پروژه بازیابی اطلاعات

فاز اول

تهیه کننده:

سارا تاجرنیا

استاد مربوطه:

دکتر نیک آبادی

بهار ۱۴۰۰

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

در گام پیش پردازش عملیات های زیر را انجام میدهیم:

- استخراج توکن
- نرمال سازی متون
- حذف کلمات پر تکرار
- ریشه یابی

استخراج توکن:

برای استخراج توکن در دیتا های خوانده شده توالی کاراکترها را به نشانه های کلمه برش میدهیم. تا بتوانیم آنها را راحت تر طبقه بندی کنیم. این عملیات به صورت تابع word_tokenize در گیت اضافه شده پیاده سازی شده.

تحریم آمریکا در برابر ایران ⟵ تحریم، آمریکا، در، برابر، ایران

باید توجه شود که نیم فاصله ها در استخراج توکن گزاره ها جداسازی نمیشوند و با کد u200c\ مشخص میشود.

نرمال سازی متون:

نرمال سازی عبارات برای کلماتی استفاده میشوند که یک معنا دارند ولی به شکل های مختلفی نوشته میشود. ما باید عبارات را نرمال سازی کنیم تا جستجو در آنها دقیق تر، راحت تر و بهتر باشد. این عملیات به صورت تابع Normalize در گیت اضافه شده پیاده سازی شده.

3 دوره ⟵ ۳ دوره

حذف کلمات پر تکرار:

وجود کلمات پر تکرار در جستجو باعث به وجود آمدن خطای بیشتر میشود. برای مثال وجود کلمه به، و، یا و ... در برخی متون بسیار زیاد است اما تعدد آن طرفا به معنای مربوط بودن گزاره به سند نیست در نتیجه سبب افزایش خطا میشود.
مجموعه این کلمات پرتکرار به صورت stowords_list در hazm وجود دارند.

ریشه یابی:

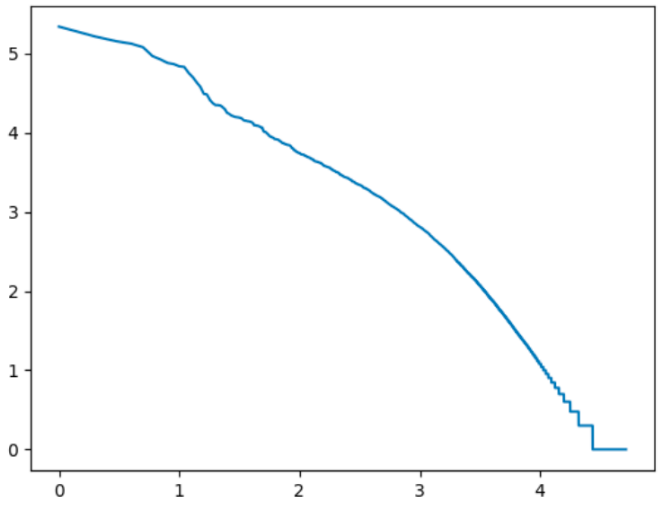
در عبارات اصولا کلماتی که ریشه یکسان دارند ممکن است تفاوت چندانی در نتیجه ی اسناد به صورت آماری نداشته باشند. البته روش این ریشه یابی در زبان های مختلف میتواند متفاوت باشد. این در صورتی است که بخواهیم اشکال مختلف یک ریشه مطابقت داشته باشند. تابع این عملیات در گیت اضافه شده پیاده سازی شده است.

