سوال اول) در مرحله پیش پردازش داده ها ۴ گام اصلی طی شده است .

Initial stages of text processing

- Tokenization
 - Cut character sequence into word tokens
 - Deal with "John's", a state-of-the-art solution
- Normalization
 - Map text and query term to same form
 - You want U.S.A. and USA to match
- Stemming
 - We may wish different forms of a root to match
 - authorize, authorization
- Stop words
 - We may omit very common words (or not)
 - the, a, to, of

به ترتیب ۱ - نرمال سازی ۲ - استخراج توکن ۳-حذف کلمات پرتکرار(stop word) ۴-ریشه یابی انجام شده است. از کتابخانه hazm و parsivar برای پیش پردازش ها استفاده می کنیم .

در شکل زیر نحوه استفاده و فراخوانی کتابخانه ها برای پیش پردازش اسناد آمده است .

```
import pandas as pd
from hazm import Normalizer
from hazm import stopwords_list
from parsivar import FindStems
from parsivar import Tokenizer

class Preprocessing:

def __init__(self):...

# normalizes text before extracting tokens
def getTokens(self, text: str):
    return self.__tokenizer.tokenize_words(self.normalize(text))

def normalize(self, text):
    return self.__normalizer.normalize(text)

def getStems(self, words):
    return [self.__stemmer.convert_to_stem(word) for word in words]

def deleteStopWords(self, words):
    return [word for word in words if word not in self.__stopWords]
```

1-نرمال سازی: کلمات در یک قالب استاندارد نوشته می شوند و به هنجار می شوند. در این مرحله شکل های مختلف یک کلمه به یک شکل یکسان تبدیل می شوند. متون از لحاظ فاصله، نیم فاصله، تاریخ ها و ... یکسان می شوند. مثلا کلمه قرآن یا قرآن به یک شکل قرآن تبدیل می شود. یا euni-codeهای مختلفی که برای حرف 'ک' وجود دارد به یک شکل تبدیل می شود. برای نرمال سازی از کتابخانه hazm استفاده شده است. دلیل انجام نرمال سازی این است که پردازش ها را آسان تر می کند و پاسخگویی به پرسمان ها با دقت بهتری انجام می شود

```
from PreprocessingUtils import Preprocessing

preprocessing = Preprocessing()

print(preprocessing.normalize('عمر من من من من المناده المنادة المناده المناده المناده المناده المناده المناده المناده المنادة المناده المناده المناده المنادة المنادة
```

Y-استخراج توکن: یکی از قسمت های اصلی ساخت موتور های بازیابی اطلاعات ؛ ساختن توکن هاست . در واقع به وسیله توکن های می توانیم لیست شاخص های مکانی را بسازیم و توسط آن به کوئری کاربر پاسخ دهیم. برای استخراج توکن از کتابخانه parsivar استفاده می کنیم .

```
from PreprocessingUtils import Preprocessing

preprocessing = Preprocessing()

print(preprocessing.getTokens('عدد انگلیسی 13 به فارسی تبدیل می شود'))

test 

/home/mohammad/Documents/uni/term8/information-retrieval/project/search-er

['شود∪2000عدد', 'انگلیسی', '۱۳', 'به', 'فارسی', 'تبدیل', 'می']
```

۳-حذف کلمات پرتکرار: ایست واژه ها غالبا کلماتی هستند که خیلی در جست و جو به کار نمی آیند و بار معنایی کمی دارند. کلماتی مثل با ، تا ، و ، که و ... که خیلی هم پرتکرار هستند را حذف میکنیم . این کار باعث می شود شاخص معکوس ساخته شده حجم کمتری را اشغال کند . می دانیم مدیریت حافظه یکی از بحث های چالشی در ساخت موتور بازیابی اطلاعات است. البته در موتورهای جست و جوی پیشرفته همه ی ایست واژه ها را حذف نمی کنند بلکه به آن ها وزن می دهند.

