

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

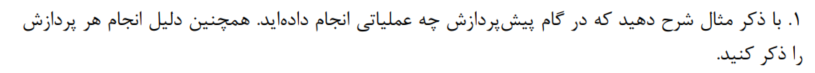
دانشکده مهندسی کامپیوتر

بازیابی اطلاعات

(پاییز ۱۴۰۰)

**گزارش فاز اول**

**محمد چوپان 9831125**



در گام پیش پردازش ابتدا محتوای هر سند را نرمال میکنیم برای مثال :



این کار باعث می شود تا تشخیص ما برای هر کلمه صحیح تر باشد و هنگامی که کلمات را از هم جدا میکنیم کلماتی که نیاز به نیم فاصله یا غیره دارند درست انتخاب شوند.

سپس کلمات را tokenize میکنیم :

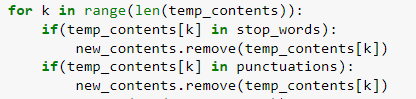


که به صورت زیر در می آید :



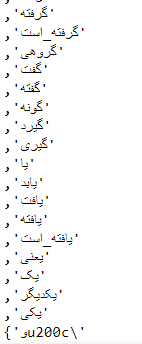
دلیل انجام این کار هم این است که هنگام جستجو query ها جستجو بر اساس token ها است.

پس از آن از کلمات پرتکرار لیستی تهیه کرده و آن ها را حذف میکنیم :



و همچنین علائم نگارشی را نیز حذف میکنیم تا در جستجو ما تاثیری نداشته باشند.

لیست کلمات پر تکرار:

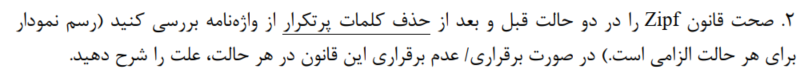


در نهایت ریشه یابی میکنیم کلمات را :

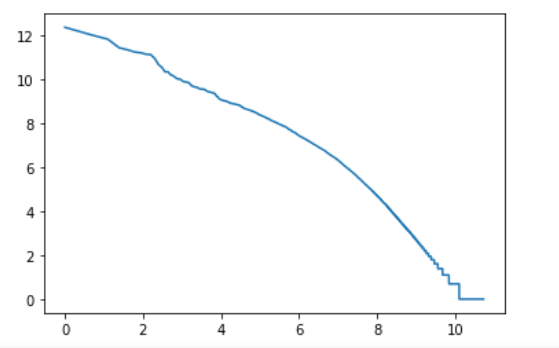


این برای این است که ریشه کلمات استفاده شده در متن یا فعل ها امکان دارد یکسان باشد و در جستجو کاربر نباید تاثیر زیادی بگذارد.

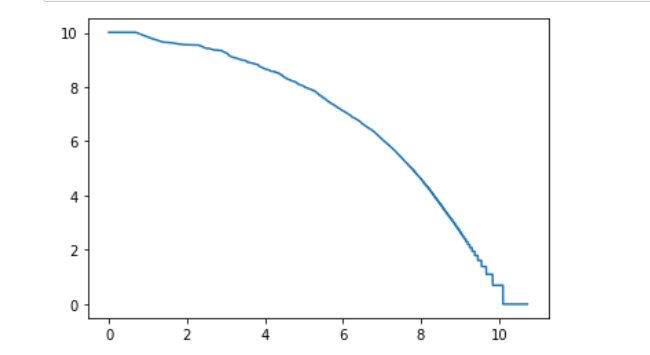




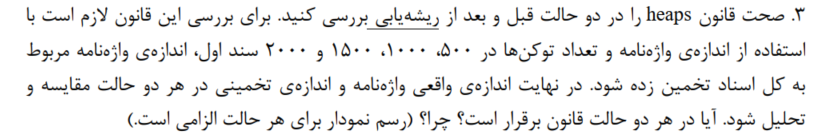
**قبل از حذف کلمات پر تکرار:**



**بعد از حذف کلمات پرتکرار:**



همانطور که میبینم در هر دو حالت از این قانون پیروی میکنند.



ابتدا تعداد واژه ها و توکن ها را قبل و بعد از ریشه یابی در همه حالات پیدا میکنیم.

**بعد از ریشه یابی :**

برای 500 سند :

تعداد کل واژه ها : 142400

تعداد توکن ها :6599

برای 1000 سند : 292131

تعداد توکن ها : 9509

برای 1500 سند :

تعداد کل واژه ها : 442653

تعداد توکن ها :11308

برای 2000 سند :

تعداد کل واژگان : 581805

تعداد توکن ها : 12824



ما با استفاده از 1000 و 1500 مقدار ها را به دست می اوریم .