آورده است ⟵ آورد#آور

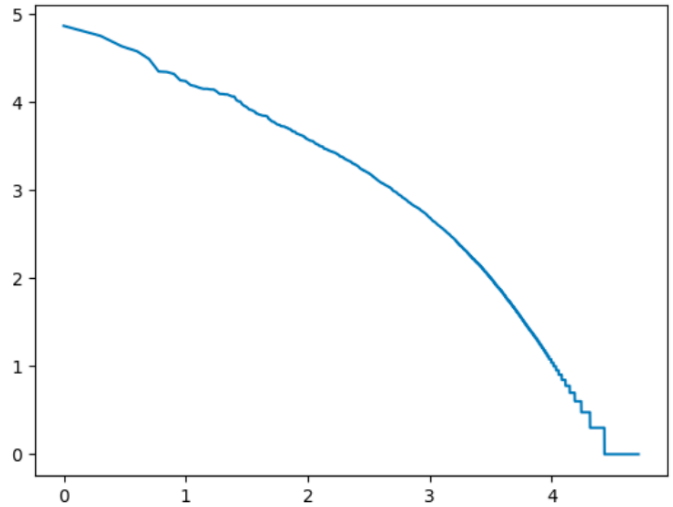
قهرمانی ⟵ قهرمان

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

قبل از حذف کلمات پر تکرار:



بعد از حذف کلمات پر تکرار:



۳. صحت قانون heaps را در دو حالت قبل و بعد از ریشه‌یابی بررسی کنید. برای بررسی این قانون لازم است با استفاده از اندازه‌ی واژه‌نامه و تعداد توکن‌ها در ۵۰۰، ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ سند اول، اندازه‌ی واژه‌نامه مربوط به کل اسناد تخمین زده شود. در نهایت اندازه‌ی واقعی واژه‌نامه و اندازه‌ی تخمینی در هر دو حالت مقایسه و تحلیل شود. آیا در هر دو حالت قانون برقرار است؟ چرا؟ (رسم نمودار برای هر حالت الزامی است.)

جدول تعداد واژه ها و توکن ها قبل و بعد از ریشه گیری:

	۵۰۰ سند اول	۱۰۰۰ سند اول	۱۵۰۰ سند اول	۲۰۰۰ سند اول
تعداد کل واژه ها	۱۳۳۱۰۰	۲۷۲۷۰۱	۴۱۳۱۵۹	۵۴۳۴۹۸
تعداد توکن ها قبل از ریشه یابی	۹۹۶۱	۱۴۷۵۷	۱۷۸۰۴	۲۰۲۳۶
تعداد توکن ها بعد از ریشه یابی	۷۵۰۵	۱۰۸۲۳	۱۲۸۷۰	۱۴۵۴۷

برای حالت قبل از ریشه یابی:

$\log M = \log k + b \log T$

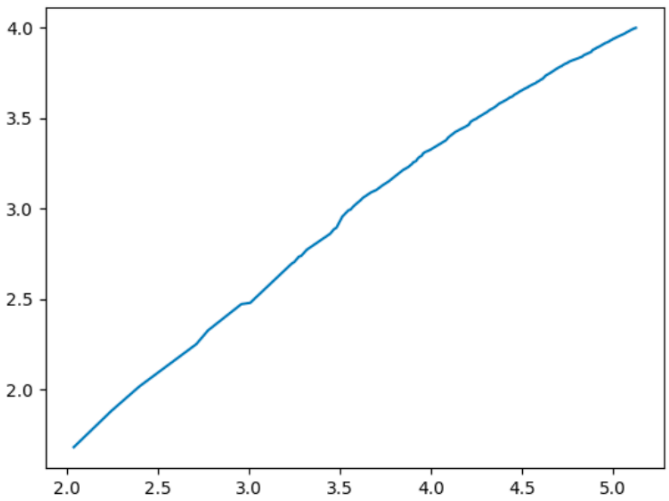
*log 9961 =log k + b *log 133100*

*log 14757 = log k + b *log 272701*

➔ *1/b=* $\frac{\left(\log_{10}(272701) - \log_{10}(14757)\right)}{\left(\log_{10}(14757) - \log_{10}(9961)\right)}$ *b=~ 0.53*

➔ *Log(k) = (b * log(272701)) – log(14757)*

➔ *k ~=19.41243*



برای حالت بعد از ریشه یابی:

