به نام خدا

تحلیل داده های حجیم

تمرین سری اول

گزارش سوال سوم

استاد: دکتر غلامپور

دانشجو:سجاد هاشم بیکی

پاییز1401

-1،2 :

ابتدا فایل را میخوانیم و سپس با استفاده از تابعjson.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 درمیاوریم.(**words_rdd**)

-4:

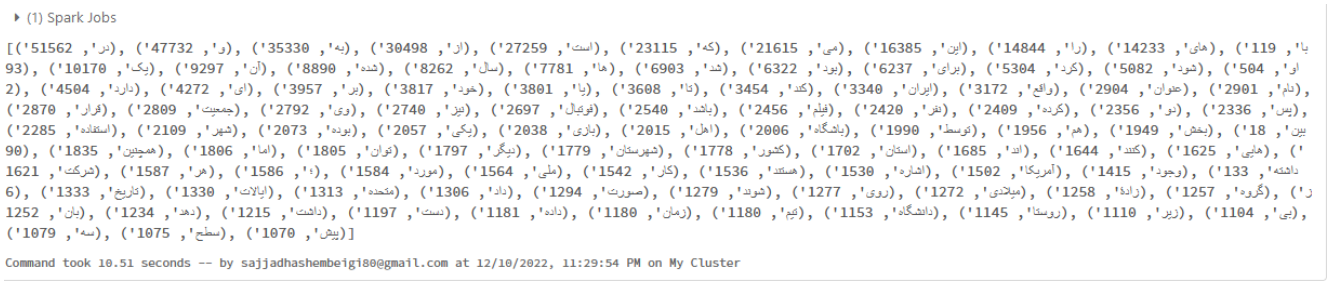
برای شمارش کلمات هر کلمه را به صورت یک جفت key value درمیاوریم و سپس با دستور reduce by key جفت های مشابه را جمع میکنیم و تعداد کلمات به دست میاید.

و سپس **words_count_rdd** را میسازیم.

:5-

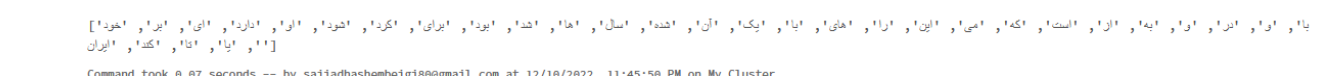
دراین بخش جفت های ساخته شده را با دستور countbykey میشماریم و سپس با دستور sorted به صورت نزولی مرتب میکنیم و در لیست sorted_words_list ذخیره میکنیم.

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



:6-

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



: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 اوردیم. دراین قسمت با دستورfilter کلماتی که کمتر از 20 بار تکرار شده اند را پیدا میکنیم و لیست کلمات غیررایج را میسازیم(**uncommon_words_list**)

:10-

در این قسمت بایستی کلمات غیررایج را که در قسمت قبل به دست اوردیم،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 ان به صورت کامنت شده داخل نوتبوک قرار دارد.

:11-

در این قسمت تیتر و متن تمیز شده هرمقاله را که از قبل بدست اوریم،به همراه لینک هرمقاله داخل یک دیشکنری قرار میدهیم.در نهایت لیستی از دیکشنری ها داریم.که این لیست را تبدیل به articles_cleaned_rdd میکنیم.

-Exploration

:1-

در این قسمت تعداد کلمات را بعد از تمیز کردن بدست میاوریم.سپس کلماتی که سه حرف دارند و حروف انها در بازه ی کد اسکی مربوط به حروف فارسی قرار دارند را پیدا میکنیم و لیست (unique_3_Letter_words_fa) را میسازیم. که طول لیست همان تعداد کلمات سه حرفی و متمایز فارسی موجود در متن مقالات میباشد.