حال با استفاده از این دو معادله :

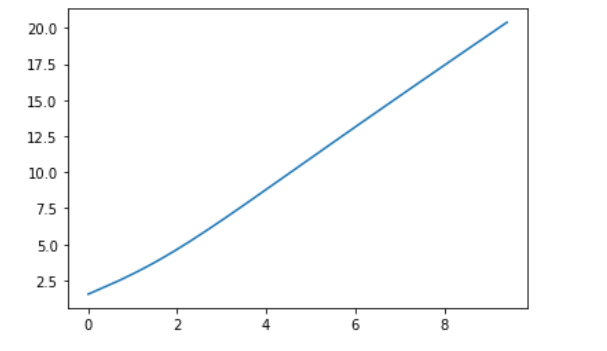


برای تعداد واژه های 4870956 مقدار تخمین زده برابر است با : 30713

مقدار واقعی:

تعداد واژه ها : 4870956

تعداد توکن ها : 45707



**قبل از ریشه یابی :**

تعداد کل : 142400

تعداد توکن ها : 9959

برای 1000 :

تعداد کل :292131

تعداد توکن ها :14755

برای 1500 :

تعداد کل :442653

تعداد توکن ها : 17802

برای 2000 :

تعداد کل : 581805

تعداد توکن ها :

20234

با استفاده از مقادیر 500 و 1000 :



که با استفاده ار این اعداد :

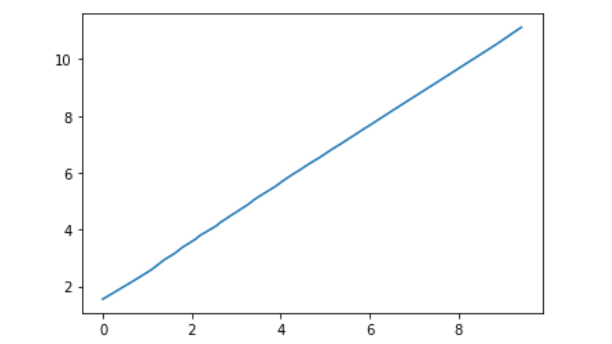


با استفاده از تخمین ها مقدار توکن ها برابر است یا :68677

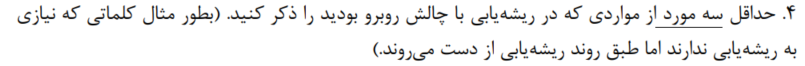
حال مقدار واقعی:

تعداد کل واژه ها :4870956

تعداد توکن ها : 72868



در هر دو حالت میبینیم که مقدار تخمین زده شده فاصله زیاد دارد و این به این دلیل است که کتاب خانه های ریشه یابی ما به درستی عمل نمی کنند و از در کل از قانون heaps تبعیت نمی شود.



برای مثال :  که در اینجا این کلمه به طور کامل از دست می رود.

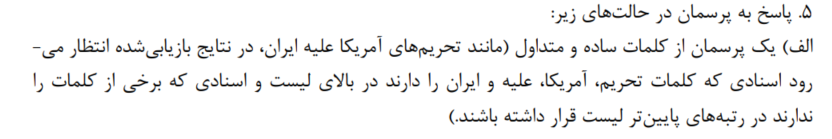
 و یا کلا دو واژه با معنا های متفاوت به دست می آید.

یا واژه های انگیلیسی که نوشتار فارسی دارند مانند :  یا 

تمامی این واژه ها از دست می روند.

 که پس از استفاده از ریشه یابی از بین می روند.

یا کلماتی که دارای تر یا ترین هستند اما صفت برتری ندارند مانند کبوتر که تر آن حذف می شود اما این غلط است.



