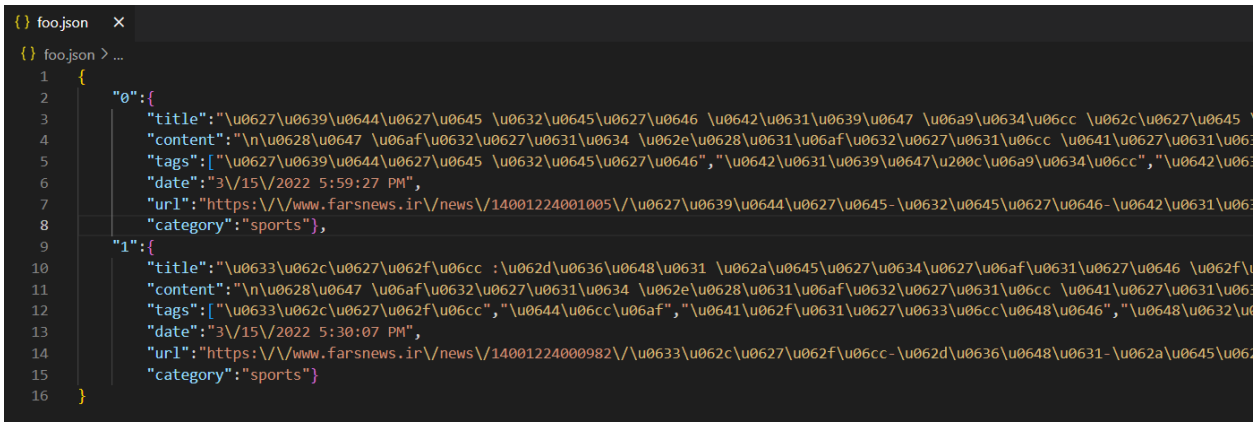
امیرعلی بلباسی ۹۸۳۱۱۰۹

دو خبر اول را در فایل دیگری برای تست آوردم:



فایل کلمات پرتکرار که در کتابخانه استفاده شده (hazm) موجود است:

In [1]:

```
%pip install hazm

from __future__ import unicode_literals
from hazm import *
import codecs
import json
```

```
Requirement already satisfied: hazm in d:\applications\anaconda\lib\site-
packages (0.7.0)
Requirement already satisfied: nltk==3.3 in
d:\applications\anaconda\lib\site-packages (from hazm) (3.3)
Requirement already satisfied: six in d:\applications\anaconda\lib\site-
packages (from nltk==3.3->hazm) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

In [2]:

```
f = open('foo.json')
#f = open('IR_data_news_12k.json')
data = json.load(f)
f.close()

contents = []
for i in data:
    tmp = data[i]['content']
    contents.append(tmp[0:len(tmp)-16])
print('#1')
for i in contents:
    print(i)

normalizer = Normalizer()
for i in range(len(contents)):
    contents[i] = normalizer.normalize(contents[i])
print('#2')
for i in contents:
    print(i)

tokens = []
for i in range(len(contents)):
    tokens.append(word_tokenize(contents[i]))
print('#3')
for i in tokens:
    print(i)
    print()

lemmatizer = Lemmatizer()
for i in range(len(tokens)):
```

```
    for j in range(len(tokens[i])):
        tokens[i][j] = lemmatizer.lemmatize(tokens[i][j])
print('#4')
for i in tokens:
    print(i)
    print()

"""
stemmer = Stemmer()
for i in range(len(tokens)):
    for j in range(len(tokens[i])):
        tokens[i][j] = stemmer.stem(tokens[i][j])
print('#5')
for i in tokens:
    print(i)
    print()
"""

f = codecs.open('stopwords.dat', encoding='utf-8')
stopWords = []
for l in f.readlines():
    stopWords.append(l.strip('\n'))
f.close()
print('A#')
print(stopWords)
print()

for i in range(len(stopWords)):
    stopWords[i] = normalizer.normalize(stopWords[i])
print('B#')
print(stopWords)
print()

for i in range(len(stopWords)):
    stopWords[i] = lemmatizer.lemmatize(stopWords[i])
print('C#')
print(stopWords)
print()

tmpTokens = []
for i in range(len(tokens)):
    tmpTokensDoc = []
    for j in range(len(tokens[i])):
        if(tokens[i][j] not in stopWords): tmpTokensDoc.append(tokens[i][j])
    tmpTokens.append(tmpTokensDoc)
```