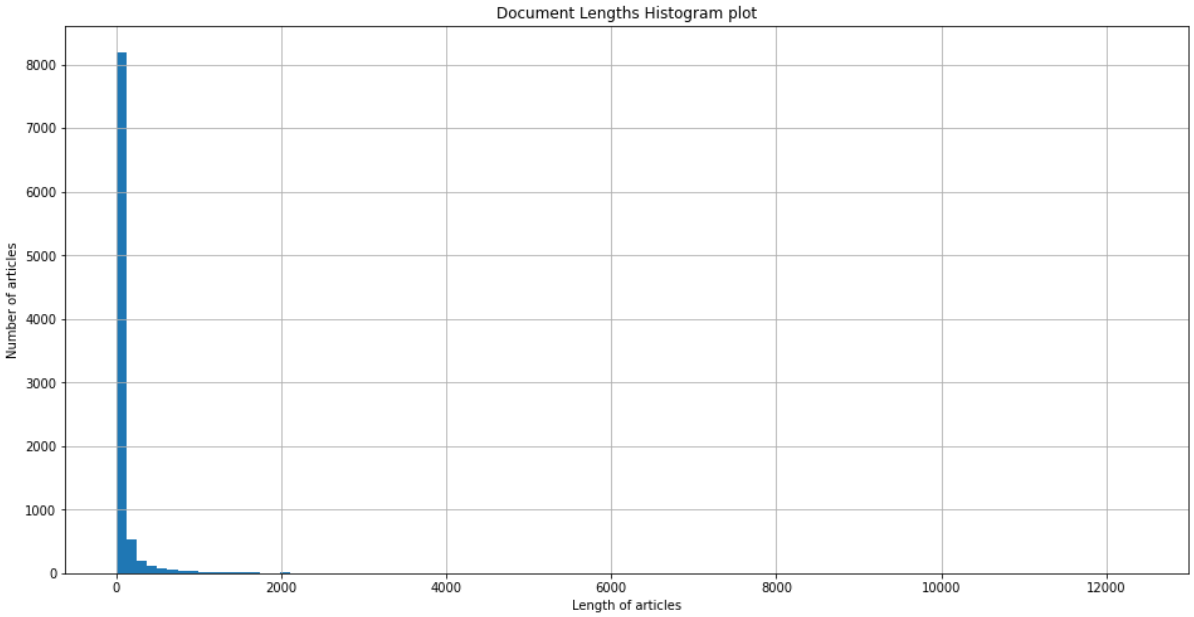
:2

در این قسمت،ابتدا تمام کلمات سه حرفی انگلیسی را بدست میاوریم و سپس با توجه به تعداد تکرار هر کلمه انها را مرتب میکنیم.20 کلمه اول لیست مطلوب ماست.که در زیر امده 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(تیتر و متن) را بدست اورده و سپس هستوگرام ان را رسم میکنیم که درزیر امده است.



:4-

طول مقالات را مرتب کرده و 5 تای اول انها که بیشترین طول را دارند بدست میاوریم.

تیتر و لینک این 5 مقاله بدین صورت میباشد:

```
['حسین بن علی', 'ولادیمیر لنین', 'اراک', 'سوسیالیسم', 'آزمون توریدگ']
```

```
['https://fa.wikipedia.org/wiki?curid=19165', 'https://fa.wikipedia.org/wiki?curid=10927', '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') ,('اقتصاد', '2.0364644418381492') ,('حقوق', '1.6099797419767568') ,('پزشکی', '1.3327646870668515') ,('سیاست', '3.006717134022817') ,('تاریخ', '10.406226676617976')]
0.8529693997227851 ,')
```

TF-IDF + Searching

:1-

در این قسمت بدست میاوریم که هر کلمه در چند مقاله بکار رفته است(در واقع پارامترn در فرمول tfifd) و درنهایت words_df_rdd را میسازیم.10 نمونه از ان بدین شکل است:

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

:2

در این قسمت پارامتر IDF را برای هر کلمه بدست میاوریم که نتیجه ان 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 ','وامادگی'),
 ('1.033 ','توسط'),
 ('2.511 ','نرمی'),
 ('1.482 ','گیرد')]
```

:3

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

```
Out[317]: [(0, ('0.004761904761904762 ','اتوجایرو')),
 (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 ','دام')),
 (0, ('0.004761904761904762 ','اختراع')),
 (0, ('0.004761904761904762 ','هواپیما'))]
```