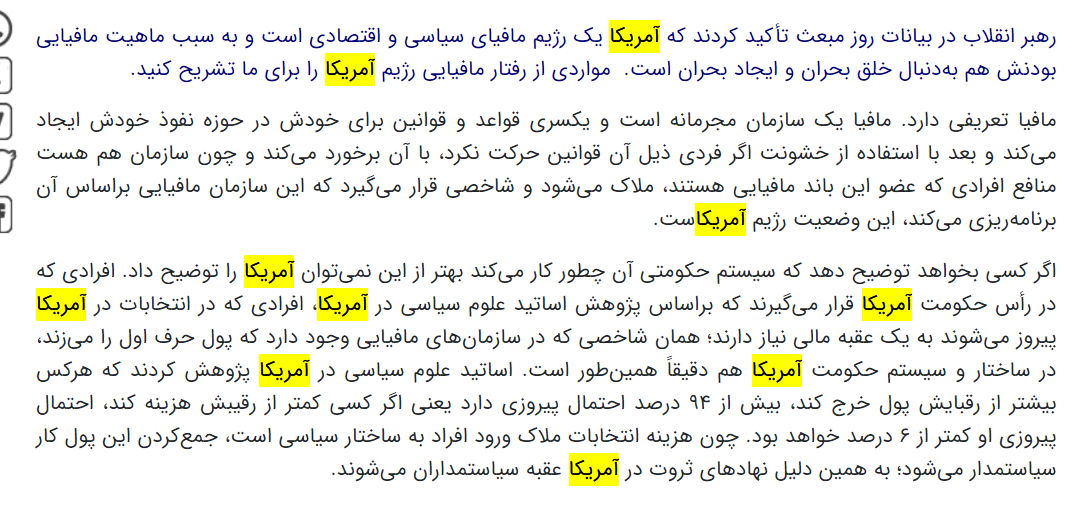
پاسخ اول :

پاسخ نا مرتبط

تیتر :

بحران اوکراین، مصداق ماهیت مافیایی آمریکاست

لینک : [https://www.farsnews.ir/news/14001214001141/بحران-اوکراین-مصداق-ماهیت-مافیایی-آمریکاست](https://www.farsnews.ir/news/14001214001141/%D8%A8%D8%AD%D8%B1%D8%A7%D9%86-%D8%A7%D9%88%DA%A9%D8%B1%D8%A7%DB%8C%D9%86-%D9%85%D8%B5%D8%AF%D8%A7%D9%82-%D9%85%D8%A7%D9%87%DB%8C%D8%AA-%D9%85%D8%A7%D9%81%DB%8C%D8%A7%DB%8C%DB%8C-%D8%A2%D9%85%D8%B1%DB%8C%DA%A9%D8%A7%D8%B3%D8%AA)



پاسخ 2 :

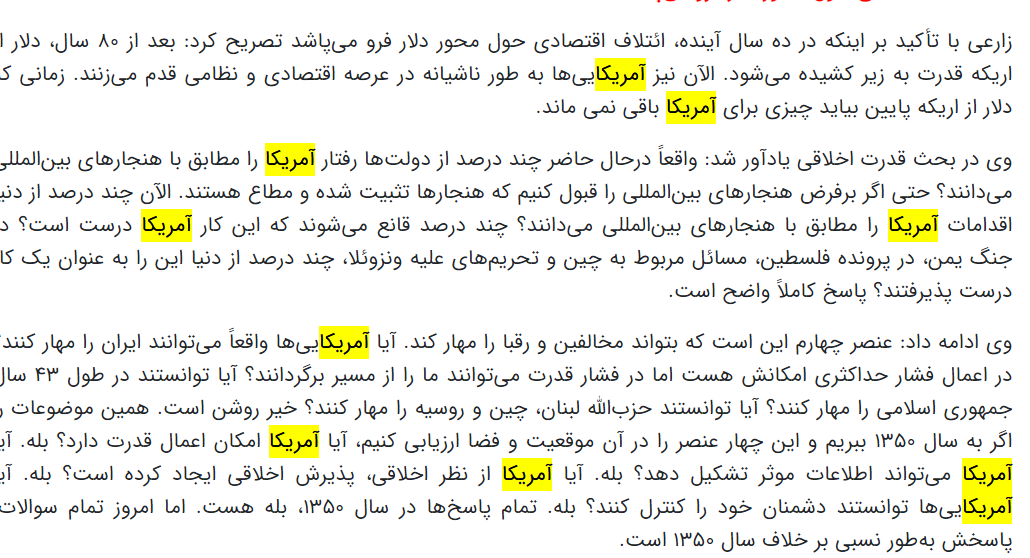
پاسخ نا مرتبط

تیتر :

میزگرد «تلاطم در اوکراین»| ائتلاف اقتصادی حول محور دلار فرو می‌پاشد

لینک :

[https://www.farsnews.ir/news/14001215000516/میزگرد-تلاطم-در-اوکراین|-ائتلاف-اقتصادی-حول-محور-دلار-فرو-می‌پاشد](https://www.farsnews.ir/news/14001215000516/%D9%85%DB%8C%D8%B2%DA%AF%D8%B1%D8%AF-%D8%AA%D9%84%D8%A7%D8%B7%D9%85-%D8%AF%D8%B1-%D8%A7%D9%88%DA%A9%D8%B1%D8%A7%DB%8C%D9%86%7C-%D8%A7%D8%A6%D8%AA%D9%84%D8%A7%D9%81-%D8%A7%D9%82%D8%AA%D8%B5%D8%A7%D8%AF%DB%8C-%D8%AD%D9%88%D9%84-%D9%85%D8%AD%D9%88%D8%B1-%D8%AF%D9%84%D8%A7%D8%B1-%D9%81%D8%B1%D9%88-%D9%85%DB%8C%E2%80%8C%D9%BE%D8%A7%D8%B4%D8%AF)



که در این متن ایران به تعداد 31 عدد و تحریم نیز موجود است.