```
tokens = tmpTokens.copy()
print('#5')
for i in tokens:
    print(i)
    print()
#1
```

در نامه ای رسمی به (AFC) به گزارش خبرگزاری فارس، کنفدراسیون فوتبال آسیا فدراسیون فوتبال ایران و باشگاه گیتی پسند زمان قرعه کشی جام باشگاه های فوتسال آسیا را رسماً اعلام کرد. بر این اساس ۲۵ فروردین ماه ۱۴۰۱ مراسم قرعه کشی جام باشگاه های فوتسال آسیا در مالزی برگزار می شود. باشگاه گیتی پسند بعنوان قهرمان فوتسال ایران در سال ۱۴۰۰ به این مسابقات راه پیدا کرده است. پیش از این گیتی پسند تجربه ۳ دوره حضور در جام باشگاه های فوتسال آسیا را داشته که هر سه دوره به فینال مسابقات راه پیدا کرده و یک عنوان قهرمانی و دو مقام دومی بدست آورده است.

به گزارش خبرگزاری فارس، سید حمید سجادی در حاشیه مراسم گرامیداشت روز جوان در جمع خبرنگاران در رابطه با عرضه سهام سرخابی‌ها در بورس اظهار داشت: منتظر طی روند هستیم و بعدا اطلاع رسانی خواهیم کرد. وی در مورد حضور تماشاگران در مسابقات فوتبال اظهار داشت: حضور تماشاگران در لیگ برتر فوتبال تابع نظر فدراسیون ،.سازمان لیگ و ستاد ملی مبارزه با کرونا است
#2

در نامه‌ای رسمی به (AFC) به گزارش خبرگزاری فارس، کنفدراسیون فوتبال آسیا فدراسیون فوتبال ایران و باشگاه گیتی پسند زمان قرعه کشی جام باشگاه‌های فوتسال آسیا را رسما اعلام کرد. بر این اساس ۲۵ فروردین ماه ۱۴۰۱ مراسم قرعه کشی جام باشگاه‌های فوتسال آسیا در مالزی برگزار می‌شود. باشگاه گیتی پسند بعنوان قهرمان فوتسال ایران در سال ۱۴۰۰ به این مسابقات راه پیدا کرده است. پیش از این گیتی پسند تجربه ۳ دوره حضور در جام باشگاه‌های فوتسال آسیا را داشته که هر سه دوره به فینال مسابقات راه پیدا کرده و یک عنوان قهرمانی و دو مقام دومی بدست آورده است.

به گزارش خبرگزاری فارس، سید حمید سجادی در حاشیه مراسم گرامیداشت روز جوان در جمع خبرنگاران در رابطه با عرضه سهام سرخابی‌ها در بورس اظهار داشت: منتظر طی روند هستیم و بعدا اطلاع رسانی خواهیم کرد. وی در مورد حضور تماشاگران در مسابقات فوتبال اظهار داشت: حضور تماشاگران در لیگ برتر فوتبال تابع نظر فدراسیون، سازمان لیگ و ستاد ملی مبارزه با کرونا است.
#3

```
['به', 'گزارش', 'خبرگزاری', '،', 'فارس', 'کنفدراسیون', 'فوتبال', 'آسیا', '(',
'AFC', ')', 'در', 'ای\u200cنامه', 'رسمی', 'به', 'فدراسیون', 'فوتبال',
'ایران', 'و', 'باشگاه', 'گیتی', 'پسند', 'زمان', 'قرعه', 'کشی', 'جام',
'بر', '.', 'کرد', 'اعلام', 'رسما', 'را', 'آسیا', 'فوتسال', 'های\u200cباشگاه',
'جام', 'کشی', 'قرعه', 'مراسم', '۱۴۰۱', 'ماه', 'فروردین', '۲۵', 'اساس', 'این',
'شود\u200cمی', 'برگزار', 'مالزی', 'در', 'آسیا', 'فوتسال', 'های\u200cباشگاه',
'در', 'ایران', 'فوتسال', 'قهرمان', 'بعنوان', 'پسند', 'گیتی', 'باشگاه', '.',
'پیش', '.', 'کرده_است', 'پیدا', 'راه', 'مسابقات', 'این', 'به', '۱۴۰۰', 'سال',
```

```
'از', 'این', 'گیتی', 'پسند', 'تجربه', '۳', 'دوره', 'حضور', 'در', 'جام',
'دوره', 'سه', 'هر', 'که', 'داشته', 'را', 'آسیا', 'فوتسال', 'های\u200cباشگاه',
'به', 'فینال', 'مسابقات', 'راه', 'پیدا', 'کرده', 'و', 'یک', 'عنوان',
'قهرمانی', 'و', 'دو', 'مقام', 'دومی', 'بدست', 'آورده_است', '.']

