محمد توکلی

فاز اول پروژه

۱)

در گام پیش پردازش مراحل زیر را طی کرده ایم:

- متون را نرمالایز کرده ایم
- Token هارا استخراج کرده ایم
- ریشه یابی کرده ایم
- کلمات پرتکرار (stop-words) هارا حذف کرده ایم

برای سه مورد اول از کتابخانه parsivar و در مورد آخر از کتابخانه hazm استفاده شده است.

که توابع آن به صورت زیر پیاده سازی شده اند:

```
def normalize(input_text):
    output_text = my_normalizer.normalize(input_text)
    return output_text


def get_tokens(input_text):
    words = my_tokenizer.tokenize_words(my_normalizer.normalize(input_text))
    return words


def get_steams(input_words):
    words = []
    for word in input_words:
        words.append(my_stemmer.convert_to_stem(word))
    return words


def delete_stop_words(input_words):
    news_words = []
    stop_words = stopwords_list()
    for word in input_words:
        if word not in stop_words:
            news_words.append(word)

    return news_words
```

علت نرمالایز سازی : استفاده از نرمال سازی باعث میشود تمام متون از نظر فاصله و نیم فاصله، تاریخ ها، اعداد و ... یکسان شوند و مثلا در متنی عدد ها فارسی و در متن دیگر انگلیسی نباشند و این قواعد در تمام متون به صورت یکسان رعایت شود.

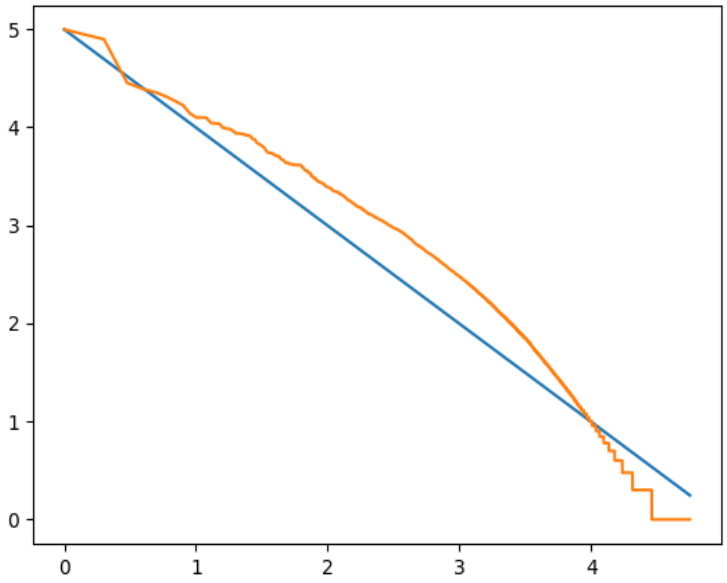
علت استخراج توکن : قسمت اصلی ساخت موتور های جست و جو ساختن توکن هاست که درواقع به وسیله توکن های میتوانیم لیست شاخص های مکانی را بسازیم و توسط آن به کوئری کاربر پاسخ دهیم.

علت ریشه یابی : ریشه یابی باعث میشود کلمه هایی مانند: رفت، رفتم، رفتی، رفتیم، رفتند، دارم میرم، داشتم میرفتم و.... و تمامی اشکال فعل ها و اسم ها تنها به یک کلمه تبدیل شود و زمانی که کاربر هرکدام از موارد ذکر شده را وارد کرد اسناد مربوط به دیگر حالت ها نیز برگردانده شوند.

