به نام یکدانه دردانه

سید عارف طباطبایی 9831040

فاز 1:

زیر بخش 1:

ابتدا با json.load داک‌های خبری را می‌خوانیم و آن‌ها را به صورت tuple های سه تایی شامل title، url و content در می‌آوریم.

در هر content، ابتدا punctuation ها را حذف می‌کنیم. سپس با استفاده از کتابخانه parsivar، متن مربوطه را ابتدا normalize و سپس tokenize می‌کنیم. سپس توکن‌ها را ریشه‌یابی می‌کنیم و در ادامه نیز stopwords را حذف می‌کنیم. برای مثال خواهیم داشت:

به ازای متن ورودی:

```
'!تقدیم به تو . امیدوارم امروز حالت خوب باشه ?سلام خوبی'
```

نتیجه:

```
'سلام', 'خوبی', 'امیدوار', 'امروز', 'حالت', 'خوب', 'باشه', 'تقدیم', 'تو'
```

زیر بخش 2:

ساختمان شاخص مکانی را با استفاده از تابع defaultdict در کتابخانه collections ایجاد می‌کنیم:

```
# Create the inverted index
index = defaultdict(lambda: {'num': 0, 'positions': defaultdict(lambda: [0, []])})
```

شامل num و positions که در اینجا num به تعداد بار استفاده از یک کلمه در تمام doc ها و positions به تعداد بار استفاده از آن کلمه در هر doc و موقعیت مکانی‌های آن‌ها اشاره دارند.

مطابق تکه کد زیر تمام doc ها را زیر و رو کرده و شاخص مکانی را می‌سازیم:

```
for i, doc in preprocessed_data.items():
    if int(i) % 1000 == 0:
        print(f'{int(i)} have been processed.')
    content = doc['content']
    for j, token in enumerate(content):
        index[token]['num'] += 1
        index[token]['positions'][i][0] += 1
        index[token]['positions'][i][1].append(j)
```

در انتها مطابق زیر برای مثال، شاخص مکانی را برای کلمه خبر چاپ می‌کنیم:

```
term = index['خبر']
print(f"Total frequency: {term['num']}")
print("Positions:")
for doc_id, positions in term['positions'].items():
    print(f"  Document {doc_id}: total_number: {positions[0]}, positions: {positions[1]}")
print()
```

نتیجه:

```
Total frequency: 1881
Positions:
  Document 14: total_number: 1, positions: [6]
  Document 19: total_number: 1, positions: [13]
  Document 29: total_number: 1, positions: [420]
  Document 48: total_number: 1, positions: [360]
  Document 56: total_number: 1, positions: [131]
  Document 78: total_number: 2, positions: [12, 25]
  Document 79: total_number: 1, positions: [55]
  Document 135: total_number: 1, positions: [26]
  Document 142: total_number: 1, positions: [838]
  Document 148: total_number: 2, positions: [320, 682]
  Document 150: total_number: 1, positions: [22]
  Document 200: total_number: 1, positions: [31]
  Document 203: total_number: 1, positions: [67]
  Document 206: total_number: 1, positions: [26]
  Document 211: total_number: 2, positions: [18, 61]
  Document 216: total_number: 2, positions: [124, 135]
  Document 218: total_number: 2, positions: [32, 93]
  Document 219: total_number: 1, positions: [80]
  Document 220: total_number: 1, positions: [94]
  Document 222: total_number: 1, positions: [25]
  Document 223: total_number: 2, positions: [62, 86]
  Document 226: total_number: 1, positions: [13]
```

```
  Document 236: total_number: 1, positions: [122]
  Document 238: total_number: 1, positions: [154]
  Document 259: total_number: 1, positions: [64]
  Document 300: total_number: 2, positions: [50, 55]
  Document 304: total_number: 2, positions: [150, 244]
  Document 306: total_number: 1, positions: [163]
  Document 314: total_number: 8, positions: [54, 58, 73, 77, 98, 154, 158, 172]
  Document 319: total_number: 1, positions: [198]
  Document 353: total_number: 1, positions: [33]
  Document 354: total_number: 1, positions: [338]
```

و....

زیر بخش 3:

پرسمان 1:

```
Query:  باشگاه های فوتسال آسیا
Terms:  ['فوتسال', 'آسیا', 'باشگاه']
Document: 0 in Title of:  اعلام زمان قرعه کشی جام باشگاه های فوتسال آسیا Relevance Score: 3
Document: 797 in Title of:  مدیرعامل پیکان: پتانسیل تهران برای میزبانی جام باشگاه‌ها بیشتر است/حافظیه در حد لیگ یک و دو هم نیست Relevance Score: 3
Document: 1306 in Title of:  تشریح جزئیات دریافت مجوز حرفه‌ای برای تیم‌های فوتسال و تشکیل کمیته ویژه در هفته‌های پایانی لیگ Relevance Score: 3
Document: 1535 in Title of:  چرا تیم ملی فوتسال به نایبلند اعزام نمی شود؟ Relevance Score: 3
Document: 1671 in Title of:  از فوتبال به فوتسال رسید/باشگاه‌های ایران به هوش باشند AFC مجوز حرفه‌ای Relevance Score: 3
```

برای مثال در عنوان داک شماره 0، می‌توان عبارت سرچ شده را مشاهده کرد.