['به', 'گزارش', 'خبرگزاری', 'فارس', '،', 'سید', 'حمید', 'سجادی', 'در',
'در', 'خبرنگاران', 'جمع', 'در', 'جوان', 'روز', 'گرامیداشت', 'مراسم', 'حاشیه',
'رابطه', 'با', 'عرضه', 'سهام', 'های\u200cسرخابی', 'در', 'بورس', 'اظهار',
'داشت', ':', 'منتظر', 'طی', 'روند', 'هستیم', 'و', 'ابعدا', 'اطلاع',
'خواهیم_کرد', '.', 'وی', 'در', 'مورد', 'حضور', 'در', 'تماشاگران',
'فوتبال', 'اظهار', 'داشت', ':', 'حضور', 'تماشاگران', 'در', 'لیگ', 'برتر',
'فوتبال', 'تابع', 'نظر', 'فدراسیون', '،', 'سازمان', 'لیگ', 'و', 'ستاد',
'ملی', 'مبارزه', 'با', 'کرونا', 'است', '.']

#4
['به', 'گزارش', 'خبرگزاری', 'فارس', '،', 'کنفدراسیون', 'فوتبال', 'آسیا', '(',
'AFC', ')', 'در', 'نامه', 'رسم', 'به', 'فدراسیون', 'فوتبال', 'ایران', 'و',
'باشگاه', 'جام', 'کشید#کش', 'قرعه', 'زمان', 'پسند', 'گیتی', 'باشگاه',
'اساس', 'این', 'بر', '.', 'کرد#کن', 'اعلام', 'رسما', 'را', 'آسیا', 'فوتسال',
'باشگاه', 'جام', 'کشید#کش', 'قرعه', 'مراسم', '۱۴۰۱', 'ماه', 'فروردین', '۲۵',
'گیتی', 'باشگاه', '.', 'شد#شو', 'برگزار', 'مالزی', 'در', 'آسیا', 'فوتسال',
'به', '۱۴۰۰', 'سال', 'در', 'ایران', 'فوتسال', 'قهرمان', 'بعنوان', 'پسند',
'گیتی', 'این', 'از', 'پیش', '.', 'کرد#کن', 'پیدا', 'راه', 'مسابقات', 'این',
'فوتسال', 'باشگاه', 'جام', 'در', 'حضور', 'دوره', '۳', 'تجربه', 'پسند',
'مسابقات', 'فینال', 'به', 'دوره', 'سه', 'هر', 'که', 'داشته', 'را', 'آسیا',
'راه', 'مقام', 'دو', 'و', 'قهرمان', 'عنوان', 'یک', 'و', 'کرده', 'پیدا',
'دوم', 'بدست', 'آورد#آور', '.']

['به', 'گزارش', 'خبرگزاری', 'فارس', '،', 'سید', 'حمید', 'سجاد', 'در',
'در', 'خبرنگار', 'جمع', 'در', 'جوان', 'روز', 'گرامیداشت', 'مراسم', 'حاشیه',
'رابطه', 'با', 'عرضه', 'سهام', 'های\u200cسرخابی', 'در', 'بورس', 'اظهار',
'داشت', ':', 'منتظر', 'طی', 'روند', 'هست#', 'و', 'ابعدا', 'اطلاع',
'رسان', 'رساند#کن', 'کرد#کن', '.', 'وی', 'در', 'مورد', 'حضور', 'تماشاگر', 'در',
'مسابقات', 'فوتبال', 'اظهار', 'داشت', ':', 'حضور', 'تماشاگر', 'در', 'لیگ',
'برتر', 'فوتبال', 'تابع', 'نظر', 'فدراسیون', '،', 'سازمان', 'لیگ', 'و',
'ستاد', 'ملی', 'مبارزه', 'با', 'کرونا', 'است#', '.']

