

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

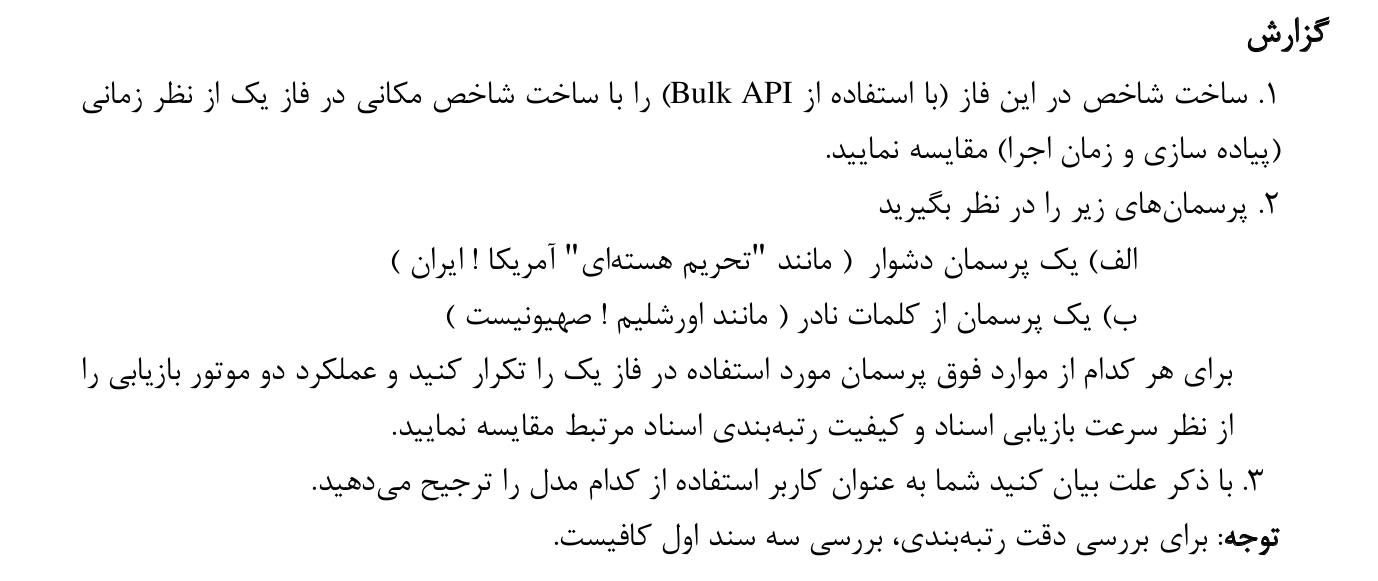
دانشکده مهندسی کامپیوتر

بازیابی اطلاعات

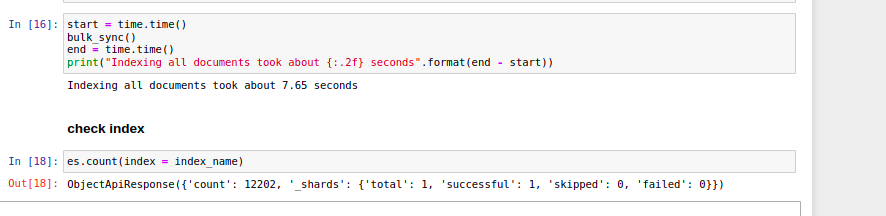
(پاییز ۱۴۰۰)

**گزارش فاز سوم**

**محمد چوپان 9831125**

  
سوال 1 :

با توجه به سرعت و پیاده سازی که Bulk API دارد که حدودا در 5 الی 6 خط پیاده سازی می شود خیلی ساده تر و راحت تر است از شاخص مکانی و سرعت بیشتری نیز دارد نسبت به شاخص مکانی همانطور که قابل مشاهده است تنها 7.65 ثانیه طول کشیده است . اما در فاز یک حدودا 20 دقیقه زمان میبرد.



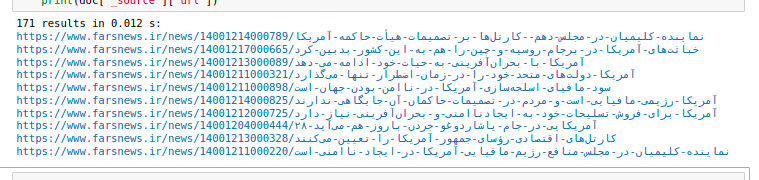
سوال 2 :

پاسخ های فاز 1 :

در فاز یک برای هر دو پرسش پاسخی یافت نشده است . اما برا این فاز :