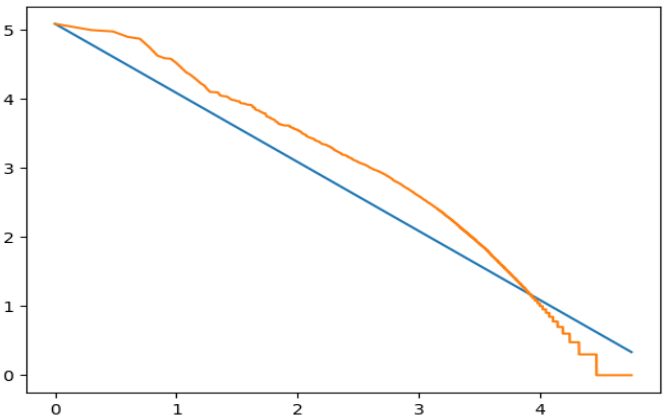
علت حذف کلمات پرتکرار : علت این کار هم این است که برای جست و جو کاربر کلماتی نظیر است، تو، "."، کوچک و ... از لیست توکن ها حذف شوند زیرا برای کاربر وجود این کلمات در کوئری در نتایج برگردانده شده تاثیری تقریبا ندارد و کاهش حجم بسیار ارزنده تر از مقدار کمی کاهش دقت در نتایج بازگردانده شده است.

۲)

نمودار برای قبل از حذف کردن stop-word ها به صورت زیر است:



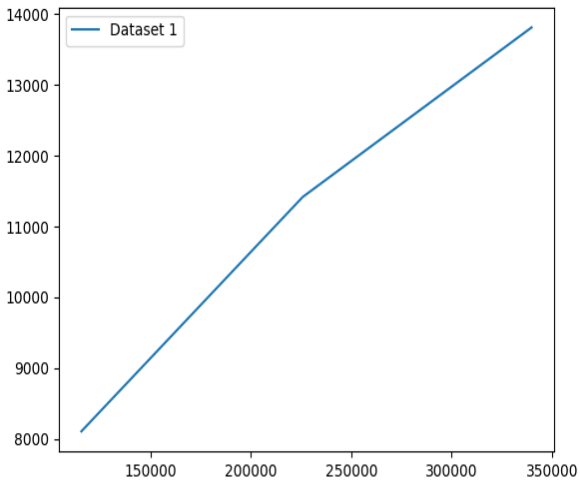
نمودار برای بعد از حذف کردن stop-word ها به صورت زیر است:

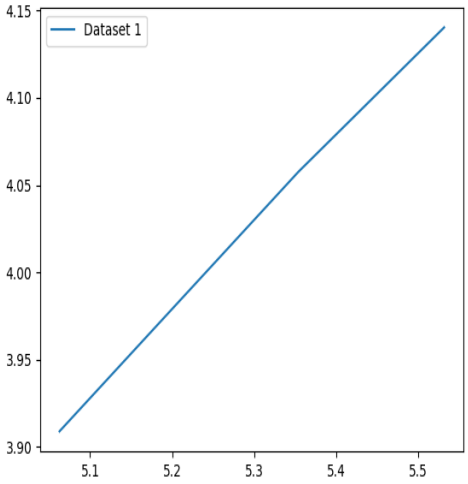


میبینیم که در دو حالت قانون برقرار است و نمودار ها تقریبا بر هم منطبق هستند.

۳)

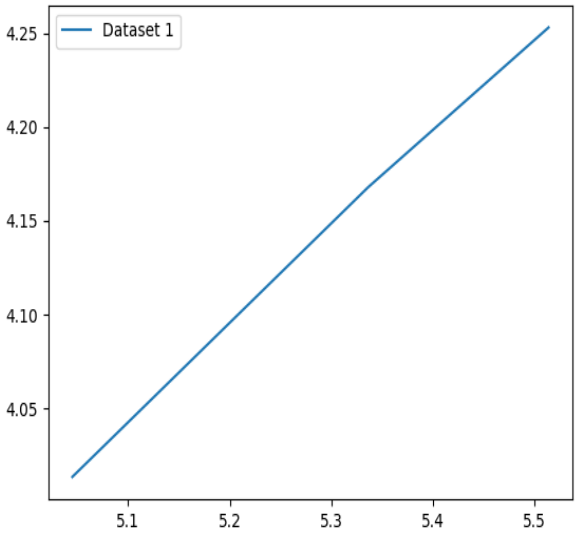
نمودار های زیر مربوط به بعد از ریشه یابی است:

