

به نام خدا



تحلیل داده های حجیم

تمرین سری اول

گزارش سوال سوم

استاد: دکتر غلامپور

دانشجو: سجاد هاشم بیکی

پاییز 1401

-2، 1 :

ابتدا فایل را میخوانیم و سپس با استفاده از تابع `json.loads` محتوا را تبدیل به متون فارسی میکنیم و کار پیش پردازش را شروع میکنیم.

یک دیکشنری شامل کاراکترهایی که میخواهیم حذف کنیم درست کرده و سپس هر کاراکتر را با `space` جاگذاری میکنیم.

داده های پاکسازی شده را در لیست ذخیره و سپس به `rdd` تبدیل میکنیم.

-3:

در قسمت قبل یک `rdd` ساخته شد که شامل مقالات به صورت `splited` شده بود. این `rdd` را با `flatMap` به یک `rdd` که شامل کلمات است درمیآوریم. (`words_rdd`)

-4:

برای شمارش کلمات هر کلمه را به صورت یک جفت `key value` درمیآوریم و سپس با دستور `reduce by key` جفت های مشابه را جمع میکنیم و تعداد کلمات به دست میآید.

و سپس `words_count_rdd` را میسازیم.

## 5-

در این بخش جفت های ساخته شده را با دستور `countbykey` می شماریم و سپس با دستور `sorted` به صورت نزولی مرتب میکنیم و در لیست `sorted_words_list` ذخیره میکنیم.

صد عضو اول این لیست همان `100 most common words` میباشد . که در لیست `list_of_top_100_comm_words` ذخیره شده اند.

► (1) Spark Jobs

```
با', '119', 'های', '14233', 'را', '14844', 'این', '16385', 'می', '21615', 'که', '23115', 'است', '27259', 'از', '38498', 'به', '35330', 'در', '47732', 'در', '51562', 'او', '504', 'خود', '5082', 'کرد', '5384', 'برای', '6237', 'بود', '6322', 'شد', '6983', 'ها', '7781', 'سال', '8262', 'خود', '8890', 'ان', '9297', 'یک', '10170', '93', 'نام', '2901', 'خزان', '2904', 'واقع', '3172', 'ایران', '3340', 'کند', '3454', 'تا', '3608', 'یا', '3801', 'خود', '3817', 'بر', '3957', 'ای', '4272', 'دارد', '4504', '2', 'پس', '2336', 'نو', '2356', 'کرده', '2409', 'نفر', '2420', 'فیلم', '2456', 'باند', '2540', 'فوق', '2697', 'نیز', '2740', 'وی', '2792', 'جمعیت', '2809', 'غزار', '2870', 'بین', '18', 'نمش', '1949', 'هم', '1956', 'و', 'نوسط', '1990', 'باشگاه', '2006', 'اغل', '2015', 'بازی', '2038', 'یکی', '2057', 'بود', '2073', 'شیر', '2109', 'استفاده', '2285', '90', 'هلی', '1625', 'کند', '1644', 'اند', '1685', 'استان', '1702', 'کنور', '1778', 'شهرستان', '1779', 'دیگر', '1797', 'خوان', '1805', 'اما', '1806', 'همچین', '1835', '1621', '133', 'رچود', '1415', 'آمریکا', '1502', 'اشاره', '1530', 'هستد', '1536', 'کار', '1542', 'ملی', '1564', 'مورد', '1584', 'ا', '1586', 'هر', '1587', 'درکت', '1621', 'د', 'گروه', '1257', 'زاده', '1258', 'میلادی', '1272', 'روی', '1277', 'خوند', '1279', 'صورت', '1294', 'داد', '1306', 'متحد', '1313', 'ایالات', '1330', 'تاریخ', '1333', '6', '1252', 'بی', '1104', 'زیر', '1110', 'روستا', '1145', 'دانشگاه', '1153', 'نیم', '1180', 'زمان', '1180', 'داده', '1181', 'نست', '1197', 'داشت', '1215', 'دهد', '1234', 'بان', '1252', '1079', 'پیش', '1070', 'سخت', '1075', 'سه', '1079']
```

Command took 10.51 seconds -- by sajjadhashembeigi18@gmail.com at 12/10/2022, 11:29:54 PM on My Cluster

## 6-

در قسمت قبل بخش قابل توجهی از کلمات پرتکرار، معنی خاصی به مقالات نمیدهند و آنها را به `stopwords` اضافه میکنیم تا بعدا آنها را از مقالات حذف کنیم. میتوان همه کلمات پرتکرار یا بخشی از آنها را `stopwords` در نظر گرفت. که من در اینجا بخشی از آنها را در نظر گرفتم.