A#
['در', 'به', 'از', 'که', 'این', 'را', 'با', 'است', 'برای', 'آن', 'یک',
'خود', 'تا', 'کرد', 'بر', 'هم', 'نیز', 'گفت', 'می\u200cشود', 'وی', 'شد',
'ما', 'اما', 'یا', 'شده', 'باید', 'هر', 'آنها', 'بود', 'او', 'دیگر',
'مورد', 'می\u200cکند', 'شود', 'کند', 'وجود', 'بین', 'پیش', 'است_شده',
'پس', 'نظر', 'اگر', 'همه', 'یکی', 'حال', 'من', 'هستند', 'کنند', 'نیست',
'چه', 'بی', 'می', 'بخش', 'می\u200cکنند', 'همین', 'افزود', 'هایی',
'داشت', 'سه', 'بسیار', 'بیشتر', 'داد', 'روی', 'همچنین', 'راه', 'دارند',
```

'ولی', 'جدید', 'بعد', 'شدن', 'اینکه', 'میان', 'هیچ', 'تنها', 'سوی', 'چند',
'کرده_است', 'نه', 'اول', 'دهد\u200cمی', 'کردند', 'برخی', 'کردن', 'حتی',
'بار', 'درباره', 'تمام', 'افراد', 'طور', 'چنین', 'شما', 'بیش', 'نسبت',
'طی', 'بزرگ', 'دوم', 'ندارد', 'چون', 'کرده', 'تواند\u200cمی', 'بسیاری',
'خواهد_شد', 'دیگری', 'گوید\u200cمی', 'آنان', 'البته', 'بدون', 'همان', 'حدود',
'قبل', 'ویژه', 'کل', 'وارد', 'توان\u200cمی', 'رشد', 'یعنی', 'قابل', 'کنیم',
'چرا', 'بوده_است', 'سازی', 'لازم', 'هنوز', 'گذاری', 'نیاز', 'براساس',
'پیدا', 'تغییر', 'حالی', 'جای', 'کم', 'گرفت', 'وقتی', 'شوند\u200cمی',
'رو', 'بیان', 'آیا', 'تعداد', 'زیادی', 'فقط', 'مدت', 'باعث', 'تحت', 'اکنون',
'برابر', 'دهد', 'جاری', 'بلکه', 'نوع', 'بودن', 'کرده_اند', 'عدم', 'شدند',
'آقای', 'خصوص', 'شاید', 'گیری', 'زیر', 'امر', 'مربوط', 'اخیر', 'بوده', 'مهم',
'اثر', 'کنید', 'سایر', 'سوم', 'اولین', 'کنار', 'فکر', 'بودند', 'کننده',
'ضمن', 'مثل', 'پی', 'دارای', 'حل', 'ممکن', 'گیرد\u200cمی', 'باز', 'مانند',
'رسد\u200cمی', 'آنچه', 'امکان', 'طول', 'موجب', 'کسی', 'منظور', 'دور', 'اجرا',
'تعیین', 'گفته', 'شوند', 'جمع', 'خیلی', 'علاوه', 'گونه', 'تاکنون', 'رسید',
'سال', 'گرفته', 'شده_اند', 'چهار', 'علت', 'داشته_باشد', 'خواهد_بود', 'طرف',
'تهیه', 'تبدیل', 'مناسب', 'زیرا', 'مشخص', 'می\u200cتوانند', 'نزدیک', 'جریان',
'روند', 'بنابراین', 'دهند\u200cمی', 'یافت', 'نخستین', 'بالا', 'پنج', 'ریزی',
'عالی', 'چیزی', 'نخست', 'بیشتری', 'ترتیب', 'شده_بود', 'خاص', 'خوبی', 'خوب',
'شروع', 'فرد', 'کامل', 'غیر', 'رود\u200cمی', 'دهند', 'آخرین', 'دادن', 'جدی',
'بهترین', 'شامل', 'گیرد', 'بخشی', 'باشند', 'تمامی', 'بهتر', 'داده_است', 'حد',
'نبود', 'کسانی', 'کرد\u200cمی', 'داریم', 'علیه', 'باشد\u200cمی', 'دانست',
'ناشی', 'داشتند', 'دهه', 'شد\u200cمی', 'ایشان', 'آنجا', 'گرفته_است', 'دچار',
'می\u200cآید', 'لحاظ', 'آنکه', 'داده', 'بعضی', 'هستیم', 'اند', 'برداری',
'نباید', 'می\u200cکنیم', 'نشست', 'سهم', 'همیشه', 'آمد', 'اش', 'وگو',
'می\u200cکنم', 'حداقل', 'طبق', 'جا', 'خواهد_کرد', 'نوعی', 'چگونه', 'رفت',