برای حذف ایست واژه ها از لیست کلمات کتابخانه hazm استفاده شده است .

```
from PreprocessingUtils import Preprocessing

preprocessing = Preprocessing()

print(preprocessing.deleteStopWords(['ندرخت', 'و', 'لو']))

test ×

/home/mohammad/Documents/uni/term8/information-retrieval/project

['درخت']

index | - | |

total 67M

-rw-rw-r-- 1 mohammad mohammad 29M Apr 21 18:40 inverted-index-without-stopwords.dat

-rw-rw-r-- 1 mohammad mohammad 39M Apr 21 18:38 inverted-index-with-stopwords.dat
```

اختلاف حجم شاخص معكوس بعد از حذف ايست واژه ها

* ریشه کلمات را ذخیره می کنیم . این کار باعث می شود کلماتی که یک مفهوم را دارند و یکی هستند به شکل های مختلف ذخیره نشوند و موجب می شود موتور جستجوی ما عملکرد بهتری داشته باشد. برای مثال باعث می شود کلمه خواستن، خواستم و خواستیم را با یک دید به آنها نگاه کنیم و همه این ها را یک واژه ببینیم. ریشه یابی ممکن است اطلاعاتی را نیز از دست بدهیم ولی در اکثر کاربردها مفید است . برای ریشه یابی از کتابخانه parsivar استفاده شده است .

سوال دوم)

قانون zipf : تعداد تکرارهای کلمه i ام که بر اساس تکرار در کل اسناد مرتب شده است ؛ ارتباط عکس با i دارد .

- Equivalent: cf_i = K/i where K is a normalizing factor, so
 - $\log \operatorname{cf}_i = \log K \log i$
 - Linear relationship between log cf_i and log i

قانون zipf در صورت حذف ایست واژه ها :

پرتکرار ترین کلمات در حالتی که ایست واژه ها حذف شده اند.

```
import pickle
import math
import matplotlib.pyplot as plt

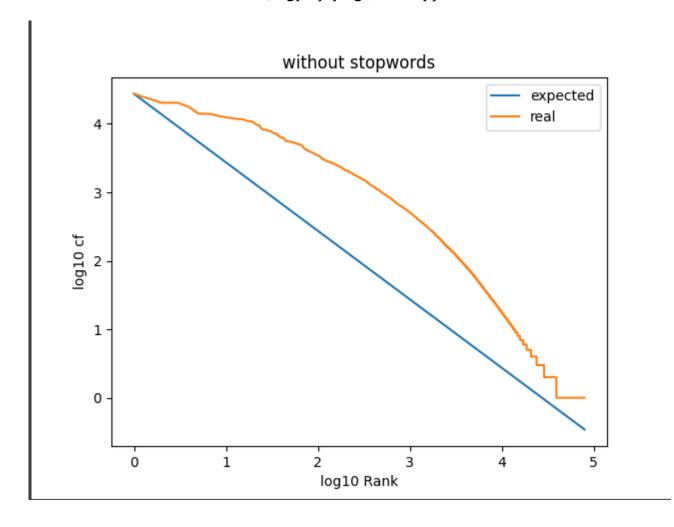
inverted_index = pickle.load(open("../inverted-index-with-stopwords.dat", "rb"))
inverted_index = {k: v for k, v in sorted(inverted_index.items(), key=lambda item: item[1].allFrequency, reverse=True)}
key_list = list(inverted_index.keys())
all_frequency_list = [x.allFrequency for x in inverted_index.values()]
max_frequency = all_frequency_list[0]

cf_real = [math.log10(i) for i in all_frequency_list]
rank = [math.log10(i) for i in range(len(key_list))]
cf_expected = [math.log10(max_frequency / (i + 1)) for i in range(len(key_list))]

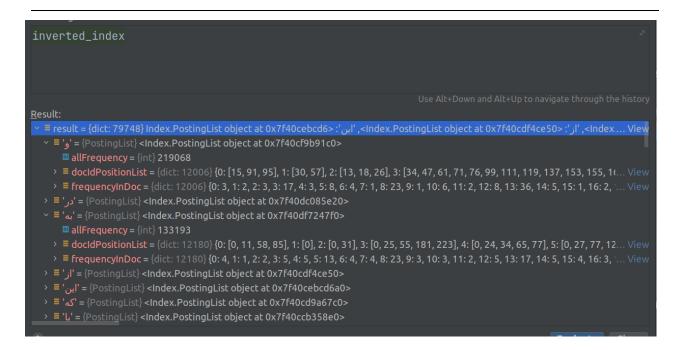
plt.title('with stopwords')

plt.plot(rank, cf_expected)
plt.plot(rank, cf_expected)
plt.legend(["expected", "real"])
plt.xlabel("log10 Rank")
plt.ylabel("log10 Rank")
plt.ylabel("log10 Rank")
plt.show()
```