با', 'او', 'انرا', 'او', 'به', 'از', 'است', 'که', 'می', 'این', 'را', 'های', 'یا', 'یک', 'آن', 'شده', 'سال', 'ها', 'خود', 'برای', 'کرد', 'خود', 'ان', 'دارد', 'ای', 'انرا', 'خود', 'یا', 'انرا', 'اگر', 'انرا', 'ایران']

Command took 0.87 seconds -- by sajjadhashembeigi18@gmail.com at 12/10/2022, 11:45:58 PM on My Cluster

7-:

در این بخش حروف ربط و اضافه را (stopwords) با دستور filter از مقالات حذف میکنیم و سپس articles\_without\_stopwords\_rdd را میسازیم.

8,9-:

در قسمت 4 تعداد هر کلمه را به دست آوردیم. در این قسمت با دستور filter کلماتی که کمتر از 20 بار تکرار شده اند را پیدا میکنیم و لیست کلمات غیررایج را (uncommon\_words\_list) میسازیم

ریاضت' و گفتند' و مقبضان' و نهمتهای' و کانت' و فراغتی' و اه' و 'dharma' و برآز' و واماندگی' و بپرخاند' و کواثری' و سم' و بشریت' و یابند' و مقصدات' و ادم' و پنجگانه' و خوریم' و اینکرد' و احتیاط' [  
نقدانی' و صفات' و یارونده' و اخونده' و یزد' و گری' و گریگ' و رامنی' و جسمانی' و گردم' و الوتر' و ییتانده' و گلیسمای' و '۳۵۰۰۰' و مورمین' و زانورده' و '۱۰۳' و منظمی' و دانشات' و اسکندیاری' و کا  
سایها' و رسیم' و منشاء' و پیشگمان' و '12' و 'اصنیماها' و '1969' و 'مأموریت' و 'التمسین' و 'دارده' و 'محاسبه' و 'ایرونیسمیک' و 'هوازی' و 'هواپوشی' و 'کثرلی' و 'استیکری' و 'مستثنی' و '۱۷۱ کیلومتر' و  
آلزیلیان' و 'لیجان' و 'لیلیا' و 'وناز' و 'مروج' و '۷۵۴' و '۵۴' و 'مختصات' و 'محمور' و 'جوانشیر' و 'مسند' و 'خراج' و 'انقراض' و 'Adharbāyagān' و 'نکو' و 'گرم' و 'ریشه' و 'دگیلی' و 'II/1' و '۲۲۸' و  
ش' و '۱۹۱۷' و 'نکارها' و 'خدا' و 'اورینوها' و 'چاندان' و 'نزارها' و 'فراپاغ' و 'طلیله' و '۱۹۲۸-۱۹۵۳' و 'ایاز' و 'لیلی' و 'عینی' و '۱۳۷۲' و 'فصلیه' و 'اعتبار' و 'مساعت' و 'پیار' و 'کلم' و '۲۰۳۸۲۰۰۰' و 'گا  
ومیش' و 'سای' و 'پرسن' و 'پرزهای' و 'عربی' و 'روسانی' و 'کتری' و '۸۰۲۳۸۰۶۷۲' و '۱۹۰۶' و '۱۰۸' و 'پیماد' و 'اصانده' و 'هفتای' و 'گژیوران' و 'ارویلی' و 'کرش' و '۳۶۴۰۰۰۰' و 'ارستان' و 'ارقام' و 'براندگین  
یورویژن' و 'پزگها' و 'گرافرو' و 'Contest' و '۲' و '۱۰۰۹۲۲' و 'ظیمان' و 'اورویوران' و 'ضلعی' و 'قیل' و 'قسم' و 'قری' و 'اصلی' و 'کردهای' و 'زیرفشار' و 'کلیبر' و '۱۲۰۰۰۰' و '۴۴۰۹۱' و 'عدا' و 'حق' و '۹۸'  
ن' و 'آتششانی' و 'خیز' و 'جی' و '۸' و 'جی' و '۱۷۶۷' و 'برآید' و 'وشتاک' و 'تخلیگران' و 'تیپی' و 'خن' و 'القم' و 'منی' و '۲۶۰۰۰' و 'صانبی' و 'اسکورا' و 'یفتگی' و 'وسیلی' و 'پستندار' و '۷۲/۴۴' و 'افقل' و '۲'  
مراج' و 'مسیدانه' و 'هوجو' و 'جدال' و 'انزوی' و 'قلم' و 'نورده' و '۱۸۵۴' و 'کوتیار' و 'بنام' و 'درهای' و 'مجلله' و '۱۹۳۹' و 'پستند' و 'اُم' و 'مختمین' و 'ویرا' و 'Tale' و '۳' و 'پیلو' و 'ندرا' و 'آزانه  