'هنگام', 'فوق', 'روش', 'ندارند', 'سعی', 'بندی', 'کلی', 'شمار', 'کافی',
'آورد', 'پشت', 'چیز', 'داشته_است', 'کوچک', 'سمت', 'زیاد', 'همچنان', 'مواجه',
'حالا', 'روبه', 'سال\u200cهای', 'دادند', 'می\u200cکردند', 'عهده', 'نیمه',
'جایی', 'دیگران', 'سی', 'بروز', 'یکدیگر', 'آمده_است', 'جز', 'کنم', 'سپس',
'کنندگان', 'خودش', 'همواره', 'یافته', 'شان', 'صرف', 'نمی\u200cشود', 'رسیدن',
'چهارم', 'یابد', 'متر', 'ساز', 'داشته', 'کرده_بود', 'باره', 'نحوه', 'کردم',
'شخصی', 'داشته_باشند', 'پخش', 'محسوب', 'کمی', 'متفاوت', 'سراسر',
'تو', 'داشتن', 'نظیر', 'آمده', 'گروهی', 'فردی', 'ع', 'همچون', 'خطر',
'خویش', 'کدام', 'دسته', 'سبب', 'عین', 'آوری', 'متاسفانه', 'بیرون', 'دار',
'ابتدا', 'شش', 'افرادی', 'می\u200cگویند', 'سالهای', 'درون', 'نیستند',
'پر', 'یافته_است', 'خاطرنشان', 'گاه', 'جمعی', 'اغلب', 'دوباره',
'می\u200cیابد', 'لذا', 'زاده', 'گردد', 'اینجا']

B#

['و', 'در', 'به', 'از', 'که', 'این', 'را', 'با', 'است', 'برای', 'آن', 'یک',
'خود', 'تا', 'کرد', 'بر', 'هم', 'نیز', 'گفت', 'می\u200cشود', 'وی', 'شد',
'دارد', 'ما', 'اما', 'یا', 'شده', 'هر', 'باید', 'آنها', 'او', 'بود', 'دیگر',
'مورد', 'می\u200cکند', 'شود', 'کند', 'وجود', 'بین', 'پیش', 'شده_است',
'پس', 'نظر', 'اگر', 'همه', 'یکی', 'حال', 'من', 'هستند', 'کنند', 'نیست',

'هایی', 'افزود', 'همین', 'کنندu200c\می', 'بخش', 'می', 'بی', 'چه', 'باشد',
'داشت', 'سه', 'بسیار', 'بیشتر', 'داد', 'روی', 'همچنین', 'راه', 'دارند',
'ولی', 'جدید', 'بعد', 'شدن', 'اینکه', 'میان', 'هیچ', 'تنها', 'سوی', 'چند',
'کرده_است', 'نه', 'اول', 'دهدu200c\می', 'کردند', 'برخی', 'کردن', 'حتی',
'بار', 'درباره', 'تمام', 'افراد', 'طور', 'چنین', 'شما', 'بیش', 'نسبت',
'طی', 'بزرگ', 'دوم', 'ندارد', 'چون', 'کرده', 'تواندu200c\می', 'بسیاری',
'خواهد_شد', 'دیگری', 'گویدu200c\می', 'آنان', 'البته', 'بدون', 'همان', 'حدود',
'قبل', 'ویژه', 'کل', 'وارد', 'توانu200c\می', 'رشد', 'یعنی', 'قابل', 'کنیم',
'چرا', 'بوده_است', 'سازی', 'لازم', 'هنوز', 'گذاری', 'نیاز', 'براساس',
'پیدا', 'تغییر', 'حالی', 'جای', 'کم', 'گرفت', 'وقتی', 'شوندu200c\می',
'رو', 'بیان', 'آیا', 'تعداد', 'زیادی', 'فقط', 'مدت', 'باعث', 'تحت', 'اکنون',
'برابر', 'دهد', 'جاری', 'بلکه', 'نوع', 'بودن', 'کرده_اند', 'عدم', 'شدند',
'آقای', 'خصوص', 'شاید', 'گیری', 'زیر', 'امر', 'مربوط', 'اخیر', 'بوده', 'مهم',
'کنید', 'سایر', 'سوم', 'اولین', 'کنار', 'فکر', 'بودند', 'کننده', 'اثر',
'مثل', 'پی', 'دارای', 'حل', 'ممکن', 'گیردu200c\می', 'باز', 'مانند', 'ضمن',
'آنچه', 'امکان', 'طول', 'موجب', 'کسی', 'منظور', 'دور', 'اجرا', 'رسدu200c\می',