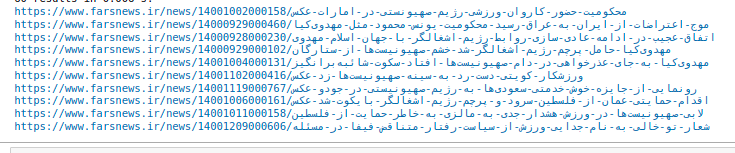


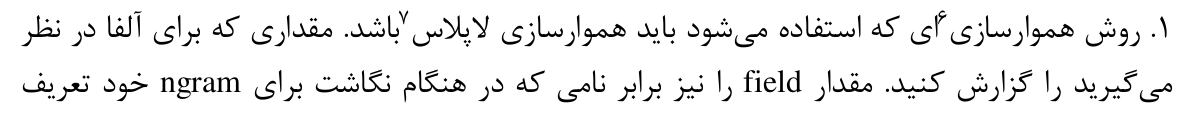
کوئری به شکل بالا است و حالا پاسخ آن :

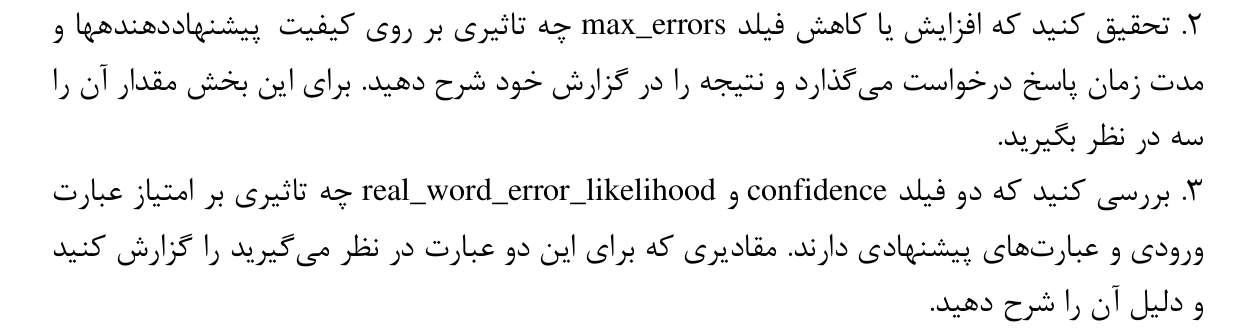
  
پاسخ ها به شکل زیر است و نسباتا مزتبط است.

برای کوئری دوم :

  
پاسخ ها :

  
میبینم که مجددا پاسخ ها نسباتا در 3 تای اول مرتبط هست.

  
مقدار آلفا را برابر 0.8 در نظر گرفته ایم .

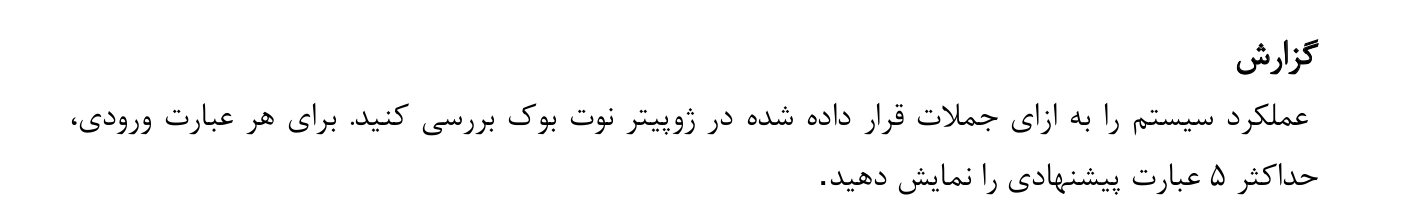
  
سوال 2 :

مقدار این متغیر یعنی بیشترین تعداد اشتباه ها در ورودی را چند در نظر بگیرید که اگر ما خیلی زیاد کنیم امکان دارد جواب ها پرت تولید کند و مقادیر 1 یا 2 بهتر است که با توجه ورودی های داده شده ما 3 قرار دادیم.

سوال 3 :

مقدار confidence را برابر 1 قرار دادیم که این مقدار به این معناست که برای عبارت ورودی یک امتیاز در نظر میگیرد مقدار این متغیر ضرب در امتیاز عبارت ورودی را threshold قرار می دهد و امتیاز های بالای آن را بر میگرداند و اگر 0 باشد n تای برتر را بر میگرداند.

مقدار real\_word را برابر 0.5 در قسمت اول گذاشته ایم و به این معناست که احتمال غلط املایی یک اصطلاح را حتی اگر در دیکشنری ما وجود داشته باشد است. یعنی اگر 0.95 باشد یعنی 5٪ کلمات واقعی غلط املایی دارند. منطقا هرچه که این مقدار بیشتر باشد امتیاز عبارت ورودی بیشتر می شود. و خروجی ها کمتر می شود اگر غلط ها همه واقعی باشند.