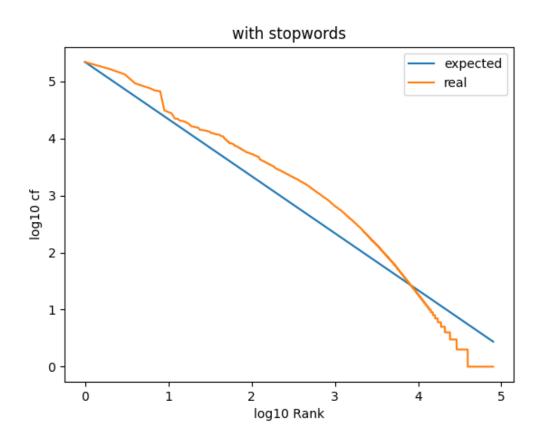
کد مربوط به کشیدن نمودار قانون zipf



قانون zipf بدون حذف ايست واژه ها:



پرتکرارترین کلمات بدون حذف ایست واژه ها



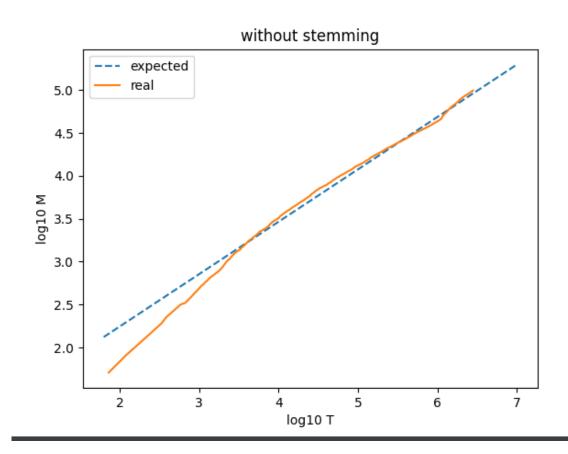
همانطور که مشاهده می شود زمانی که ایست واژه ها وجود دارند منحنی با دقت بهتری روی نمودار این قانون فیت می شود ولی در حالتی که ایست واژها حذف شده اند فاصله بیشتری با خط راست دارد . پس این قانون برای حالتی که ایست واژه ها حذف نشده اند تخمین بهتری می زند و اختلاف کمتری با مقدار واقعی دارد .

سوال سوم)

قانون heaps: یک قانون تجربی در زبان شناسی می باشد که ارتباط بین تعداد کلمات دیکشنری و تعداد کل کلمات اسناد را بیان می کند . در ادامه این قانون را روی اسناد ، قبل و بعد از ریشه یابی بررسی می کنیم .

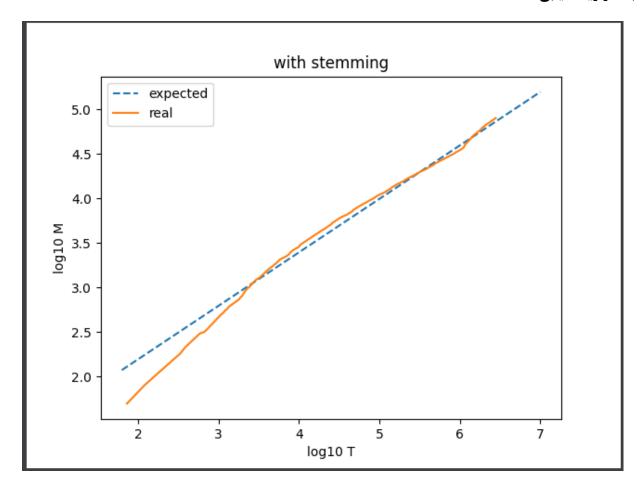
• $\log M = \log k + b \log T$

بدون ریشه یابی :



در تصاویر بالا نمودار خطی روی منحنی فیت شده است و مقادیر b,k محاسبه شده است . سپس مقادیر واقعی و مقدار تخمینی بعد از بررسی شدن تعداد مشخصی از اسناد با هم مقایسه شده اند .

بعد از ریشه یابی:



به وضوح در تصاویر بالا همه اطلاعات بعد از بررسی شدن تعداد مشخصی اسناد نشان داده شده است .

همانطور که مشخص است هر چه تعداد بیشتری از اسناد بررسی می شود این قانون تخمین بهتری برای اندازه دیکشنری دارد . همچنین در هر دو حالت قبل و بعد از ریشه یابی می توان مشاهده کرد که با دقت خوبی روی نمودار خطی فیت شده اند . قبل از ریشه یابی شیب و عرض از مبدا نمودار تخمینی کمی بیشتر از بعد از حالتی است که ریشه یابی صورت گرفته است .