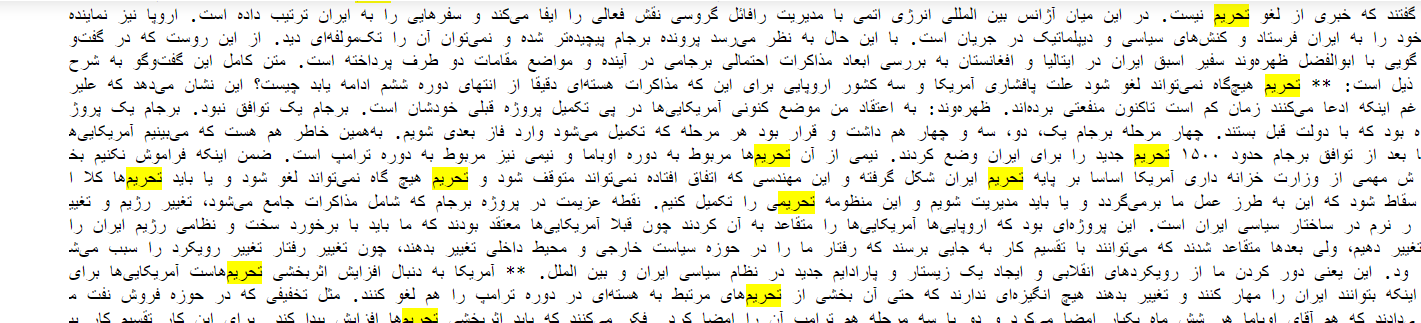
پاسخ 3 :

پاسخ مرتبط

اهرم‌سازی از افغانستان در برجام/ نقطه عزیمت آمریکا در مذاکرات جامع با ایران چیست؟

[https://www.farsnews.ir/news/14000803000676/اهرم‌سازی-از-افغانستان-در-برجام-نقطه-عزیمت-آمریکا-در-مذاکرات-جامع-با](https://www.farsnews.ir/news/14000803000676/%D8%A7%D9%87%D8%B1%D9%85%E2%80%8C%D8%B3%D8%A7%D8%B2%DB%8C-%D8%A7%D8%B2-%D8%A7%D9%81%D8%BA%D8%A7%D9%86%D8%B3%D8%AA%D8%A7%D9%86-%D8%AF%D8%B1-%D8%A8%D8%B1%D8%AC%D8%A7%D9%85-%D9%86%D9%82%D8%B7%D9%87-%D8%B9%D8%B2%DB%8C%D9%85%D8%AA-%D8%A2%D9%85%D8%B1%DB%8C%DA%A9%D8%A7-%D8%AF%D8%B1-%D9%85%D8%B0%D8%A7%DA%A9%D8%B1%D8%A7%D8%AA-%D8%AC%D8%A7%D9%85%D8%B9-%D8%A8%D8%A7)

11864

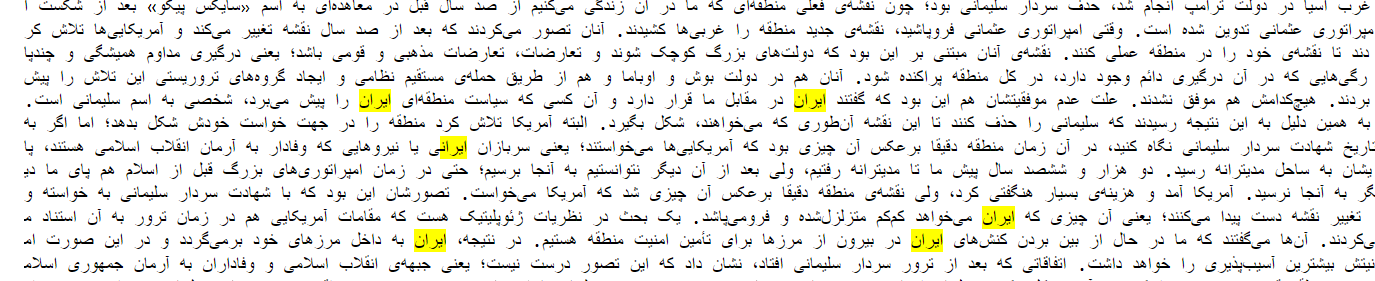


پاسخ 4 :