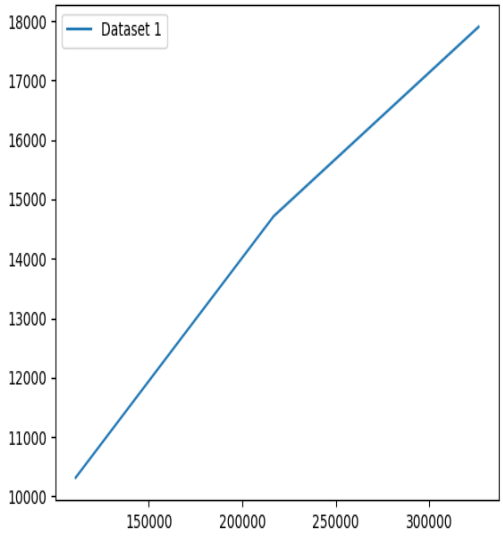




که یکی از آنها لوگاریتمی و دیگری بدون لوگاریتم گیری رسم شده

نمودار های زیر مربوط به قبل از ریشه یابی است:





که یکی از آنها لوگاریتمی و دیگری بدون لوگاریتم گیری رسم شده

میبینیم که در دوحالت نمودار ها منطبق بر هم هستند.

در حالت قبل از ریشه یابی داریم:

تعداد توکن ها: ۳۲۶۳۹۲

تعداد کلمه ها: ۱۷۹۰۴

تعداد کل توکن ها: ۱۹۵۹۵۵۲

تعداد کل کلمه ها: ۷۲۸۹۰

که به ازای b = 0.67 رابطه برقرار می شود.

در حالت بعد از ریشه یابی داریم:

تعداد توکن ها: ۳۳۹۸۸۳

تعداد کلمه ها: ۱۳۸۱۲

تعداد کل توکن ها: ۲۰۷۸۶۸۴

تعداد کل کلمه ها: ۵۵۸۹۰

که به ازای b = 0. 66 رابطه برقرار می شود.

۴)

- کلمه دادند که بعد از ریشه یابی به کلمه "ده" تبدیل شد که میتواند با عدد ده اشتباه گرفته شود
- کلمه است که بعد از ریشه یابی به کلمه "اس" تبدیل شده است و لزومی برای ریشه یابی وجود نداشت
- کلمه می افتد که به معنای اتفاق افتادن هست بعد از ریشه یابی به کلمه "افت" تبدیل شده است که در اسناد ورزشی ممکن است به معنای افت تیم باشد و بهتر بود از ریشه افتاد به جای افت استفاده میکردیم

۵)

الف)

```
C:\Python38\python.exe "D:/College/Term 7/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [153, 410, 733, 767, 976, 994, 1008, 1073, 1331, 1577, 1729, 1743, 1753, 1769, 1772, 1787, 1830, 1831, 1847, 1898, 1938, 1944, 2025, 2030, 2037, 2041, 2110, 2131, 2181, 2217, 2263, 2277, 2283, 229
doc_id:  153
title:  توضیحات مسؤول مسابقات لیگ یک درباره شایعه سکته ناظر بازی
====================
doc_id:  410
title:  اعلام آخرین اقدامات تراکتور برای بازشدن پنجره/ احتمال استفاده از بازیکنان جدید برابر گل‌گهر
====================
doc_id:  733
title:  گزارش تمرین پرسپولیس| روحیه شاد قبل از مصاف با الهلال/ پا به توپ شدن گل‌محمدی و باقری
====================
doc_id:  767
title:  واکنش نصیرزاده به اظهارات مالک ماشین سازی: به جای فرافکنی به تعهداتتان عمل کنید
====================
doc_id:  976
title:  باج بن سلمان به انگلیس و قطر/ پروژه فوتبالی آل سعود چگونه رقم خورد؟
====================
doc_id:  994
title:  پنجره نقل و انتقالات باشگاه فولاد خوزستان باز شد
====================
doc_id:  1008
title:  معاون بین‌الملل و  مدیر کمیته حرفه ای سازی باشگاه استقلال منصوب شدند
====================
doc_id:  1073
title:  برای هتل محل اقامت AFC باشگاه پرسپولیس خواستار ارسال تاییدیه
====================
doc_id:  1331
title:  تائید حضور تماشاگر در دیدار تیم ملی فوتبال مقابل کره جنوبی توسط وزیر ورزش
====================
doc_id:  1577
title:  نامه پرسپولیس به کنفدراسیون فوتبال آسیا برای اسکان شاگردان گل محمدی در ریاض
```

