.ir/news/14001214000432/رئیس-دانشکده-هسته‌ای-دانشگاه-شهید-بهشتی-انرژی-هسته‌ای-نیاز-قطعی-امروز

Title: رئیس دانشکده هسته‌ای دانشگاه شهید بهشتی: انرژی هسته‌ای نیاز قطعی امروز و فردای ایران است

***

Result 3

Document ID: 11495

URL: https://www.farsnews.ir/news/14000813000113/گزارش-فارس-از-اجتماع-میدانی-یوم-الله-۳-آبان-در-تهران-تاکید-بر-ضرورت

Title: گزارش فارس از اجتماع میدانی یوم الله ۱۳ آبان در تهران/ تاکید بر ضرورت هوشیاری ملت‌های منطقه برابر توطئه‌های آمریکا

***

Result 4

Document ID: 11341

URL: https://www.farsnews.ir/news/14000816000568/ماجرای-دستگیری-دو-دانشمند-ایرانی-در-آمریکا-سلیمانی-مرا-در-بند

Title: ماجرای دستگیری دو دانشمند ایرانی در آمریکا/ سلیمانی: مرا در بند قاتلین و قاچاقچیان بین‌المللی زندانی کردند

***

Result 5

Document ID: 11489

URL: https://www.farsnews.ir/news/14000813000157/آمریکایی‌ها-عادت-کرده‌اند-مدام-از-ایران-شکست-بخورند-روایت-سرلشکر

Title: آمریکایی‌ها عادت کرده‌اند مدام از ایران شکست بخورند/ روایت سرلشکر سلامی از مواجهه سپاه با آمریکا در دریای عمان

تحلیل: این کوئری علی رغم آنچه در صورت پروژه گفته پیچیدگی زیادی ندارد چرا که Double quotation به راحتی از آن حذف می‌شود و تعداد کلمات بالا در کوئری سیستم ما را کند نمی‌کند زیرا فارغ از کوئری ماتریس ها ساخته می‌شوند. ۵ نتیجه ای که آمده همگی مرتبط هستند ولی همچنان مشکل کلمه ایران وجود دارد.

ث) اورشلیم ! صهیونیست

No result for given query, because of this word: اورشلیم

تحلیل: هیچ سندی کلمه اورشلیم را دارا نمی‌باشد برای همین هیچ نتیجه ای خروجی نداده است.