سوال چهارم) بحث ریشه یابی یکی از پر چالش ترین قسمت های ساخت موتور بازیابی اطلاعات است . کلماتی که به خودی خود، حالت ریشه دارند ولی بعد از فراخوانی تابع stem به شکل دیگری تبدیل می شوند که ممکن هست با کلمات دیگر اشتباه گرفته شوند. به طور کلی در ریشه یابی ممکن است اطلاعاتی از دست دهیم .

در ۱) فعل ها ۲)کلمات هم خانواده ۳)کلمات مفرد و جمع با چالش های زیادی روبرو هستیم .

مثلا علم ، عالم ، تعليم ، معلم و ... همگی به علم تبديل شوند .

در فعل ها ممکن است اطلاعاتی ازدست برود مثلا "خواهم رفت" به کلمه ایی دیگر تبدیل شود ؛ در این صورت دیگر یک فعل آینده نیست . در کلمات مفرد و جمع نیز ممکن است اطلاعاتی را از دست بدهیم .

واژه "رستم" یک اسم است که بعد از ریشه یابی به واژه "رس" تبدیل میشود که با واژه "رس" (خاک رس) اشتباه گرفته میشود. مثال بعدی واژه "آتش" است که بعد از ریشه یابی میبینیم تبدیل به واژه "آت" شده است که به اشتباه "ش" انتهایی آن حذف شده است و تبدیل به واژه بی معنایی شده است.

البته كتابخانه parsivar دقت بيشتري براي ريشه يابي دارد كه براي ساخت شاخص معكوس از آن استفاده شده است .

سوال ينجم)

الف) پرسمان : تحریم های امریکا علیه ایران

'تحریم' - A2 | grep میرسد مذاکرات و cat <u>IR data news 12k.json</u> | jq | grep ما '6929' - A2 | grep تحریم' rontent: ه گزارش خبرنگار سیاسی خبرگزاری فارس، به نظر میرسد مذاکرات وین به روزهای آخر خود نزدیک شده است و براساس اعلاا": "content:" برخی منابع تا رسیدن به توافق راه زیادی باقی نمانده و تنها نیاز است آمریکا به آنچه در برجام برای برداشتر مل کند و تضمین معتبری هم برای انجام این تعهدات داده شود. براساس اعلام مسئولان سیاست خارجه در کنار دریافت تضمین معتبر، برداشتر مه تحریمها از نکات و مطالبات اصلی ایران در مذاکرات است. اما به نظر میرسد آمریکا در مقابل عمل به این تعهد برجامی مقاومت میکند. * تحریمهای آمریکا علیه ایران آمریکا پس از تسخیر لانه جاسوسی آمریکا در تهران، «کارتر» در تاریخ 8 نوامبر 1979م، با استناد به قا

همانطور که مشاهده می شود مستندی که اول نمایش داده شده است دارای پرسمان کاربر است . بقیه موارد نیز داری پرسمان کاربر هستند .