پاسخ های داده شده :

  
  
برای عبارت اول :

فوتبال را درست کرده اما لیگ را خیر در همه خروجی ها

برای دومی اصلا پیشنهادی نداده

برای سومی تمامی پیشنهادات اشتباه است.

برای 4 و5 هم همینطور

برای عبارت 6 پاسخ درستی داده است.

برای 7 انقلاب را تصحبح کرده است اما بقیه را خیر.

برای 8 پیشنهاد کاملا درست داده است با بیشترین امتیاز

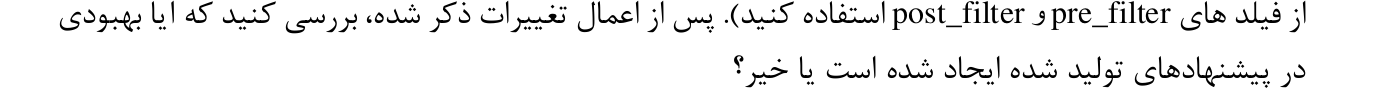
برای 9 پیشنهاد درست داده است.

برای 10 و 11 پاسخی ندارد.

برای 12 پیشنهاد اول صحیح است.

برای 13 ندارد و برای 14و15 کاملا پیشنهاد اشتباه داده است و برای 16 نیز پیشنهاد اول درست است.

به نظر من برای برای یک غلط غیر واقعی نسبتا خوب عمل کرده است اما برای بقیه خیر.



  
  
برای اولی تغییری نکرده .

2 هنوز پاسخی ندارد.

سه پیشنهاد ها بهتر شده

4 امید تصحیح شده تنها

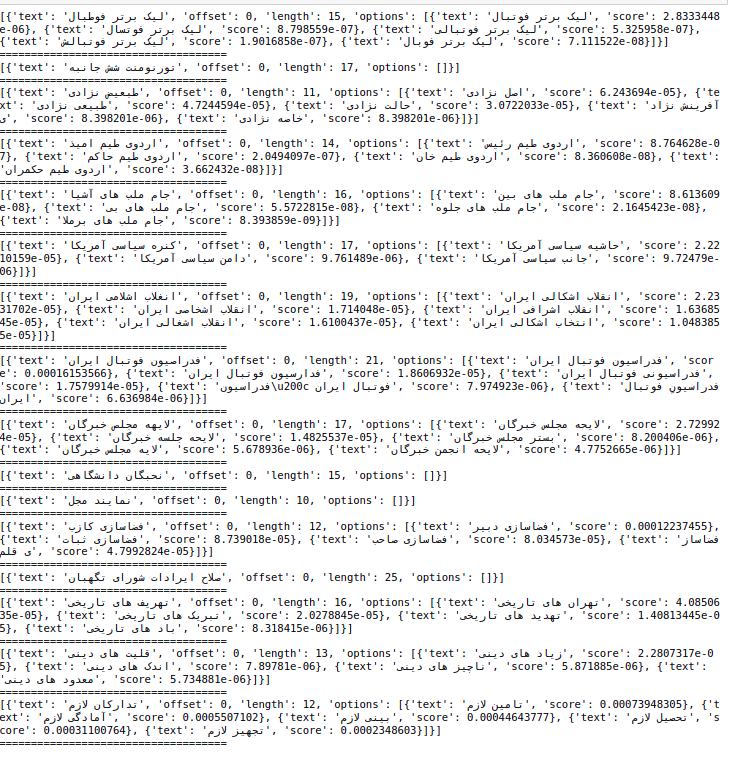
انقلاب اسلامی تصحیح شده است.

نخبگان دانشگاهی پیشنهاد درست داده است.