پرسمان 2:

```
Query:  باشگاه های فوتسال ! آسیا
Terms:  ['فوتسال', 'باشگاه']
Forbidden Terms:  ['آسیا']
phase words:  {}
Document: 81 in Title of:  ماجدی: فوتبال کشور به تغییرات نیاز دارد Relevance Score: 2
Document: 426 in Title of:  اتفاق عجیب در لیگ نوجوانان/ قهرمانی سایپا پس گرفته شد!+نامه Relevance Score: 2
Document: 453 in Title of:  دبیر هیات فوتبال اصفهان: ابلاغی برای حضور تماشاگران نداشتیم/امیدوارم روزی مساوات رعایت شود Relevance Score: 2
Document: 467 in Title of:  سرمربی شهرخودرو جریمه شد/محرومیت برای بازیکن فوتسال Relevance Score: 2
Document: 493 in Title of:  اوج حساسیت در هفته آخر لیگ برتر فوتسال/ انصراف شهروند از بازی مقابل مس Relevance Score: 2
Document: 1339 in Title of:  فولاد در کشتی تبعداری می‌کند/گرشاسبی: در جذب بازیکن رکورد زده‌ایم Relevance Score: 2
Document: 1594 in Title of:  همه باشگاه‌ها ملزم به رعایت ابلاغیه AFC جایزی: مستند/مجوز زندی بتن را کمیته داوران تایید کرد Relevance Score: 2
Document: 1780 in Title of:  خوراکچی: داور دو گل صحیح ما را مردود اعلام کرد/کواپ مظلوم واقع شد Relevance Score: 2
Document: 2033 in Title of:  شکایت باشگاه گیتی پسند از سرمربی مس سونگون ورزقان Relevance Score: 2
Document: 2320 in Title of:  بازیکنان فوتسال مس برای راستی آزمایی تست کرونا به تهران می آیند/ مصاف گیتی پسند و مس 29 بهمن Relevance Score: 2
Document: 2391 in Title of:  عطابخش: به جز ۲-۳ بازیکن سایر نفراتمان می‌توانند در لیگ امید بازی کنند/ پیکان در بودجه مضیقه دارد Relevance Score: 2
Document: 2435 in Title of:  حضور مدیرعامل جدید در تمرین پیکان و وعده به شاگردان حسینی+عکس Relevance Score: 2
Document: 2453 in Title of:  تصمیمات سازمان لیگ فوتسال راهگشا یا بلای جان؟ Relevance Score: 2
Document: 2478 in Title of:  بیانیه مس سونگون در خصوص تصمیم سازمان لیگ فوتسال Relevance Score: 2
Document: 2583 in Title of:  اقدام ارزشمند گیتی پسند در پی مثبت شدن کرونای بازیکنان مس سونگون Relevance Score: 2
```

برای جلوگیری از نمایش داک های شامل کلمه آسیا، مقدار score آن را 1- می‌کنیم. برای مثال داک شماره 0 را در زیر مشاهده می‌کنید:

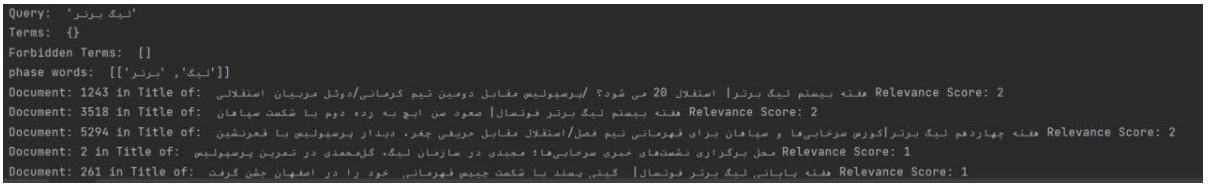
```
Document: 6763 in Title of:  جایزی: نامه ای درباره انصراف شهروند از حضور در لیگ برتر فوتسال ارسال نشده است Relevance Score: 1
Document: 9712 in Title of:  تاکید رئیس‌جمهور در تدوین بودجه ۱۴۰۱ بر عدالت بود/ عزم جدی دولت برای حل مشکلات Relevance Score: 1
Document: 0 in Title of:  اعلام زمان قرعه کشی جام باشگاه های فوتسال آسیا Relevance Score: -1
Document: 8 in Title of:  مدیر بی‌دستاورد، رئیس شد/ روزگار تلخ‌تر از تلخ برای والیبال ارومیه؟ Relevance Score: -1
Document: 29 in Title of:  مدیرعامل آلومینیوم: مرفاوی می‌داند هیچ باشگاهی با یک نفر صحبت نمی‌کند/استقلالی‌ها لیگ را گرفته‌اند Relevance Score: -1
Document: 50 in Title of:  !یک شاکی دیگر به والیبال ارومیه اضافه شد/مجرد: قراردادم را گُم، پشتم را خالی و بلاکم کردند Relevance Score: -1
Document: 54 in Title of:  اکبری: سیاست پیکان کسب قهرمانی نبود/شهداب یزد برای قهرمانی بسته شده بود Relevance Score: -1
```