'رسید', 'تاکنون', 'گونه', 'علاوه', 'خیلی', 'جمع', 'شوند', 'گفته', 'تعیین',
'طرف', 'خواهد_بود', 'داشته_باشد', 'چهار', 'علت', 'شده_اند', 'گرفته', 'ساله',
'جریان', 'نزدیک', 'توانندu200c\می', 'مشخص', 'زیرا', 'مناسب', 'تبدیل', 'تهیه',
'ریزی', 'پنج', 'بالا', 'نخستین', 'یافت', 'دهندu200c\می', 'بنابراین', 'روند',
'خوب', 'خوبی', 'خاص', 'شده_بود', 'ترتیب', 'بیشتری', 'نخست', 'چیزی', 'عالی',
'جدی', 'دادن', 'آخرین', 'دهند', 'رودu200c\می', 'غیر', 'کامل', 'فرد', 'شروع',
'حد', 'داده_است', 'بهتر', 'تمامی', 'باشند', 'بخشی', 'گیرد', 'شامل', 'بهترین',
'دانست', 'باشدu200c\می', 'علیه', 'داریم', 'کردu200c\می', 'کسانی', 'نبود',
'دچار', 'گرفته_است', 'آنجا', 'ایشان', 'شدu200c\می', 'دهه', 'داشتند', 'ناشی',
'برداری', 'اند', 'هستیم', 'بعضی', 'داده', 'آنکه', 'لحاظ', 'آیدu200c\می',
'وگو', 'اش', 'آمد', 'همیشه', 'سهم', 'نشست', 'کنیمu200c\می', 'نباید',
'رفت', 'چگونه', 'نوعی', 'خواهد_کرد', 'جا', 'طبق', 'حداقل', 'کنمu200c\می',
'کافی', 'کلی', 'شمار', 'بندی', 'سعی', 'ندارند', 'روش', 'فوق', 'هنگام',
'آورد', 'پشت', 'چیز', 'داشته_است', 'کوچک', 'سمت', 'زیاد', 'همچنان', 'مواجه',
'نیمه', 'عهده', 'کردندu200c\می', 'دادند', 'هایu200cسال', 'روبه', 'حالا',
'سپس', 'کنم', 'جز', 'آمده_است', 'یکدیگر', 'بروز', 'سی', 'دیگران', 'جایی',
'رسیدن', 'شودu200cنمی', 'صرف', 'شان', 'یافته', 'همواره', 'خودش', 'کنندگان',
'کردم', 'نحوه', 'باره', 'کرده_بود', 'داشته', 'ساز', 'متر', 'یابد', 'چهارم',
'سراسر', 'متفاوت', 'کمی', 'پخش', 'محسوب', 'داشته_باشند', 'شخصی', 'تو',
'خطر', 'همچون', 'ع', 'فردی', 'گروهی', 'آمده', 'نظیر', 'داشتن', 'کاملا',
'دار', 'بیرون', 'متاسفانه', 'آوری', 'عین', 'سبب', 'دسته', 'کدام', 'خویش',
'نیستند', 'درون', 'سالهای', 'گویندu200c\می', 'افرادی', 'شش', 'ابتدا',
'دوباره', 'اغلب', 'جمعی', 'گاه', 'خاطرنشان', 'پر', 'یافته_است',
'اینجا', 'گردد', 'زاده', 'لذا', 'یابدu200c\می']

C#

['و', 'در', 'به', 'از', 'که', 'این', 'را', 'با', 'است#', 'برای', 'آن', 'یک',
'شد#شو', 'وی', 'شد#شو', 'گفت#گو', 'نیز', 'هم', 'بر', 'کرد#کن', 'تا', 'خود',
'او', 'بود#باش', 'آن', 'هر', 'باید', 'شده', 'یا', 'اما', 'ما', 'داشت#دار',

'پیش', 'بین', 'وجود', 'کند', 'شد#شو', 'کرد#کن', 'مورد', 'دو', 'دیگر',
'کرد#کن', 'من', 'هست#', 'حال', 'یک', 'همه', 'اگر', 'نظر', 'پس', 'شد#شو',
'افزود#افزا', 'همین', 'کرد#کن', 'بخش', 'می', 'بی', 'چه', 'بود#باش', 'نیست',
'سه', 'بسیار', 'بیشتر', 'داد', 'روی', 'همچنین', 'راه', 'داشت#دار', 'هایی',
'جدید', 'بعد', 'شدن', 'اینکه', 'میان', 'هیچ', 'تنها', 'سو', 'چند', 'داشت',
'کرد#کن', 'نه', 'اول', 'داد#ده', 'کرد#کن', 'برخی', 'کردن', 'حتی', 'ولی',
'بار', 'درباره', 'تمام', 'افراد', 'طور', 'چنین', 'شما', 'بیش', 'نسبت',