پاسخ نا مرتبط

کارشناس مسائل بین الملل: جبهه مقاومت بعد از شهادت سردار سلیمانی قوی‌تر شد

[https://www.farsnews.ir/news/14001006000153/کارشناس-مسائل-بین-الملل-جبهه-مقاومت-بعد-از-شهادت-سردار-سلیمانی-قوی‌تر](https://www.farsnews.ir/news/14001006000153/%DA%A9%D8%A7%D8%B1%D8%B4%D9%86%D8%A7%D8%B3-%D9%85%D8%B3%D8%A7%D8%A6%D9%84-%D8%A8%DB%8C%D9%86-%D8%A7%D9%84%D9%85%D9%84%D9%84-%D8%AC%D8%A8%D9%87%D9%87-%D9%85%D9%82%D8%A7%D9%88%D9%85%D8%AA-%D8%A8%D8%B9%D8%AF-%D8%A7%D8%B2-%D8%B4%D9%87%D8%A7%D8%AF%D8%AA-%D8%B3%D8%B1%D8%AF%D8%A7%D8%B1-%D8%B3%D9%84%DB%8C%D9%85%D8%A7%D9%86%DB%8C-%D9%82%D9%88%DB%8C%E2%80%8C%D8%AA%D8%B1)



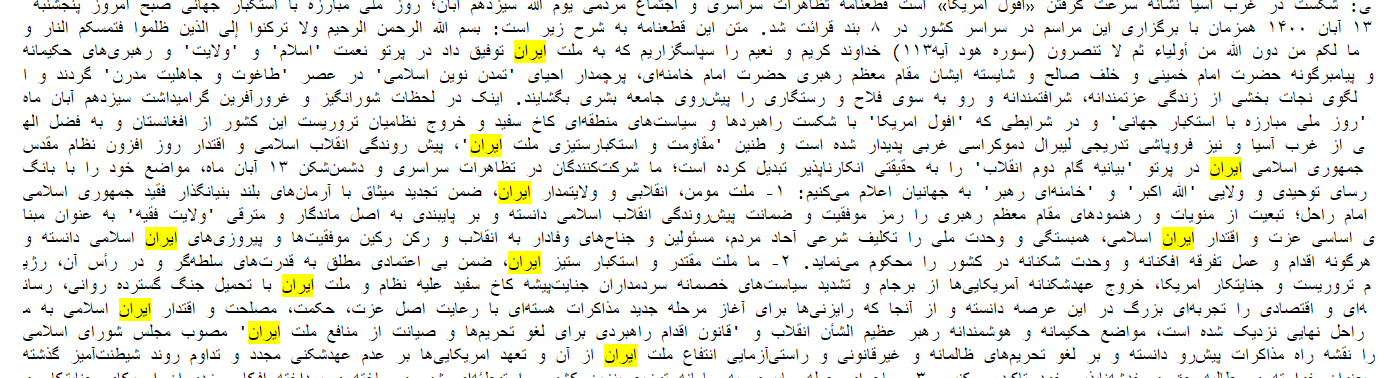
پاسخ 5 :

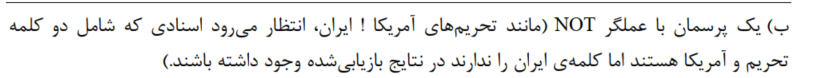
پاسخ نا مرتبط

گزارش فارس از اجتماع میدانی یوم الله 13 آبان در تهران/ تاکید بر ضرورت هوشیاری ملت‌های منطقه برابر توطئه‌های آمریکا

[https://www.farsnews.ir/news/14000813000113/گزارش-فارس-از-اجتماع-میدانی-یوم-الله-3-آبان-در-تهران-تاکید-بر-ضرورت](https://www.farsnews.ir/news/14000813000113/%DA%AF%D8%B2%D8%A7%D8%B1%D8%B4-%D9%81%D8%A7%D8%B1%D8%B3-%D8%A7%D8%B2-%D8%A7%D8%AC%D8%AA%D9%85%D8%A7%D8%B9-%D9%85%DB%8C%D8%AF%D8%A7%D9%86%DB%8C-%DB%8C%D9%88%D9%85-%D8%A7%D9%84%D9%84%D9%87-3-%D8%A2%D8%A8%D8%A7%D9%86-%D8%AF%D8%B1-%D8%AA%D9%87%D8%B1%D8%A7%D9%86-%D8%AA%D8%A7%DA%A9%DB%8C%D8%AF-%D8%A8%D8%B1-%D8%B6%D8%B1%D9%88%D8%B1%D8%AA)