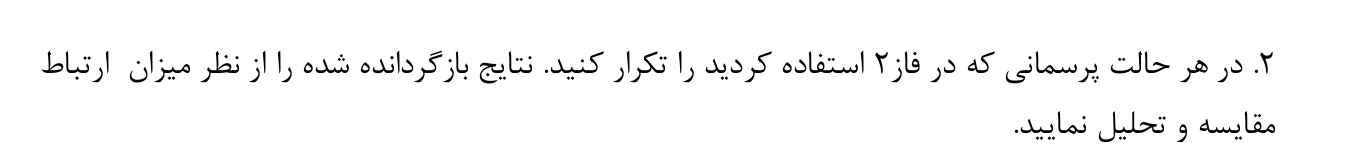
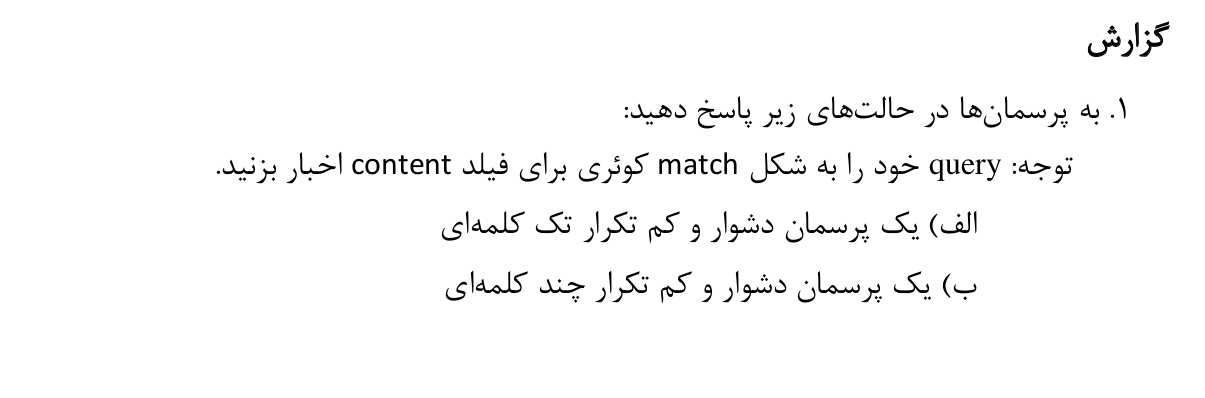
برای نماینده مجلس هنوز جوابی نداریم.

به طور کلی میبینم که برای کلماتی که ایراد در ابتدا یا انتها آن ها بوده بهتر عمل کرده است و نسبت به قبلی پاسخ ها بهتر است.

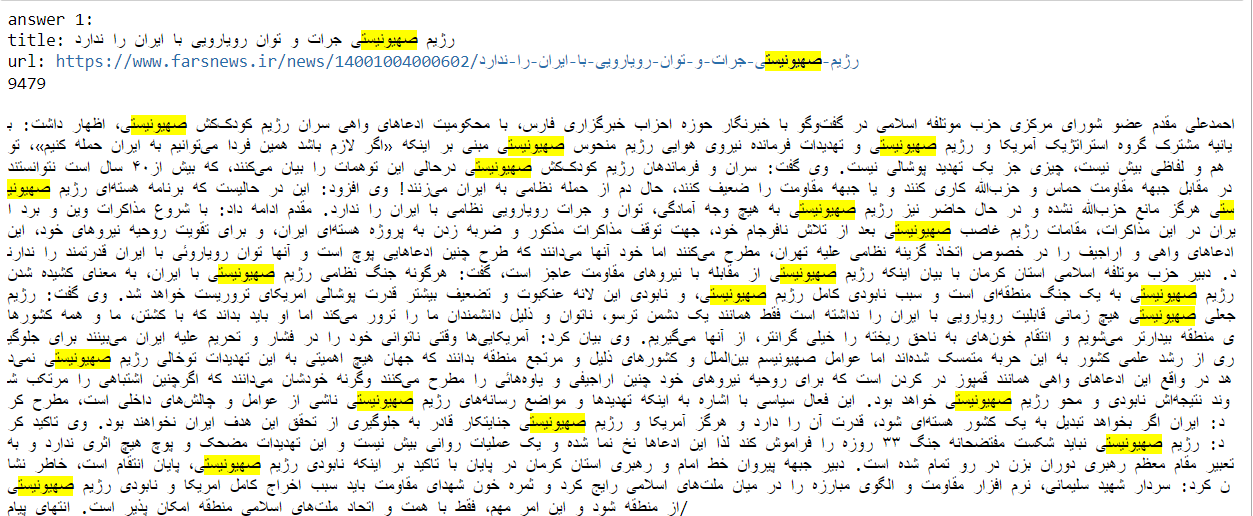
برای بخش مترادف ها :