ب)پرسمان: تحریم های امریکا! ایران

جت الاسلام مجتبی ذوالنور نماینده مردم قم در مجلس در گفتوگو با فیرنگار بارلمانی خبرگاری فارس، با اشاره به م۱۳ ": "content": "content": "ساله تحریمهای مدایل بر امان به م۱۳ ": "content": "ساله تحریمهای مدایل با در این در حوزههای مختلف و ریشه های اصلی وقوع پدیده تحریم اظهار داشت: فلسفه ومع تحریمها، افرایش فشار بر مردم و نظام بوده است تا تابآوری مردم از بین برود و مقاومت آنها بشکند و در نهایت تسلیم شوند. وی افرود: دشمن با این هدف تحریمها را ادامه سنده و هر چیزی که دشمنان ما احساس کنند که میتواند در مساطل کردن مسؤولان و نیز در تنگنا قرار دادن مردم فرز و افق شود، از آن ا ادامه سنده و در نهاید بن بر مرد تسلیم شوند. عمو کمیسیون امنیت ملی و سیاست خارجی مجلس تمریح کرد: مذاکرات هسته ای از زمان دولت خا تمی و زمان بر آن این المنیت ملی و سیاست خارجی مجلس تمریح کرد: مذاکرات هسته ای از زمان دولت خا در در این که آفای روحانی دبیر شورای عالی امنیت ملی بوده آن که آفیان لاریجانی، طریف و غرافچی مذاکرات را دنبال ک در در النو در توضیح این نکته گفت: امیدوار شدن در حراها بین مین مدورا شده که آفیان لاریجانی، خبیر در زمانهایی نیز مستاطی شده است. و در توضیح این نکته گفت: امیدوار شدن دشمن امیدوار شده که اصداس کند ما در وضعیت شکننده داخلی هستیم و قدرت چا در در در در خوان می این میز مداکری در داخلی را دیده این در در خوان می و در است که بخبارچکی و انسجام داخلی را دیده این در در خوان میان موان مقاومت مداکرکنند که امانیان در میان میز مداکره بود و آمد آفازادها به غرب نماینده میرد و میر نمایش و نمایس کند مداکرکند در به میرد مداکرکند در به میرد مداکرکند و به خربی این میز مداکره بود و حتی اطلاعاتی در دستر این و زمید مداکند که ما نفس نظام را در دا شدن میرد کرد از این در در این میز مداکرک بدر در تمیک از مأموریکان فتیه 88 بوده است که یک نوع نفرخه و شکیدگی بر از در اخل کشور ایجاد کرد و به دشمن سیگال داد. * این اینک اینک این امیده این اینک امید و نمید در نمایل ما فرار دارد و کم نمید میشود، گفت: امروز، یک را هبرد لغو تحریمها داری در مدید میشود این برد میان نمید می در داخل کشور ایان در داخل کشور در مو از در در مدید مین ایک بر در میدان در بید مشید که از در کردن در مداکران با بیان بیده در در اخل کشور در ادار کشید که از در کردن در مداکران با جدیل میزیم تا می

در مستند ۷۶۱۵ فقط یک کلمه ایرانی وجود که هنگام ساخت شاخص به ایران تبدیل نشده است . به همین دلیل در پاسخ پرسمان ظاهر شده است . همچنین کلمه تحریم و امریکا در این مستند موجود است .

پ)"کنگره ضدتروریست"

```
(6929: [[344, 345]]}

1 documents found that have at least 1 word of user query)

The results are ranked in the following order
{1: {6929: 1}}

rank: 1 , doc_id: 6929 , frequency: 1

title: توضیحات منبع آگاه درباره وقفه مذاکرات وین URL: https://www.farsnews.ir/news/14001222000450/
```

به گزارش خبرنگار سیاسی خبرگزاری فارس، به نظر میرسد مذاکرات وین به روزهای آخر خود نزدیک شده است و براساس اغالاه"

به گزارش خبرنگار سیاسی خبرگزاری فارس، به نظر میرسد مذاکرات وین به روزهای آخر خود نزدیک شده است و براساس اغالاه" "content":

مل کند و تضمین معتبری هم برای انجام این تعهدات داده شود. براساس اغلام مسئولان سیاست خارجه در کنار دریافت تحمین معتبر، برداشتن محتبر، برداشتن معتبر، برداشتن معتبر، برداشتن هم برای انجام این تعهدات داده شود. براساس اغلام مسئولان سیاست خارجه در کنار دریافت تحمین معتبر، برداشتن همه معتبر، برداشتن معتبر، برداشتن همه معتبر، برداشتن معتبر، برداشتن معتبر، برداشتن همه معتبر، برداشتن معتبر، برداشتن معتبر، برداشتن المعتبر، برداشتن المعتبر، برداشتن عدام داراییهای خود در تاریخ ۹ نوامبر 1979م. با استناد به فا نون «کنترل صدور تسلیحات نظامی»، کشتی حامل لوازم یدکی نظامی متعلق به ایران را توقیف کرد. ارزا باکنهای آمریکا خارج خواهد کرد. کا اوج گیری کشمکشهای سیاسی بر سر مسئله تصرف سفارت، دولت موقت، اغلام کرد تمام داراییهای خود را از بانکهای آمریکا خارج خواهد کرد. کا اوج گیری کشمکشهای سیاسی بر سر مسئله تصرف سفارت، دولت موقت، اغلام کرد تمام داراییهای خود را از بانکهای آمریکا خارج خواهد کرد. کا و قانون شر (IEEP) بارتر با لحاط این احتمال، در کشور شرایط اصطراری اغلام کرد تمام داراییهای خود را از بانکهای آمریکا خارج و خواهد کرد. کا معدور دستور و بزره شماری (IEEP)، تمام داراییهای ایران در آمریکا، 2 میلیارد دلار در بانک خودرال زرو» آمریکا، افروک و خودی ایران بالغ بر 12 میلیارد دلار در سایر با نکهای آمریکا و خدود 2 میلیارد دلار در این خودرال زرو» آمریکا، ماروک میلیارد دلار در سایر با نکهای آمریکا و خدود 2 میلیارد دلار در اندر خودی در اوریل ۱۹۵۹م، کارنر با مدور دستور ویزه شماری (1225)، فروش و غرضه هرگونه کالا از ایران بیز ممنوع شد، امید در ناریخ کووک کرد، اما و نوان نیز ممنوع شد، امید در ناریخ کووک کرد، اما و نوان نیز ممنوع شد، امید در ناریخ کووک کرد، در این در اسای در سایا میادن در ایران در خود در این در ایران بین ممنوع شد، امید در ایران نور خود کرد، ساز کردن اموان اعلان در ایران در میان در در ایران سایم بود. در این داری در ایران اسایم بود. در سای ۱۸۹۹م، آمریکا مادرات برخی از مواد شیمیایی با در ایران اسایم