'طی', 'بزرگ', 'دوم', 'داشت#دار', 'چون', 'کرده', 'توانست#توان', 'بسیار',
'کرد#کن', 'شد#شو', 'دیگر', 'گفت#گو', 'آن', 'البته', 'بدون', 'همان', 'حدود',
'براساس', 'قبل', 'ویژه', 'کل', 'وارد', 'توان\u200cمی', 'رشد', 'یعنی', 'قابل',
'شد#شو', 'چرا', 'بود#باش', 'سازی', 'لازم', 'هنوز', 'گذاشت#گذار', 'نیاز',
'تحت', 'اکنون', 'پیدا', 'تغییر', 'حال', 'جای', 'کم', 'گرفت#گیر', 'وقت',
'عدم', 'شد#شو', 'رو', 'بیان', 'آیا', 'تعداد', 'زیاد', 'فقط', 'مدت', 'باعث',
'بوده', 'مهم', 'برابر', 'داد#ده', 'جاری', 'بلکه', 'نوع', 'بودن', 'کرد#کن',
'اثر', 'آقا', 'خصوص', 'شاید', 'گرفت#گیر', 'زیر', 'امر', 'مربوط', 'اخیر',
'ضمن', 'کرد#کن', 'سایر', 'سوم', 'اولین', 'کنار', 'فکر', 'بود#باش', 'کننده',
'رسید#رس', 'مثل', 'پی', 'دارا', 'حل', 'ممکن', 'گرفت#گیر', 'باز', 'مانند',
'تعیین', 'آنچه', 'امکان', 'طول', 'موجب', 'کس', 'منظور', 'دور', 'اجرا',
'ساله', 'رسید', 'تاکنون', 'گونه', 'علاوه', 'خیلی', 'جمع', 'شد#شو', 'گفته',
'تهیه', 'طرف', 'بود#باش', 'داشت#دار', 'چهار', 'علت', 'شد#شو', 'گرفته',
'روند', 'جریان', 'نزدیک', 'توانست#توان', 'مشخص', 'زیرا', 'مناسب', 'تبدیل',
'عالی', 'ریخت#ریز', 'پنج', 'بالا', 'نخستین', 'یافت#یاب', 'داد#ده', 'بنابراین',
'شروع', 'خوب', 'خوبی', 'خاص', 'شد#شو', 'ترتیب', 'بیشتر', 'نخست', 'چیز',
'بهترین', 'جدی', 'دادن', 'آخرین', 'داد#ده', 'رفت#رو', 'غیر', 'کامل', 'فرد',
'حد', 'داد#ده', 'بهتر', 'تمام', 'بود#باش', 'بخشی', 'گرفت#گیر', 'شامل',
'دانست#دان', 'بود#باش', 'علیه', 'داشت#دار', 'کرد#کن', 'کسانی', 'بود#باش',
'دچار', 'گرفت#گیر', 'آنجا', 'ایشان', 'شد#شو', 'دهه', 'داشت#دار', 'ناشی',
'نباید', 'بردار', 'اند', 'هست#', 'بعضی', 'داده', 'آنکه', 'لحاظ', 'آمد#آ',
'حداقل', 'کرد#کن', 'وگو', 'اش', 'آمد#آ', 'همیشه', 'سهم', 'نشست', 'کرد#کن',
'روش', 'فوق', 'هنگام', 'رفت#رو', 'چگونه', 'نوع', 'کرد#کن', 'جا', 'طبق',
'همچنان', 'مواجه', 'کافی', 'کلی', 'شمار', 'بست#بند', 'سعی', 'داشت#دار',
'روبه', 'حالا', 'آورد#آور', 'پشت', 'چیز', 'داشت#دار', 'کوچک', 'سمت', 'زیاد',
'بروز', 'سی', 'دیگر', 'جا', 'نیمه', 'عهده', 'کرد#کن', 'داد#ده', 'سال',
'همواره', 'خودش', 'کنندگان', 'سپس', 'کرد#کن', 'جز', 'آمد#آ', 'یکدیگر',
'ساز', 'متر', 'یافت#یاب', 'چهارم', 'رسیدن', 'شد#شو', 'صرف', 'شان', 'یافته',
'داشت#دار', 'شخص', 'تو', 'کرد#کن', 'نحوه', 'باره', 'کرد#کن', 'داشته',
'آمده', 'نظیر', 'داشتن', 'کاملا', 'سراسر', 'متفاوت', 'کم', 'پخش', 'محسوب',
'عین', 'سبب', 'دسته', 'کدام', 'خویش', 'خطر', 'همچون', 'ع', 'فرد', 'گروهی',
'گفت#گو', 'افراد', 'شش', 'ابتدا', 'دار', 'بیرون', 'متاسفانه', 'آورد#آور',
'اغلب', 'جمع', 'گاه', 'خاطرنشان', 'پر', 'یافت#یاب', 'هست#', 'درون', 'سال',