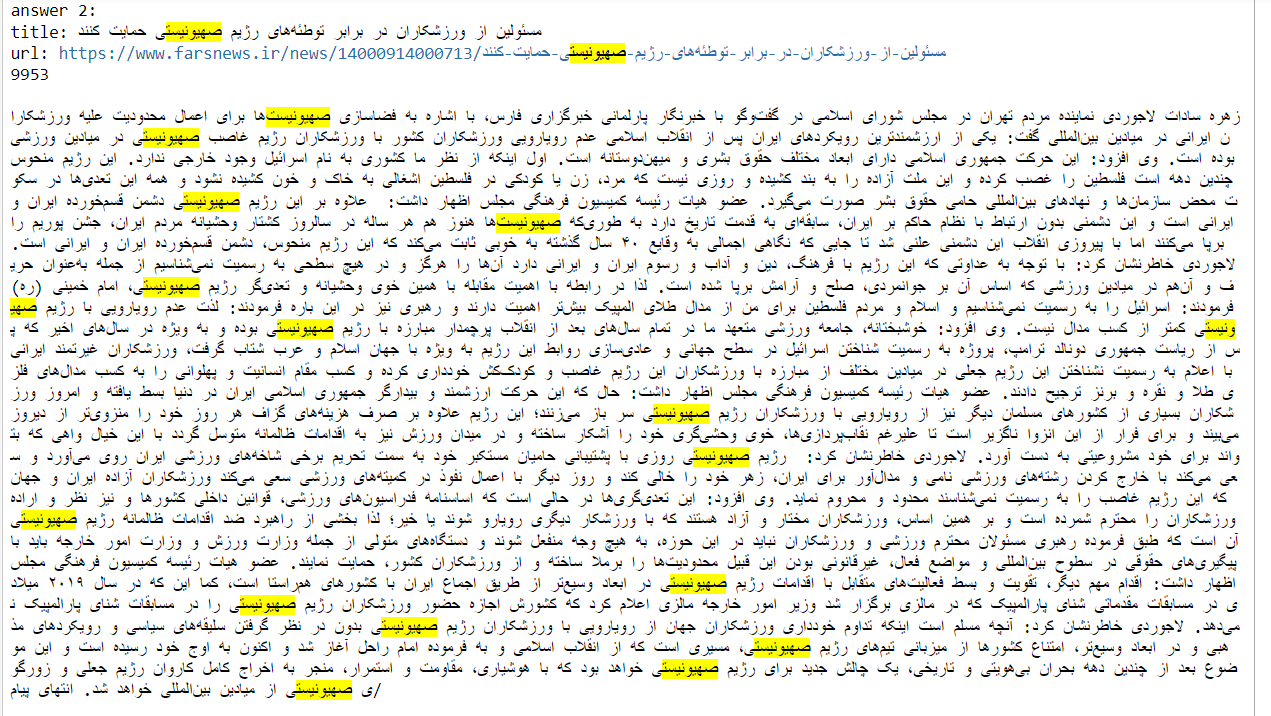


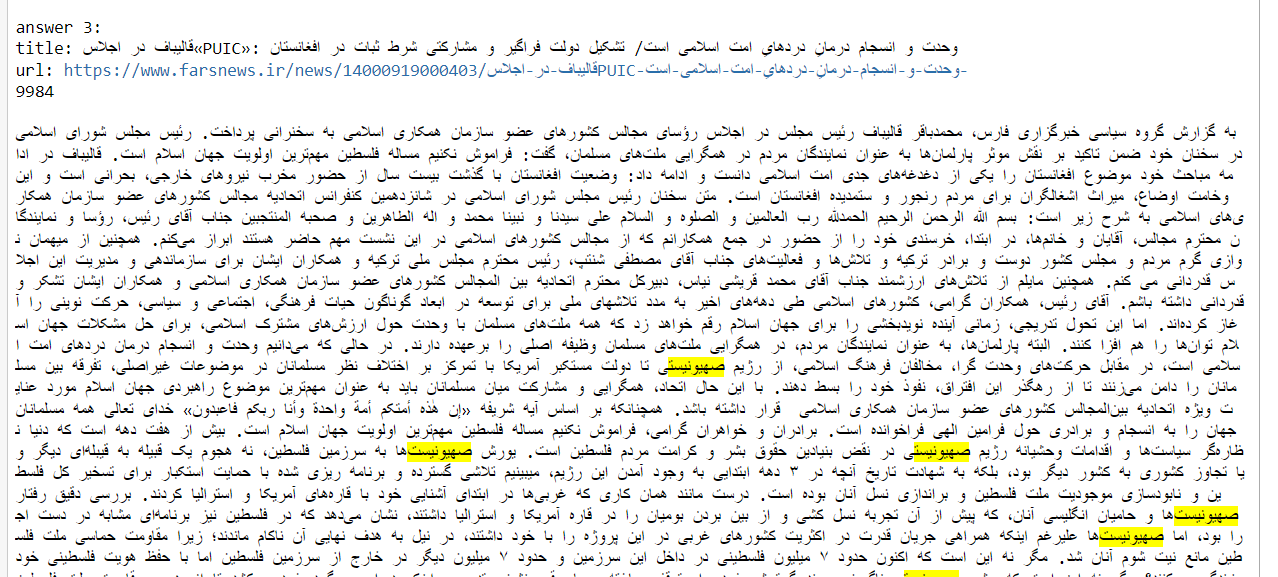
در کل نسبت به حالت قبلی تنها برای مواردی بهتر است و این شاید به دلیل اینکه لیست واژگان داده شده به آن مناسب نیست .