نی' و '۲۲۰۰۰۰' و 'مغرب' و 'دولهای' و 'فرآوری' و 'ایسان' و '۲/۵' و '۱/۵' و 'فکونمندی' و 'مفارج' و 'بهای' و '۲/۲۵' و '۲/۲۴' و 'مطمعی' و 'اجومون' و 'خاخرانه' و 'خرک' و 'رنگارنگ' و 'پیکار' و 'هانی' و 'سکویی' و 'ا  
لیکور' و 'گازوس' و 'آمازگر' و 'هانی' و 'استادهای' و 'پروکتریا' و 'رویده' و 'پوروازی' و 'مفکرانی' و 'ایبوف' و 'یتکی' و 'قزای' و 'چلریت' و 'بلقود' و 'مفکران' و 'اگر' و 'Socius' و 'زستکی' و 'قلب' و 'افشرا  
و' و 'نمشتن' و 'پرونیست' و 'مالاکسا' و 'پازش' و 'مضمم' و 'سدیکا' و 'وابلی' و 'سدیکاهاست' و 'درآوردن' و 'اندیشده' و 'کول' و 'کرخان' و 'برآمدن' و 'سالگرد' و 'هتارگر' و 'انوارد' و 'اللقاب' و 'فرهمنترین' و 'جری  
و' و 'صیویستی' و 'ارانیگلیسم' و 'نمیش' و 'تاترگزارترین' و 'دریاریان' و 'همراستا' و 'معتان' و '۱۹۱۸-۱۹۱۹' و 'دوستانان' و '۱۹۳۹-۱۹۳۹' و 'میلده' و 'احلال' و 'رفاهی' و 'پیکتا' و 'پیکستا' و 'مسنی' و 'کنفراسی' و 'سومر  
زا' و '۱۹۷۹-۱۹۷۹' و 'پازوژی' و 'میتان' و 'دردهای' و 'العتا' و 'مارگات' و 'اعظمیه' و 'استادهای' و 'ایده' و 'موفیتی' و 'کیتیلیستی' و 'بلنست' و 'کله' و 'پوشین' و 'مکرز' و 'مطلق' و 'فرایت' و 'هتکانش' و 'ایب  
تند' و 'میری' و 'لیکن' و 'مطلق' و 'میتون' و 'کتر' و 'یادهی' و 'حوم' و 'مزله' و 'های' و 'بازارها' و 'بنتانی' و 'پازردان' و 'مالینی' و 'جادی' و 'مولونخوانی' و 'ازار' و 'مهمتر' و 'بدرجای' و 'اموان' و  
اریکس' و 'ازان' و 'رایجین' و 'سرس' و 'طشی' و 'مخورد' و 'فغن' و 'کوبین' و 'وفا' و 'سرایت' و 'کامتا' و 'جیلین' و 'گاست' و 'آن' و 'پوشیر' و 'خوزستان' و 'مها' و 'نگ' و 'گده' و 'اصلی' و 'ملازید' و 'وضیبا' و 'نگ' و  
بستک' و 'اسنفا' و '۱۸۱۶' و '21' و 'Mole' و 'خلیست' و 'رسوخ' و 'گفته' و 'اکلاهایان' و 'پوستان' و 'نواختد' و 'نواختگی' و 'سوردا' و 'انلیجی' و 'ویشترین' و 'اخترشاس' و 'رصدی' و 'قصت' و 'طور' و '۱۴۰'  
کرده' و 'کهرود' و 'افراان' و 'املاخ' و 'اچ' و 'دستگرد' و '۱۳۱۷' و 'اندیشده' و 'ازده' و 'نظمیه' و '۱۲۲۷' و 'میداری' و 'هزیر' و 'مصراع' و 'پوسله' و 'یلگان' و 'اهواز' و 'اسنان' و 'افجان' و 'خمن' و 'پروچرد' و 'اد  
نرکان' و 'اورالین' و 'سیاهان' و 'رضخان' و 'دیسامانی' و 'بختاران' و 'دانشمینی' و 'پیمان' و '۳۳۶' و 'چالده' و 'کوری' و 'مسند' و 'اشیان' و 'درآیفتد' و 'پیرمزان' و 'لهج' و 'جلفای' و 'کوجانده' و 'پیشه' و 'ایرالکر  
' و 'بالایهای' و 'پیمانگری' و 'وگاز' و 'سدا' و 'نکرن' و 'مطایب' و 'چهرسازی' و 'توانیشی' و 'یاسفگی' و '۱۳۲۲' و '۱۲۲۲' و '۱۹۲۵' و '۲۰۰۰' و 'اقاسی' و 'ایتکلیف' و 'پازردانده' و 'خشد' و 'مجاورتی' و 'افراای  
حله' و 'محم' و 'شیرسنگ' و '۱۳۴۷' و 'ضرب' و 'میرحفری' و '۱۲۰۴۴۴' و 'الافعات' و 'مطمعی' و 'گنم' و 'زوبزه' و 'ممر' و 'مشک' و 'مسگر' و 'کلا' و 'شروه' و 'دوغ' و 'آفر' و 'حویو' و 'شقه' و 'شاه