11495





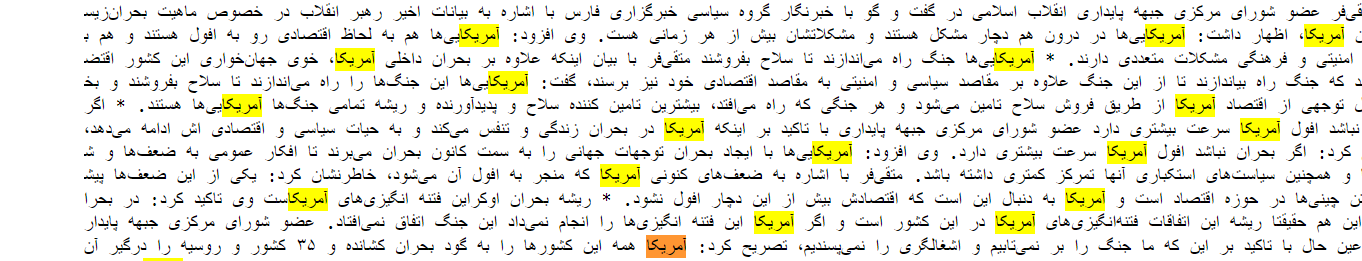
پاسخ 1 :

پاسخ نا مرتبط

آمریکا با بحران‌آفرینی به حیات خود ادامه می دهد

لینک :

[https://www.farsnews.ir/news/14001213000089/آمریکا-با-بحران‌آفرینی-به-حیات-خود-ادامه-می-دهد](https://www.farsnews.ir/news/14001213000089/%D8%A2%D9%85%D8%B1%DB%8C%DA%A9%D8%A7-%D8%A8%D8%A7-%D8%A8%D8%AD%D8%B1%D8%A7%D9%86%E2%80%8C%D8%A2%D9%81%D8%B1%DB%8C%D9%86%DB%8C-%D8%A8%D9%87-%D8%AD%DB%8C%D8%A7%D8%AA-%D8%AE%D9%88%D8%AF-%D8%A7%D8%AF%D8%A7%D9%85%D9%87-%D9%85%DB%8C-%D8%AF%D9%87%D8%AF)



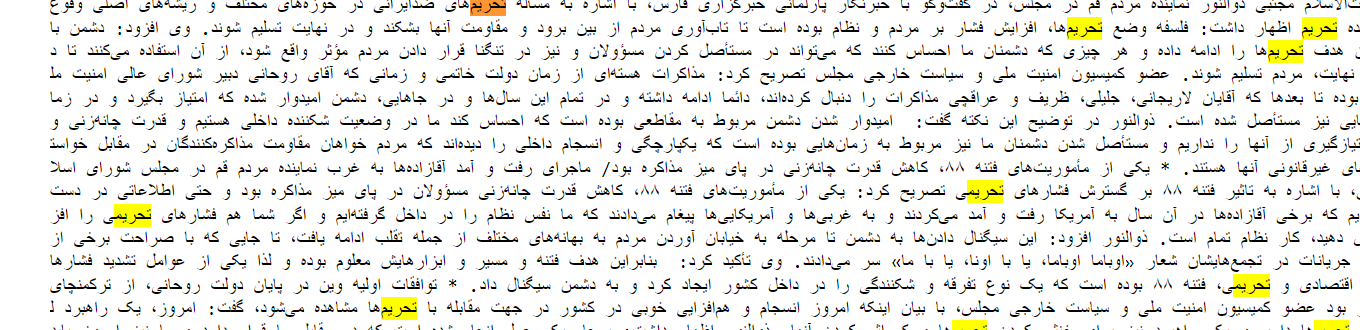
پاسخ 2 :

پاسخ مرتبط

یک مأموریت‌ فتنه 88 کاهش قدرت چانه‌زنی مسؤولان در مذاکرات بود

لینک :

[https://www.farsnews.ir/news/14001129000373/یک-مأموریت‌-فتنه-کاهش-قدرت-چانه‌زنی-مسؤولان-در-مذاکرات-بود](https://www.farsnews.ir/news/14001129000373/%DB%8C%DA%A9-%D9%85%D8%A3%D9%85%D9%88%D8%B1%DB%8C%D8%AA%E2%80%8C-%D9%81%D8%AA%D9%86%D9%87-%DA%A9%D8%A7%D9%87%D8%B4-%D9%82%D8%AF%D8%B1%D8%AA-%DA%86%D8%A7%D9%86%D9%87%E2%80%8C%D8%B2%D9%86%DB%8C-%D9%85%D8%B3%D8%A4%D9%88%D9%84%D8%A7%D9%86-%D8%AF%D8%B1-%D9%85%D8%B0%D8%A7%DA%A9%D8%B1%D8%A7%D8%AA-%D8%A8%D9%88%D8%AF)