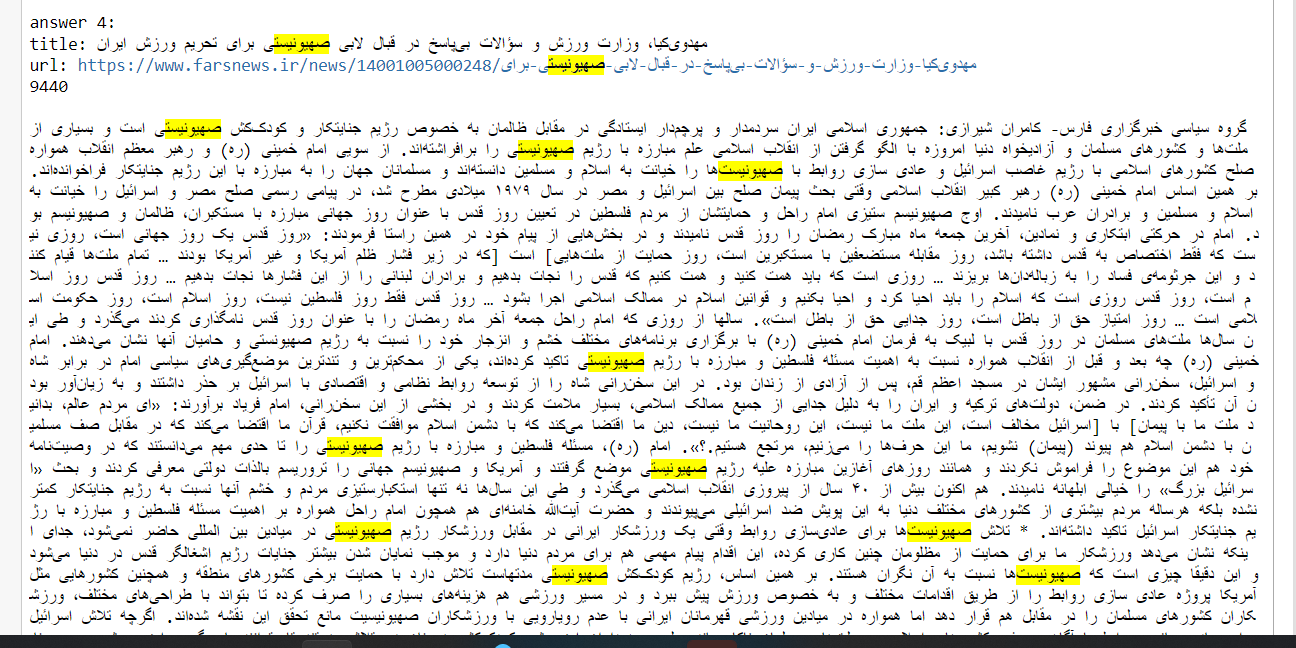
  
**ابتدا پرسمان های فاز 2 را میبینیم :**

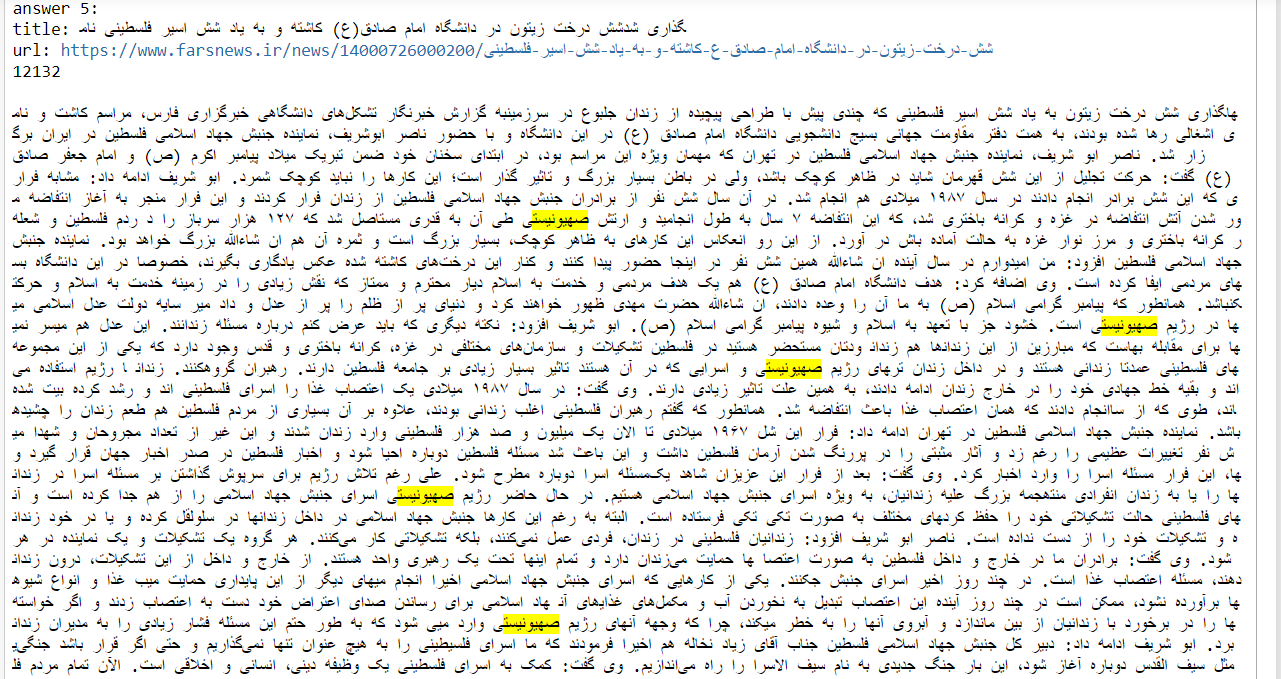
1) صهیونیست



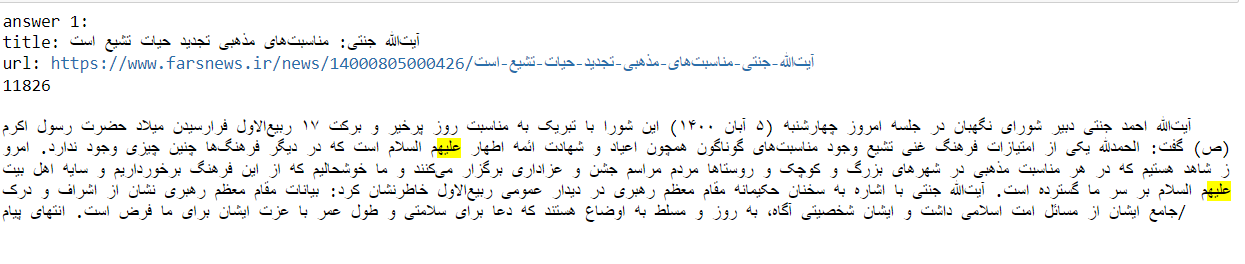




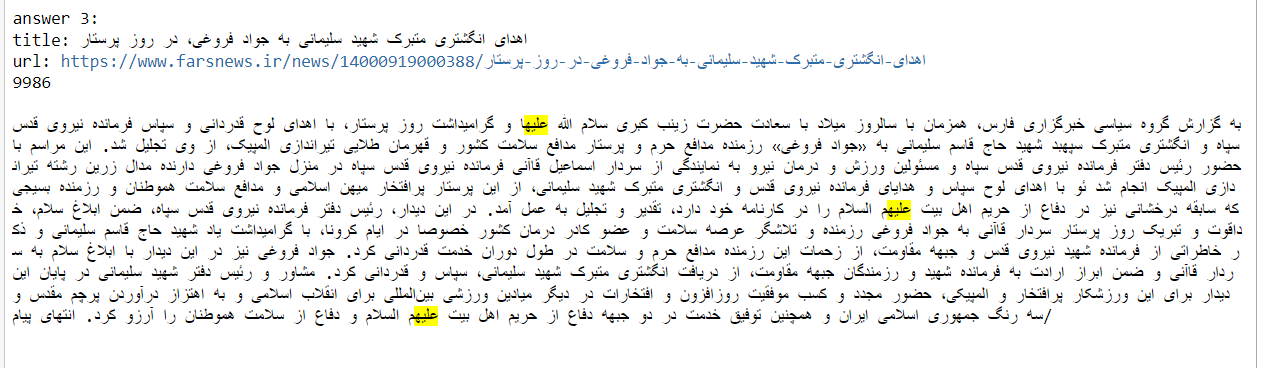


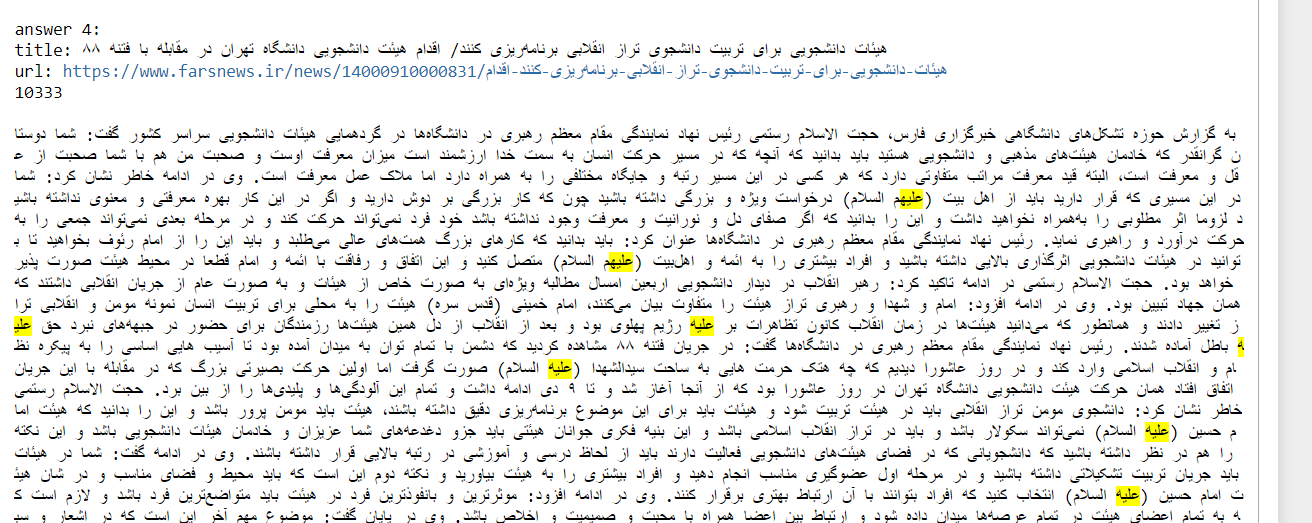