10-:

در این قسمت بایستی کلمات غیررایج را که در قسمت قبل به دست آوردیم، از مقالات حذف کنیم.

میتوان به دو صورت عمل کرد. یا همه کلمات غیررایج را حذف کرد ک ما در این گزارش و کد موجود در نوتبوک همین کار را کردیم، یا اینکه 100 تا از کلمات غیررایج را حذف کرد که کد آن به صورت کامنت شده داخل نوتبوک قرار دارد.



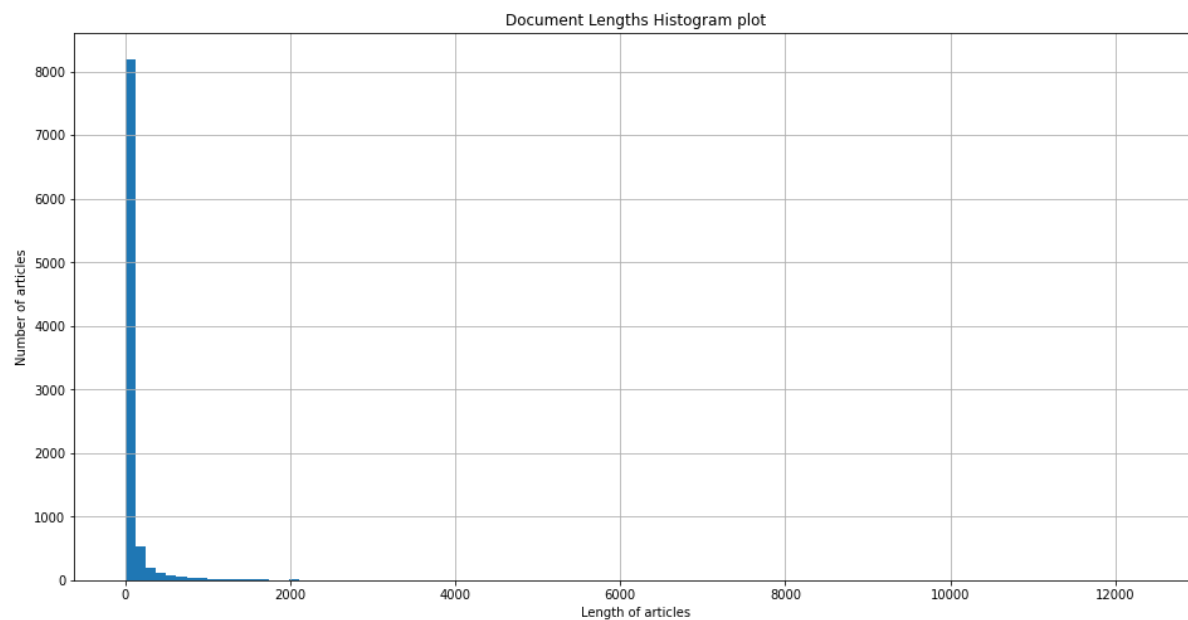
2:

در این قسمت، ابتدا تمام کلمات سه حرفی انگلیسی را بدست میاوریم و سپس با توجه به تعداد تکرار هر کلمه آنها را مرتب میکنیم. 20 کلمه اول لیست مطلوب ماست. که در زیر آمده است.