$\log M = \log k + b \log T$

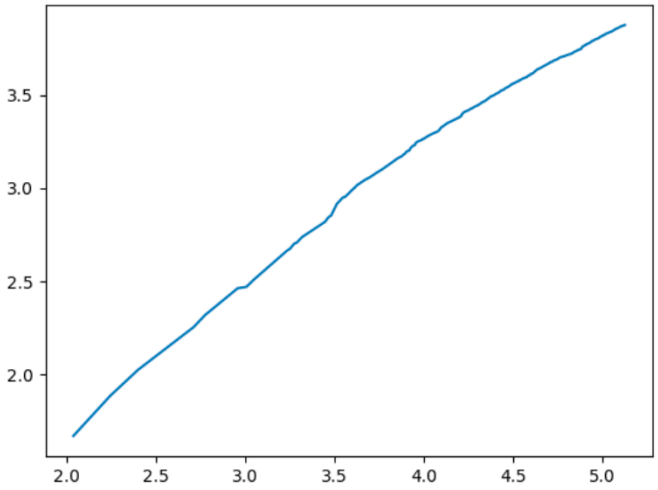
*log 7505 =log k + b *log 133100*

*log 10823 = log k + b *log 272701*

➔ $1/b = \frac{\left(\log_{10}(272701) - \log_{10}(133100)\right)}{\left(\log_{10}(10823) - \log_{10}(7505)\right)}$ *b=~ 0.51*

➔ *Log(k) = (b * log(272701)) – log(10823)*

➔ *k ~=18.28689*



۴. حداقل سه مورد از مواردی که در ریشه‌یابی با چالش روبرو بودید را ذکر کنید. (بطور مثال کلماتی که نیازی به ریشه‌یابی ندارند اما طبق روند ریشه‌یابی از دست می‌روند.)

برای ریشه یابی کلمات بسیاری از بین میرفتند برای مثال:

1- ممکن است "ان" یا "ها" به اشتباه علامت جمع تلقی شده و حذف شوند در حالی که نباید این اتفاق بیوفتد.

انتها ⟵ انت پایان ⟵ پای

2- ممکن است "تر" به اشتباه به عنوان علامت مقایسه حذف شود.

کبوتر ⟵ کبو

3- ممکن است " م " آخر کلمه به عنوان میم مالکیت به اشتباه حذف شود.

جام ⟵ جا

۵. پاسخ به پرسمان در حالت‌های زیر:
الف) یک پرسمان از کلمات ساده و متداول (مانند تحریم‌های آمریکا علیه ایران، در نتایج بازیابی‌شده انتظار می‌رود اسنادی که کلمات تحریم، آمریکا، علیه و ایران را دارند در بالای لیست و اسنادی که برخی از کلمات را ندارند در رتبه‌های پایین‌تر لیست قرار داشته باشند.)

۵ سند برتر برای عبارت تحریم‌های آمریکا علیه ایران :

- مرکز-پژوهش‌های-مجلس-مذاکرات-وین-/https://www.farsnews.ir/news/14000924000773 : 9742 به-توافقی-زودهنگام-منجر-نمی‌شود

جمله ای از سند: رای «مقصرسازی» ایران، از اعمال تحریم جدید و ابراز آمادگی در جهت اجرای سخت‌گیرانه‌تر تحریم برای «مأیوس کردن» ایران و درنهایت از تهدید به حمله نظامی برای «ایجاد هراس» از پیشرفت برنامه هسته‌ای استفاده کند.
***در این سایت به دفعات بسیار بالا هر یک از کلمات تحریم، آمریکا، علیه و ایران دیده شد و سایت بسیار شبیه و مرتبط بود.

- توضیحات-یک-منبع-آگاه-درباره-وقفه-/https://www.farsnews.ir/news/14001222000450 : 6929 مذاکرات-وین

یکی از تیترها: *تحریم‌های آمریکا علیه ایران
قسمتی از جملات سند: حدود 1.4میلیارد دلار در بانک «فدرال رزرو» آمریکا، 5.6میلیارد دلار در شعب مختلف بانک‌های این کشور در خارج از آمریکا، 2 میلیارد دلار در سایر بانک‌های آمریکا و حدود 2 میلیارد دلار در اختیار شهروندان و مؤسسه‌های خصوصی ایالات متحده بود.
***این سایت مربوط است زیرا کلمات آمریکا، ایران و تعداد زیادی کلمه تحریم در آن به کار رفته است.