ب)

```
C:\Python38\python.exe "D:/College/Term 7/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [1753, 1959, 2131, 2792, 2793, 7236], 2: [362, 390, 393, 552, 560, 724, 775, 978, 1007, 1146, 1220, 1249, 1702, 1704, 1710, 1716, 1721, 1724, 1725, 1728, 1737, 1751, 1753, 1759, 1760, 1765, 1778,
doc_id:  1753
title:  نامه جمعی از اساتید و متخصصان/ آقای رئیس‌جمهور در گام دوم انقلاب به داد «مدیریت» در کشور برسید
====================
doc_id:  1959
title:  نامه ۸ بسیج دانشجویی دانشگاه‌های تهران به معاون اول رئیس جمهور
====================
doc_id:  2131
title:  بزرگداشت شهدای مسجد قندوز در مقابل کنسولگری افغانستان/ آمریکا و آل سعود مقصران اصلی جنایت در افغانستان
====================
doc_id:  2792
title:  امروز محیط دانشگاه‌های ما عرصه دفاع مقدس است
====================
doc_id:  2793
title:  باید برای ثبت نقش دانشگاهیان در دوران دفاع مقدس کار تحقیقاتی صورت گیرد
====================
doc_id:  7236
title:  برگزاری یادبودی برای محمدسرور رجایی
====================
doc_id:  362
title:  دایی: می‌خواهم مردم مرا به عنوان انسان به یاد بیاورند نه دایی
====================
doc_id:  390
title:  تجلیل دانشگاه آزاد از  قهرمانان المپیک و کشتی‌گیران مدال‌آور+ تصاویر
====================
doc_id:  393
title:  یزدانی: فکر می‌کردم امتیاز فینال جهانی را به من می‌دهند/سبک حریف روسی تغییر کرده بود
====================
doc_id:  552
title:  میزبانان لیگ کشتی معرفی شدند/ زمان شروع مسابقات مشخص شد
====================
```

محمد توکلی

فاز اول پروژه

پ)

دانشگاه صنعتی امیرکبیر

```
C:\Python38\python.exe "D:/College/Term 7/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [1973, 2025, 2298, 2301, 2398, 2426, 2726, 2741, 2743, 2751, 2770, 2781, 2785, 2861, 2903, 3374, 4767, 4772, 5086, 6567, 6799, 6873, 6911, 7543], 2: [1749, 1772, 1802, 1972, 1973, 2025, 2038, 2053,
doc_id:  1973
title:  گزارش نظارت میدانی نمایندگان از مرزهای شمال‌غرب به کمیسیون امنیت ملی
====================
doc_id:  2025
title:  محسن اسلامی مدیرکل دفتر امور سیاسی وزارت کشور شد
====================
doc_id:  2298
title:  اعتراض هیات پارلمانی ایران به نماینده منصور هادی در اجلاس تغییرات آب و هوایی/ همکاری دولت مستعفی یمن با ائتلاف سعودی علیه یمنی‌ها
====================
doc_id:  2301
title:  هیئت پارلمانی ایران به رم سفر کرد
====================
doc_id:  2398
title:  نقض حاکمیت ملی عراق» با قراردادی 50ساله میان اقلیم و ترکیه»
====================
doc_id:  2426
title:  راهبرد دولت آیت‌الله رئیسی برای شکست تحریم‌ها ۳
====================
doc_id:  2726
title:  جان‌مایه سخنان رئیس‌جمهور
====================
doc_id:  2741
title:  آصفری: اظهارات رئیسی در سازمان ملل اقدامی برای رسوا کردن غربی ها و احقاق حقوق مردم بود
====================
doc_id:  2743
title:  راستینه: اظهارات آیت الله رئیسی مطالبه‌گری انقلابی در سازمان ملل بود/ تاکید رئیس جمهور بر منافع ملت
====================
doc_id:  2751
title:  منتظر منافع عضویت ایران در سازمان شانگهای با 40 درصد مساحت دنیا باشید/ توسعه همکاری‌های حمل و نقل، انرژی و بانکی
====================
```