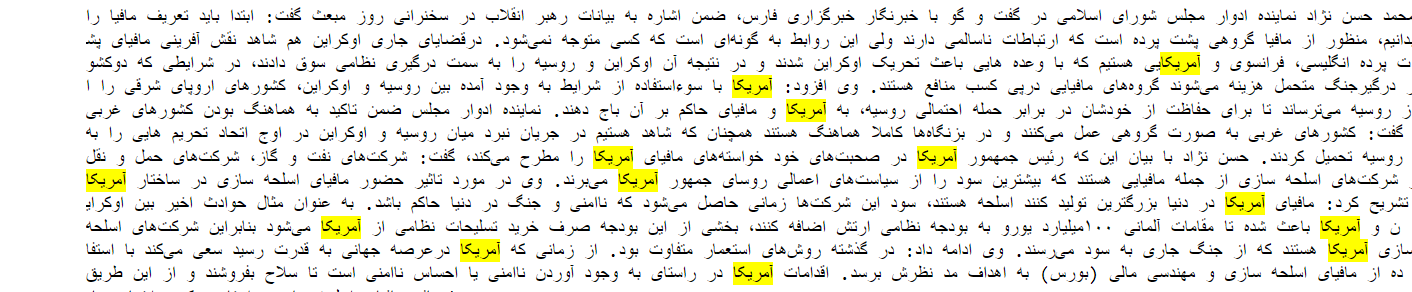
پاسخ 3 :

پاسخ نا مرتبط

سود مافیای اسلحه‌سازی آمریکا در ناامن بودن جهان است

لینک :

[https://www.farsnews.ir/news/14001211000898/سود-مافیای-اسلحه‌سازی-آمریکا-در-ناامن-بودن-جهان-است](https://www.farsnews.ir/news/14001211000898/%D8%B3%D9%88%D8%AF-%D9%85%D8%A7%D9%81%DB%8C%D8%A7%DB%8C-%D8%A7%D8%B3%D9%84%D8%AD%D9%87%E2%80%8C%D8%B3%D8%A7%D8%B2%DB%8C-%D8%A2%D9%85%D8%B1%DB%8C%DA%A9%D8%A7-%D8%AF%D8%B1-%D9%86%D8%A7%D8%A7%D9%85%D9%86-%D8%A8%D9%88%D8%AF%D9%86-%D8%AC%D9%87%D8%A7%D9%86-%D8%A7%D8%B3%D8%AA)



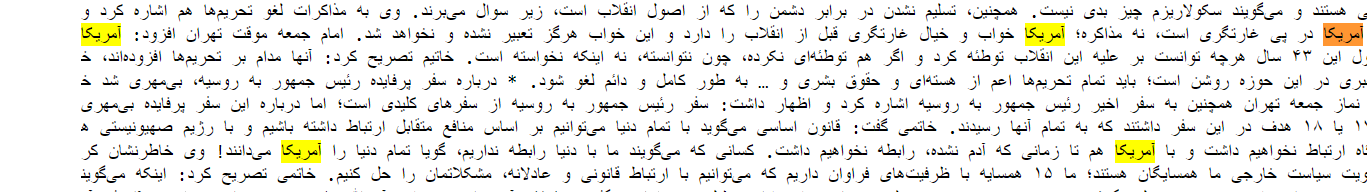
پاسخ 4 :

پاسخ مرتبط

خطیب جمعه تهران: سازمان‌ همکاری اسلامی در برابر کشتار مردم یمن سکوت کرده است

لینک :

[https://www.farsnews.ir/news/14001108000202/خطیب-جمعه-تهران-سازمان‌-همکاری-اسلامی-در-برابر-کشتار-مردم-یمن-سکوت](https://www.farsnews.ir/news/14001108000202/%D8%AE%D8%B7%DB%8C%D8%A8-%D8%AC%D9%85%D8%B9%D9%87-%D8%AA%D9%87%D8%B1%D8%A7%D9%86-%D8%B3%D8%A7%D8%B2%D9%85%D8%A7%D9%86%E2%80%8C-%D9%87%D9%85%DA%A9%D8%A7%D8%B1%DB%8C-%D8%A7%D8%B3%D9%84%D8%A7%D9%85%DB%8C-%D8%AF%D8%B1-%D8%A8%D8%B1%D8%A7%D8%A8%D8%B1-%DA%A9%D8%B4%D8%AA%D8%A7%D8%B1-%D9%85%D8%B1%D8%AF%D9%85-%DB%8C%D9%85%D9%86-%D8%B3%DA%A9%D9%88%D8%AA)