['The', 'the', 'and', 'CAD', 'for', 'DNA', 'GSI', 'NFA', 'III', 'San', 'IVB', 'AST', 'int', 'MBA', 'SIP', 'SQL', 'USS', 'bes', 'del', 'EUV']

3-:

در این قسمت طول هر مقاله (تیترو متن) را بدست آورده و سپس هستوگرام آن را رسم میکنیم که در زیر آمده است.



:4-

طول مقالات را مرتب کرده و 5 تای اول آنها که بیشترین طول را دارند بدست میاوریم.  
تیترا و لینک این 5 مقاله بدین صورت میباشد:

[ 'حسن بن علی' و 'ولایتیور لنین' و 'اراک' و 'سوسیالیسم' و 'آرمون نورینگ' ]

[ 'https://fa.wikipedia.org/wiki?curid=19165', 'https://fa.wikipedia.org/wiki?curid=18927', 'https://fa.wikipedia.org/wiki?curid=5967', 'https://fa.wikipedia.org/wiki?curid=4059', 'https://fa.wikipedia.org/wiki?curid=9002' ]

:5-

## درصد و تعداد مقالاتی که شامل 6 موضوع گفته شده میباشد بدین صورت میباشد:

[('مهندسی', '80'), ('اقتصاد', '191'), ('حقوق', '151'), ('پزشکی', '125'), ('سیاست', '282'), ('تاریخ', '976')]

[('مهندسی', '2.0364644418381492'), ('حقوق', '1.6099797419767568'), ('پزشکی', '1.3327646870668515'), ('سیاست', '3.006717134022817'), ('تاریخ', '10.406226676617976'), ('اقتصاد', '0.8529693997227851')]



## TF-IDF + Searching

1-:

در این قسمت بدست میاوریم که هر کلمه در چند مقاله بکار رفته است (در واقع پارامتر  $n$  در فرمول  $tfidf$ ) و در نهایت  $words\_df\_rdd$  را میسازیم. 10 نمونه از آن بدین شکل است:

```
Out[231]: [('نوعی', 158),  
( 'بجرحاند', 2 ),  
( 'پروانه', 26 ),  
( 'برآر', 1 ),  
( 'موارد', 356 ),  
( 'تفاوت', 105 ),  
( 'واماندگی', 1 ),  
( 'توسط', 958 ),  
( 'درمی', 29 ),  
( 'گیرد', 320 )]
```

2:

در این قسمت پارامتر  $IDF$  را برای هر کلمه بدست میاوریم که نتیجه آن بدین شکل است:

```
Out[271]: [('1.781', 'نوعی'),
            ('3.671', 'بچرخاند'),
            ('2.558', 'پروانه'),
            ('3.972', 'برآر'),
            ('1.437', 'موارد'),
            ('1.956', 'تفاوت'),
            ('3.972', 'واماندگی'),
            ('1.033', 'توسط'),
            ('2.511', 'درمی'),
            ('1.482', 'گیرد')]
```

3:

در این قسمت مقدار TF را برای هر کلمه در هر مقاله بدست میاوریم که بدین صورت میباشد:

```
Out[317]: [(0, ('0.004761904761904762', 'اتوجایزو')),
            (0, ('0.004761904761904762', 'جایرویلن')),
            (0, ('0.009523809523809525', 'ظاهر')),
            (0, ('0.009523809523809525', 'موتوری')),
            (0, ('0.009523809523809525', 'بال')),
            (0, ('0.004761904761904762', 'ظاهری')),
            (0, ('0.009523809523809525', 'ولی')),
            (0, ('0.004761904761904762', 'نام')),
            (0, ('0.004761904761904762', 'اختراع')),
            (0, ('0.004761904761904762', 'هواپیما'))]
```