محمد توکلی

فاز اول پروژه

سازمان ملل متحد

```
C:\Python38\python.exe "D:/College/Term 7/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [2793, 5021, 5022], 2: [362, 1725, 1753, 1779, 1959, 2091, 2131, 2447, 2588, 2631, 2709, 2791, 2793, 3031, 4854, 5021, 5022, 5154, 7043], 3: [311, 362, 390, 393, 480, 521, 535, 552, 560, 566, 724,
doc_id:  2793
title:  باید برای ثبت نقش دانشگاهیان در دوران دفاع مقدس کار تحقیقاتی صورت گیرد
====================
doc_id:  5021
title:  دفترچه راهنمای آزمون استخدامی دانشگاه‌ها برای بار چهارم اصلاح شد/تمدید مجدد مهلت ثبت‌نام
====================
doc_id:  5022
title:  دفترچه راهنمای آزمون استخدامی دانشگاه‌ها برای بار چهارم اصلاح شد/تمدید مجدد مهلت ثبت‌نام
====================
doc_id:  362
title:  دایی: می‌خواهم مردم مرا به عنوان انسان به یاد بیاورند نه دایی
====================
doc_id:  1725
title:  سیدمحسن دهنوی عضو هیئت امنای صندوق نوآوری و شکوفایی شد
====================
doc_id:  1753
title:  نامه جمعی از اساتید و متخصصان/ آقای رئیس‌جمهور در گام دوم انقلاب به داد «مدیریت» در کشور برسید
====================
doc_id:  1779
title:  وزیر علوم: علم و عقل دو بال دانایی است/ علم باید برای جامعه ثروت‌آفرین باشد
====================
doc_id:  1959
title:  نامه ۸ بسیج دانشجویی دانشگاه‌های تهران به معاون اول رئیس جمهور
====================
doc_id:  2091
title:  حجت‌الاسلام رستمی فقدان فعال دانشجویی دانشگاه شریف را تسلیت گفت
====================
doc_id:  2131
title:  بزرگداشت شهدای مسجد قندوز در مقابل کنسولگری افغانستان/ آمریکا و آل سعود مقصران اصلی جنایت در افغانستان
====================
```

جمهوری اسلامی ایران

```
C:\Python38\python.exe "D:/College/Term 7/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [80, 314, 617, 624, 844, 1331, 1601, 1704, 1706, 1708, 1732, 1743, 1748, 1749, 1753, 1763, 1769, 1776, 1787, 1798, 1803, 1804, 1809, 1812, 1820, 1837, 1842, 1854, 1906, 1909, 1910, 1913, 1926, 192
doc_id:  80
title:  و ماجرای قرارداد با شرکت اسرائیلی VAR واکنش تند فدراسیون فوتبال به حواشی
====================
doc_id:  314
title:  جوابیه باشگاه آلومینیوم نسبت به محرومیت 2 بازیکن ملی‌پوش
====================
doc_id:  617
title:  آمادگی قوه قضائیه برای حمایت حقوقی و معنوی از ورزشکاران/ اژه‌ای: به دعاوی حقوقی ورزشی به صورت تخصصی رسیدگی خواهد شد
====================
doc_id:  624
title:  اسکی صاحب کرسی جهانی شد
====================
doc_id:  844
title:  بیانیه فدراسیون تنیس درباره عدم صدور ویزای سه ملی پوش برای حضوردر دیویس کاپ بحرین
====================
doc_id:  1331
title:  تائید حضور تماشاگر در دیدار تیم ملی فوتبال مقابل کره جنوبی توسط وزیر ورزش
====================
doc_id:  1601
title:  تمجید فدراسیون جهانی هندبال از تاریخ‌سازی زنان ایرانی
====================
doc_id:  1704
title:  سرلشکر باقری: با وجود تمام فشارهای دشمن، جمهوری اسلامی روز به روز قوی‌تر شده است
====================
doc_id:  1706
title:  رئیس‌جمهور: اتحادیه اروپایی نباید تحت تأثیر فشارها و یکجانبه‌گرایی آمریکا قرار گیرد
====================
doc_id:  1708
title:  اروپا جوری عمل نکند که آمریکا، اروپا را حوزه نفوذ خود بداند
====================
```