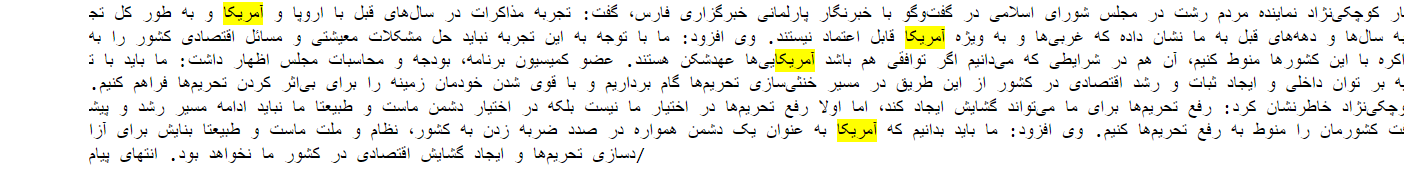
پاسخ 5 :

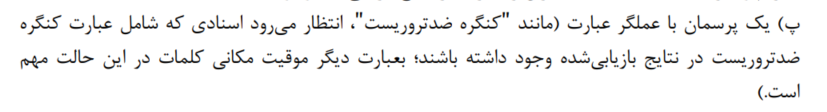
پاسخ مرتبط

مسائل اقتصادی کشور را معطل نتیجه مذاکرات نمی‌‌کنیم

لینک :

[https://www.farsnews.ir/news/14001203000383/مسائل-اقتصادی-کشور-را-معطل-نتیجه-مذاکرات-نمی‌‌کنیم](https://www.farsnews.ir/news/14001203000383/%D9%85%D8%B3%D8%A7%D8%A6%D9%84-%D8%A7%D9%82%D8%AA%D8%B5%D8%A7%D8%AF%DB%8C-%DA%A9%D8%B4%D9%88%D8%B1-%D8%B1%D8%A7-%D9%85%D8%B9%D8%B7%D9%84-%D9%86%D8%AA%DB%8C%D8%AC%D9%87-%D9%85%D8%B0%D8%A7%DA%A9%D8%B1%D8%A7%D8%AA-%D9%86%D9%85%DB%8C%E2%80%8C%E2%80%8C%DA%A9%D9%86%DB%8C%D9%85)





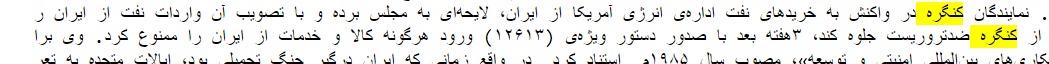
یک پاسخ یافت شد :

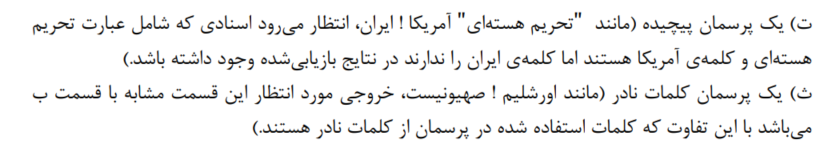
پاسخ مرتبط

توضیحات یک منبع آگاه درباره وقفه مذاکرات وین

لینک :

[https://www.farsnews.ir/news/14001222000450/توضیحات-یک-منبع-آگاه-درباره-وقفه-مذاکرات-وین](https://www.farsnews.ir/news/14001222000450/%D8%AA%D9%88%D8%B6%DB%8C%D8%AD%D8%A7%D8%AA-%DB%8C%DA%A9-%D9%85%D9%86%D8%A8%D8%B9-%D8%A2%DA%AF%D8%A7%D9%87-%D8%AF%D8%B1%D8%A8%D8%A7%D8%B1%D9%87-%D9%88%D9%82%D9%81%D9%87-%D9%85%D8%B0%D8%A7%DA%A9%D8%B1%D8%A7%D8%AA-%D9%88%DB%8C%D9%86)





برای هر دو پرسش بالا نتیجه ای یافت نشد زیرا اورشلیم در داده های ما وجود ندارد به تنهایی و برای اولی نیز به دلیل اینکه ایران نباید وجود داشته باشد نتیجه ای نداریم.