پرسمان 3:

```
Query:  'سهمیه المپیک'
Terms:  {}
Forbidden Terms:  []
phase words:  [['المپیک ', 'سهمیه']]
Document: 3907 in Title of:  تیراندازان در کدام رویدادها المپیکی می‌شوند؟ Relevance Score: 1
```

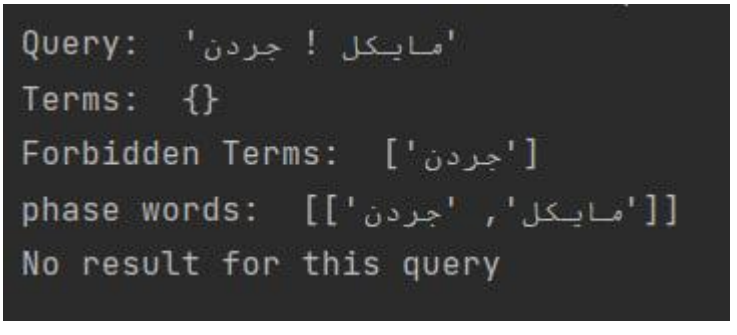
پرسمان 4:

در این پرسمان عبارت "طلای 'لیگ برتر' چغر" را سرچ کرده ام:



همان‌طور که مشاهده می‌کنید با اضافه کردن کلمه طلا به عبارت لیگ برتر، داک شماره 974 که قبلا امتیاز 1 داشت، نیز امتیاز 2 را بدست آورد. در ادامه با ممنوع کردن کلمه چغر، داک شماره 5294 از دور رقابت حذف گردید.

پرسمان 5:



در ادامه تعداد تکرار کلمات را نیز به صورت یک امتیاز مجزا در نظر می‌گیریم و اولویت را ابتدا دارا بودن تمام کلمات و سپس تکرار کلمات در نظر می‌گیریم:

برای مثال، برای کوئری " 'لیگ برتر' چغر" خواهیم داشت:

```
Query:  لیگ برتر' ! چغر'
Terms:  []
Forbidden Terms:  ['چغر']
phase words:  [['لیگ', 'برتر']]
Document: 1243 in Title of:  هفته بیستم لیگ برتر| استقلال 20 می شود؟ /پرسپولیس مقابل دومین تیم کرمانی/دوئل مربیان استقلالی Relevance Score: (2, 46)
Document: 3518 in Title of:  هفته بیستم لیگ برتر فوتسال| صعود سن ایج به رده دوم با شکست سپاهان Relevance Score: (2, 4)
Document: 5918 in Title of:  ماجرای تصمیم کبری در والیبال/سپاهان از بین دو گزینه، سومی را انتخاب کرد Relevance Score: (1, 37)
Document: 1634 in Title of:  هفته نوزدهم لیگ برتر| استقلال بی‌دفاع مقابل فجر؛ دوئل قلعه‌نویی با پرسپولیس و دربی اصفهان Relevance Score: (1, 27)
Document: 4624 in Title of:  نکته از لیگ برتر در نیم فصل اول| پرسپولیس و 5 تیم رکورددار حاشیه/ زمین سفت نویدکیا و لطف قلعه‌نویی به استقلال 16 Relevance Score: (1, 24)
Document: 6673 in Title of:  !پاسخ مدیر نظارت و بازرسی سازمان لیگ؛ آقای اصولی چه آورده ای برای فوتبال خراسان و پدیده داشتی؟ Relevance Score: (1, 24)
Document: 820 in Title of:  هفته بیست و یکم لیگ برتر|منصوریان به دنبال عافیت بخیری با پرسپولیس/تقابل استقلال و قلعه‌نویی بدون بازیکن گابنی Relevance Score: (1, 23)
Document: 5632 in Title of:  هفته سیزدهم لیگ برتر فوتبال| تداوم شکست ناپذیری استقلال در اراک؟/ نبرد قرمزپوشان تهران و تبریز Relevance Score: (1, 23)
Document: 5992 in Title of:  هفته دوازدهم لیگ برتر|سپاهان - پرسپولیس؛ نبرد آسیایی در زمین بیطرف/استقلال مقابل مانع فولادی و دربی کرمان Relevance Score: (1, 23)
```