با توجه به تصاویر بالا مشخص است که به درستی مستند بازیابی شده است

ت)پرسمان: "تحریم هسته ای" امریکا! ایران

امیر فتوتی فعال سایق دانشجویی و از ادوار اتحادیه جاسعه السامی دانشجویان در گفتوگو با خبرنگار تشکلهای دانشگاهای دانشگاهای در درباره ارتباط تعامل با کشورهای غربی و آمریکا و حل مشکلات کشور گفت: اساسا در دبیلماسی به خبرگزاری فارس، با اشاره به برخی اطهارات درباره ارتباط تعامل با کشورهای غربی و آمریکا و حل مشکلات آنها را و اسام در دبیلماسی انگاراته است آگر فوری است کنی فردن استکباری مرتبواند مشکلات فر کشور در احتد، وی ادامه داد: تنها راه حل بیشرفت با انگاراته است آگر فروی را حل کند. وی ادامه داد: تنها راه حل بیشرفت با انگاراته است آگر فروی را حل کند. وی ادامه داد: تنها راه حل بیشرفت با کنور، در ابتدا نگاه به توانمندی های درونزای آن کشور است. به فری اگر فرص را بر محت گذاره مطروحه هم بگذاریم، توجه به تجریبات هشت مسلم انگذاره مطروحه هم بگذاریم، توجه به تجریبات هما به هم نام بیشرفت با به فرموده مولا علی(ع) غیری کشور به سیبار است و متاسفاته غیرت گیرندگان چه کم. کم نیستند کشورها بن که به هر نحو همسو با سیاستهای امریکا حرکت کردند و تن به خوا وی همان نام داده اما تنها مشکلا تشان رفع نشده بلکه معیف تر و مترلزل تر شده اند. این فعال سایق دانشجوید افزود: در همین د می کند و مید از مید با بیشتات کرد مشاکل تشان رفع نشده با استرالیا متابعات و متعمره های خود نیز رحم نمی کند. فروش واکشنهای آفری خود با بیا مید با بیان به نواق مورت گرفته بشات با در آن برهه تغیین در نام با بیا در آن بره تغیین در ایا مید با بیا مید با بیا مید با بیا مید و در نظام خواسته های و نوان نیز از اندازم امریکا حتی به متحدان و مستعمره های خود نیز رحم نمی کند فروش واکشنهای آفر خود نیز رحم نمی کند فروش واکشنهای آفر خود نیز را داشت کرد با حتی در سایش کام با امریکا مید و نواند از ایا در آن برهه تغیین در معامل نگاه به بیرون بود. در بهترین میزان سازش با غرب، ایا سطح را بطه ما بیشتر از پر مید تغیین در نواند و نواند و با نواند و مید و نواند و با نواند و مید و نواند و با نواند و مید و نواند و با می نواند و امی از مشال بید و مید نواند و نواند و مید نواند و مید نواند و مید و مید نوان ان می با نوان و نواند و نواند و میان نواند و نواند اما یا گره ای این می دادند، و مید و نواند امان این میان نیزان در امان امریکا و رد نواند و مید نواند و مید نواند و مید نواند امان امان نیز به فک

جملات مستند اول بازیابی شده

ث)پرسمان : اورشلیم ! صهیونیست

```
اورشليم does not exist in the dictionary
0 documents found that have at least 1 word of user query)
------
The results are ranked in the following order
{}
```

همانطور که مشاهده می شود هیچ مستندی این کلمه را ندارد.