ت)

```
C:\Python38\python.exe "D:/College/Term 7/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [632, 1367, 3615, 3664, 3878, 4056, 4188]}
doc_id:  632
title:  خیرخواه: برخی به دنبال فلج کردن ژیمناستیک هستند/ با بایکوت فدراسیون موفقیت‌ها بیشتر شد
====================
doc_id:  1367
title:  هشدار هیات ژیمناستیک تهران در خصوص سالن‌های مختلط و اقدامات غیراخلاقی
====================
doc_id:  3615
title:  دبیر مجمع فدراسیون ژیمناستیک مشخص شد
====================
doc_id:  3664
title:  ثبت نام ۱۳ نامزد برای پست ریاست فدراسیون ژیمناستیک + اسامی
====================
doc_id:  3878
title:  جزییات تعطیلی ورزش ایران تا پایان تیرماه+ تصویر
====================
doc_id:  4056
title:  دبیر: اگر من در مباحث فنی ۱۰ باشم، درستکار ۱۰۰ است/ بنا کاملا بر اساس چرخه انتخابی عمل کرد!
====================
doc_id:  4188
title:  جزییات تعطیلی‌های ورزش ایران تا ۹ مهر ۱۴۰۰/ کدام فعالیت‌های ورزشی در تهران ممنوع است؟
====================

Process finished with exit code 0
```

ث)

```
C:\Python38\python.exe "D:/College/Term 7/Informations Retrieval/Project/Phase 1/search_words.py"
{1: [4931, 5569, 5685, 5823, 5825, 5831, 5833, 5857, 6336], 2: [327, 374, 640, 782, 1021, 1032, 1034, 1075, 1083, 1106, 1146, 1160, 1337, 1378, 1445, 1556, 1579, 1703, 1737, 1752, 1754, 1759, 1767, 178
doc_id:  4931
title:  محموله ۱.۴ میلیون دوزی واکسن کرونا وارد کشور شد
====================
doc_id:  5569
title:  محموله ۱.۴ میلیون دوزی واکسن کرونا وارد کشور شد
====================
doc_id:  5685
title:  مهم‌ترین سلاح مبارزه با کرونا
====================
doc_id:  5823
title:  واکسیناسیون؛عقلانی‌ترین راه مقابله با کرونا/پرهیز از تزریق واکسن خارج از چرخه عمومی واکسیناسیون
====================
doc_id:  5825
title:  نکاتی که باید در مورد واکسیناسیون کرونا بدانیم
====================
doc_id:  5831
title:  واکسیناسیون؛عقلانی‌ترین راه مقابله با کرونا/پرهیز از تزریق واکسن خارج از چرخه عمومی واکسیناسیون
====================
doc_id:  5833
title:  نکاتی که باید در مورد واکسیناسیون کرونا بدانیم
====================
doc_id:  5857
title:  واکسن‌های کرونا با چه داروهایی تداخل دارند؟
====================
doc_id:  6336
title:  امکان ایجاد لخته خون در واکسن  آسترازنکا چقدر است؟
====================
doc_id:  327
title:  تارتار: 3 امتیاز باارزش در آبادان بدست آوردیم/نفت این فصل از سال گذشته بهتر است
====================
```