- 11864 : https://www.farsnews.ir/news/14000803000676/اهرم‌سازی-از-افغانستان-در-برجام-نقطه-عزیمت-آمریکا-در-مذاکرات-جامع-با

تیتر: نقطه عزیمت آمریکا در مذاکرات جامع با ایران چیست؟

***این سایت همان طور که از تیتر آن هم مشخص است مربوط است و تعدد کلمات مربوط زیاد است.

- 9496 : https://www.farsnews.ir/news/14000926000385/گفت‌وگوی-مشروح-|-ترقی-آمریکا-شروط-ایران-را-نپذیرد-پشت-در-مذاکرات

تیتر: ترقی: آمریکا شروط ایران را نپذیرد، پشت در مذاکرات می‌ماند.

***کلمات مرتبط چه در تیتر چه متن بسیار زیاد است. پس سند مرتبط است.

- 9882 : https://www.farsnews.ir/news/14000919000089/چرا-غرب-مجبور-به-تمکین-از-خواسته-تهران-است-توان-هسته‌ای-تنها-یکی-از

تکه ای از جملات سند: شاید توان هسته‌ای ایران به جهت جلوه رسانه‌ای و القای خطر ایران هسته ای، بیشتر در افکار عمومی مستمسک فشار به ایران قرار گیرد، ولی بی‌شک مولفه‌های دیگر قدرت‌بخش ایران نیز کمتر از توان هسته‌ای برای آمریکا و صهیونیست‌ها نیستند.

***کلمات مربوطه به خصوص ایران به دفعات بالا تکرار شده است. پس سند مرتبط است.

ب) یک پرسمان با عملگر NOT (مانند تحریم‌های آمریکا ! ایران، انتظار می‌رود اسنادی که شامل دو کلمه تحریم و آمریکا هستند اما کلمه‌ی ایران را ندارند در نتایج بازیابی‌شده وجود داشته باشند.)

۵ سند برتر برای عبارت تحریم‌های آمریکا !ایران :

- 7249 : https://www.farsnews.ir/news/14001213000089/آمریکا-با-بحران‌آفرینی-به-حیات-خود-ادامه-می-دهد

تیتر: آمریکا با بحران‌آفرینی به حیات خود ادامه می دهد

***چه تیتر اصلی چه تیتر های فرعی همگی کاملا در مورد آمریکا است، پس مربوط است.

- 7261 : https://www.farsnews.ir/news/14001211000898/سود-مافیای-اسلحه‌سازی-آمریکا-در-ناامن-بودن-جهان-است

تیتر: سود مافیای اسلحه‌سازی آمریکا در ناامن بودن جهان است

***کل متن در مورد آمریکا است و حتی تحریم در متن آمده. سند مرتبط است.

- 6992 : https://www.farsnews.ir/news/14001211000321/آمریکا-دولت‌های-متحد-خود-را-در-زمان-اضطرار-تنها-می‌گذارد

یکی از جملات سند: وی افزود: دولتمردان بعضی کشورها که در مسیر سیاست‌های آمریکا و غرب حرکت می‌کنند، بدانند که وعده‌های آمریکا برای حمایت از آنها دروغ است و آمریکا در شرایط اضطرار آنها را تنها خواهد گذاشت..

***موضوع در مورد آمریکا و حتی با کوتاهی متن بارها و بارها تکرار شده است، سند مرتبط است.

- 11525 : https://www.farsnews.ir/news/14000812000668/ماهیت-استکباری-آمریکا-را-به-دنیا-معرفی-کنیم-تقدیر-از-اقدام-سپاه-برای

قسمتی از متن سند: معاون پارلمانی رئیس جمهور با بیان اینکه باید ماهیت استکباری آمریکا را به دنیا معرفی کنیم، گفت: در دهه گذشته برخی ها تلاش کردند چهره آمریکا را بزک کنند و بگویند با تعامل با آمریکا مشکلات کشور حل می شود اما همه نتیجه این تعامل که هیچ عایدی برای ما نداشت را دیدند.