'اینجا', 'گشت#گرد', 'زاده', 'لذا', 'یافت#یاب', 'دوباره']

#5
['(', 'آسیا', 'فوتبال', 'کنفدراسیون', '،', 'فارس', 'خبرگزاری', 'گزارش',
'گیتی', 'باشگاه', 'ایران', 'فوتبال', 'فدراسیون', 'رسم', 'نامه', ')', 'AFC',

'رسما', 'آسیا', 'فوتسال', 'باشگاه', 'جام', 'کشید#کش', 'قرعه', 'زمان', 'پسند',
'قرعه', 'مراسم', '۱۴۰۱', 'ماه', 'فروردین', '۲۵', 'اساس', '.', 'اعلام',
'کشید#کش', 'جام', 'باشگاه', 'فوتسال', 'آسیا', 'مالزی', 'برگزار', '.',
'باشگاه', 'گیتی', 'پسند', 'بعنوان', 'قهرمان', 'فوتسال', 'ایران', '۱۴۰۰',
'مسابقات', '.', 'گیتی', 'پسند', 'تجربه', '۳', 'دوره', 'حضور', 'جام',
'باشگاه', 'فوتسال', 'آسیا', 'دوره', 'فینال', 'مسابقات', 'عنوان', 'قهرمان',
'مقام', 'بدست', '.']

['گزارش', 'خبرگزاری', 'فارس', '،', 'سید', 'حمید', 'سجاد', 'حاشیه', 'مراسم',
'گرامیداشت', 'روز', 'جوان', 'خبرنگار', 'رابطه', 'عرضه', 'سهام',
'سرخابی\u200cها', 'بورس', 'اظهار', ':', 'منتظر', 'بعدا', 'اطلاع',
'رسان#رسانه', '.', 'حضور', 'تماشاگر', 'مسابقات', 'فوتبال', 'اظهار', ':',
'حضور', 'تماشاگر', 'لیگ', 'برتر', 'فوتبال', 'تابع', 'فدراسیون', '،',
'سازمان', 'لیگ', 'ستاد', 'ملی', 'مبارزه', 'کرونا', '.']

همانطور که در مثال مشاهده می‌کنید در مرحله اول فایل تست را لود کردیم سپس محتوا اخبار را استخراج کرده و پس از آن متن محتوا را نرمال می‌کنیم (در هر مرحله هم چاپ می‌کنیم تا تغییرات مشخص شود. مثلا با نرمال‌سازی space های بین "ها" و "خود کلمه" تبدیل به نیم‌فاصله می‌شود). در مرحله بعدی توکن‌ها را اسختراج کرده و در یک لیست از لیست‌ها ذخیره می‌کنیم. در مرحله بعد ریشه کلمات را جایگزین فرم‌های مختلف آن می‌کنیم (مثلا "باشگاه‌های" تبدیل به "باشگاه" می‌شود). سپس همین مراحل را روی stop word ها تکرار می‌کنیم تا مشابه هم شوند. سرآخر نیز با حلقه روی لیست توکن‌ها، آن توکن‌هایی که جزو کلمات پر تکرار هستند را حذف می‌کنیم (مانند توکن "در").

دلیل پردازش نرمال‌سازی یکدست کردن کل متون است زیرا مثلا ممکن است نویسنده در بعضی جاها نیم‌فاصله را رعایت کرده باشد و در بعضی جاها نکرده باشد. دلیل استخراج توکن‌ها هم که واضح است زیرا برای ساخت شاخص مکانی لازم اند. بعضی کلمات در جاهای مختلف یک یا چند سند با اشکال متفاوت ظاهر می‌شوند ولی همه آنها به یک مفهوم اشاره می‌کنند و یک ریشه مشترک دارند. به منظور یک جستجو دقیق‌تر، لازم است تمام این اشکال به یک شکل واحد در آورده شوند. به همین جهت پردازش ریشه‌یابی را انجام می‌دهیم. شرکت دادن کلمات پرتکرار فارسی مانند "از" در عملیات جستجو و بازیابی اسناد ، دقت ما را به مراتب پایین می‌آورد (چون این کلمات در عموم اسناد وجود دارند). به همین جهت با پیش‌پردازش، این دست کلمات را از لیست توکن‌های استخراج شده حذف می‌کنیم.