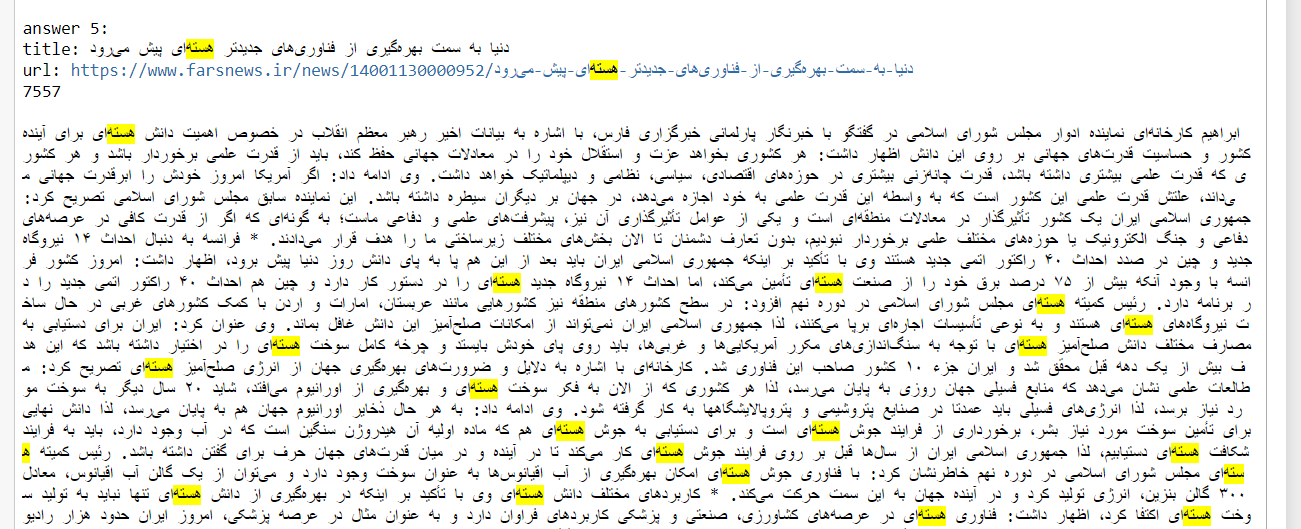
**2 ) تحریم های هسته ای آمریکا علیه ایران**











در این جستجو تنها مورد 5 مرتبط است بقیه موارد در ریشه یابی از دست رفته اند و در کل به نتیجه می رسیم که در این روش استفاده از ریشه یابی و موارد دیگر شاید خیلی درست نباشد و اینکه در کوئری های تک کلمه ای خیلی بهتر عمل می کند.

**برای این فاز :**



در این فاز نتایج مرتبط تر است اما میبنیم که برای تک کلمه‌ای با فاز 2 تقریباً برابر است و برای چند کلمه‌ای ها در این فاز خیلی بهتر عمل کرده است.