***این سند به مسائل آمریکا بسیار مرتبط است، پس در کل سند مربوط است.

- 7071 : https://www.farsnews.ir/news/14001217000665/خباثت‌های-آمریکا-در-برجام-روسیه-و-چین-را-هم-به-این-کشور-بدبین-کرد

تیتر: خباثت‌های آمریکا در برجام روسیه و چین را هم به این کشور بدبین کرد

***تیتر و متن هر دو مربوط اند پس سند مربوط است.

پ) یک پرسمان با عملگر عبارت (مانند "کنگره ضدتروریست"، انتظار می‌رود اسنادی که شامل عبارت کنگره ضدتروریست در نتایج بازیابی‌شده وجود داشته باشند؛ بعبارت دیگر موقیت مکانی کلمات در این حالت مهم است.)

۵ سند برتر برای عبارت کنگره ضدتروریست:

- 7605 : https://www.farsnews.ir/news/14001130000185/سه-نکته-مهم-درباره-مصاحبه-وزیر-خارجه-با-فایننشال‌تایمز

قسمتی از متن سند: تصویب بیانیه سیاسی با عنوان resolution یکی از فعالیت‌های معمول در کنگره آمریکا است. از ابتدای شروع به کار دور جدید کنگره (ژانویه ۲۰۲۱) تاکنون ۹۲۹ بیانیه سیاسی در مجلس نمایندگان و ۵۱۷ بیانیه سیاسی در سنا ارائه شده است (لیست بیانیه‌های سیاسی کنگره فعلی در این صفحه) اینجا (قابل مشاهده است.)
***کلمه کنگره بارها تکرار شده و مربوط است البته نه خیلی زیاد! پس میتوان گفت سند تا حدودی مربوط است.

- 10656 : https://www.farsnews.ir/news/14000904000364/سومین-کنگره-سراسری-جمعیت-جانبازان-برگزار-می‌شود

قسمتی از متن سند: بر اساس جمع بندی در این جلسه، در سومین کنگره جمعیت جانبازان انقلاب اسلامی که با توجه به شرایط شیوع کرونا با هماهنگی وزارت کشور به صورت مجازی برگزار خواهد شد، موادی از اساسنامه این جمعیت در مجمع عمومی اعضا اصلاح می‌شود. در این کنگره همچنین اعضای جمعیت نسبت به انتخاب شورای مرکزی در دوره جدید و نیز انتخاب بازرسان اقدام خواهند کرد.
***کلمه کنگره پسیار تکرار شده پس میتوان گفت به تا حدودی مربوط است.

- 7423 : https://www.farsnews.ir/news/14001206000606/محصولیِ-۱۲-میلیاردی-را-محکوم-می‌کردند-اما-با-وزرای-۱۰۰۰-میلیاردی

جمله ای از سند: **نامه اخیر 250 نماینده مجلس در واکنش به نامه نمایندگان کنگره آمریکا به بایدن بود
***تعداد محدودی کلمه کنگره آمده پس میتوان گفت کمی مربوط است.

- 11235 : https://www.farsnews.ir/news/14000819000116/جلسه-شورای-مرکزی-جدید-کارگزاران-در-غیاب-کرباسچی-مرعشی-دبیرکل-حزب-شد

تکه کوتاهی از سند: رییس کنگره حزب کارگزاران خاطرنشان کرد: کرباسچی در کنگره پیشین، بار دیگر بر این موضع پافشاری کرد و حتی گفت که این آخرین دوره‌ای است که به عنوان دبیرکل مسئولیت حزب را برعهده می‌گیرم.

***تعدادی کلمه کنگره و مانند قبلی هیچ کلمه ضدتروریست ندارد پس میتوان گفت به مقدار کمی مربوط است.

- 7547 : https://www.farsnews.ir/news/14001201000913/-مانع-بدعهدی-دوباره-آمریکا-در-توافق-صرفا-تضمین-بازدارنده-است

پاراگرافی از متن: وی با بیان اینکه این سادگی است که فرض کنیم اگر توافق در کنگره آمریکا تصویب شد، آمریکایی‌ها آن را زیر پا نمی‌گذارند،گفت: برخی به این استناد می‌کنند که کنگره آمریکا با دوسوم آرای نمایندگان می‌تواند وتوی رئیس جمهور را خنثی کند و اگر کنگره آمریکا دوسوم از چیزی را قبول داشته باشد، رئیس جمهور آمریکا نمی‌تواند آنرا ملغی کند.

***تعداد کمی کلمه کنگره در سند آمده و مقدار خیلی کمی میتوان گفت مربوط است.

ت) یک پرسمان پیچیده (مانند "تحریم هسته‌ای" آمریکا ! ایران، انتظار می‌رود اسنادی که شامل عبارت تحریم هسته‌ای و کلمه‌ی آمریکا هستند اما کلمه‌ی ایران را ندارند در نتایج بازیابی‌شده وجود داشته باشد.)

۵ سند برتر برای عبارت "تحریم هسته‌ای" آمریکا !ایران (**توجه شود که در این اسناد تحریم هسته ای آورده نشده اما ایران آورده شده است پس به طور کامل نامربوط نیست و به مقدار کمی میتوان گفت مربوط است**):

- 7249 : https://www.farsnews.ir/news/14001213000089/-آمریکا-با-بحران‌آفرینی-به-حیات-خود-ادامه-می-دهد

تیتر: آمریکا با بحران‌آفرینی به حیات خود ادامه می دهد

***آمریکا به تعدد تکرار شده و در متن اصلی کلمه ایران وجود ندارد، متن خیلی خیلی کم مرتبط است.

- 7261 : https://www.farsnews.ir/news/14001211000898/-سود-مافیای-اسلحه‌سازی-آمریکا-در-ناامن-بودن-جهان-است

تیتر: سود مافیای اسلحه‌سازی آمریکا در ناامن بودن جهان است

***آمریکا به تعدد تکرار شده و در متن اصلی کلمه ایران وجود ندارد، متن خیلی خیلی کم مرتبط است.

- 6992 : https://www.farsnews.ir/news/14001211000321/-آمریکا-دولت‌های-متحد-خود-را-در-زمان-اضطرار-تنها-می‌گذارد

تیتر: آمریکا دولت‌های متحد خود را در زمان اضطرار تنها می‌گذارد

***در این سند آمریکا چند بار تکرار شده و همچنین در متن اصلی کلمه ایران وجود ندارد، متن خیلی خیلی کم مرتبط است.

- 11525 : https://www.farsnews.ir/news/14000812000668/ماهیت-استکباری-آمریکا-را-به-دنیا-معرفی-کنیم-تقدیر-از-اقدام-سپاه-برای

قسمتی از متن سند: معاون پارلمانی رئیس جمهور با بیان اینکه باید ماهیت استکباری آمریکا را به دنیا معرفی کنیم، گفت: در دهه گذشته برخی ها تلاش کردند چهره آمریکا را بزک کنند و بگویند با تعامل با آمریکا مشکلات کشور حل می شود اما همه نتیجه این تعامل که هیچ عایدی برای ما نداشت را دیدند.

***سند دارای تعدادی آمریکا است و ایران را در متن اصلی ندارد پس سند خیلی خیلی کم مرتبط است.

- 7071 : https://www.farsnews.ir/news/14001217000665/خباثت‌های-آمریکا-در-برجام-روسیه-و-چین-را-هم-به-این-کشور-بدبین-کرد

***دارای کلمات مربوط و خالی از کلمه نامربوط است پس میتوان گفت این سند خیلی خیلی کم مرتبط است.

ث) یک پرسمان کلمات نادر (مانند اورشلیم ! صهیونیست، خروجی مورد انتظار این قسمت مشابه با قسمت ب می‌باشد با این تفاوت که کلمات استفاده شده در پرسمان از کلمات نادر هستند.)

‘‘تحریم هسته‌ای’’ آمریکا !ایران

***به دلیل موجود نبودن سندی که دارای کلمه اورشلیم باشد و صهیونیست را نداشته باشد برنامه چیزی را برنمیگرداند.