به نام خدا

گزارش فاز اول بازیابی اطلاعات

امیررضا رجبی 9831126

**سوال یک )**

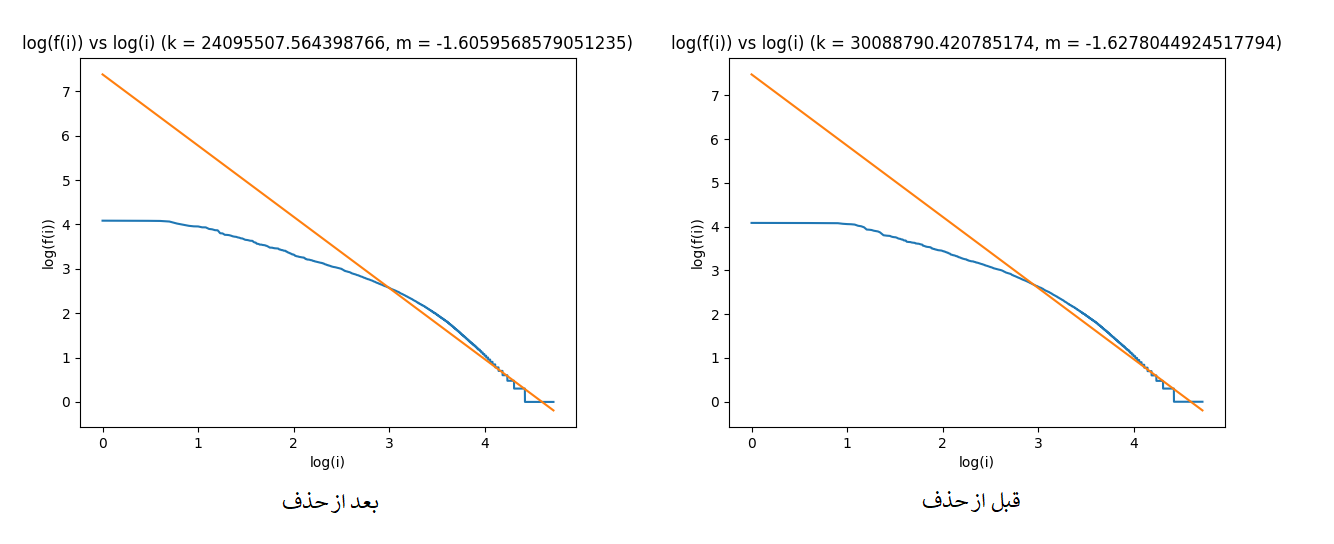
**نرمال سازی متن** : نرمال سازی متن اشاره به اصلاحات نگارشی دارد که متن ها را به حالت استانداردی (تقریبا) تبدیل کند برای این کار از کتابخانه hazm استفاده شده است .

مثالی از اصلاحات نگارشی تبدیل کتاب ها (فاصله) به کتاب‌ها(نیم فاصله)

**ریشه یابی** : کلمات را به ریشه شان برای راحتی کار و پردازش بهتر تبدیل کردم (بالا رفتن سرعت) بیابیم 🡨 یافت برای این کار از کتابخانه parsivar استفاده شده است .

**حذف کلمات پر تکرار** : کلماتی همچون از در متن ها بسیار تکرار میشوند پس برای توجه بیشتر به قسمت های مهمتر متن . حذف از از لیست استاپ ورد های hazm استفاده شده است .

**توکن ها را جدا کردن** : برای ساخت اینورتد ایندکس به کار میاید بری مثال : معلم روی تخته نوشت تبدیل می شود به (معلم ، روی ، تخته ، نوشت) از کتابخانه hazm استفاده می کنم .**سوال دو )**

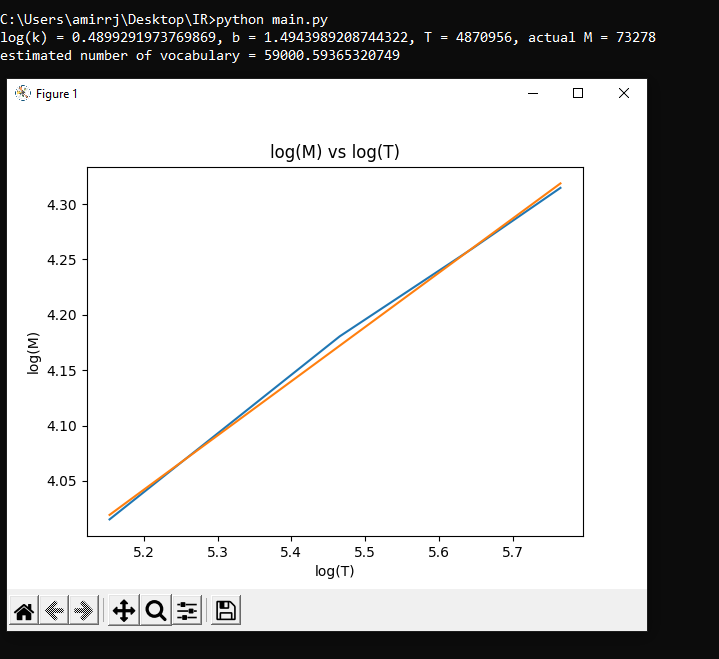


در دو حالت به طور تقریبی بر قرار نیست شیب نزدیک به یک است اما حدود 1.5 اما بعد از حذف استاپ ورد ها بیشتر در این قانون سر انگشتی بر قرار است ، چون استاپ ورد ها با تعداد بسیارشان توازن این قانون را بهم می‌زنند.

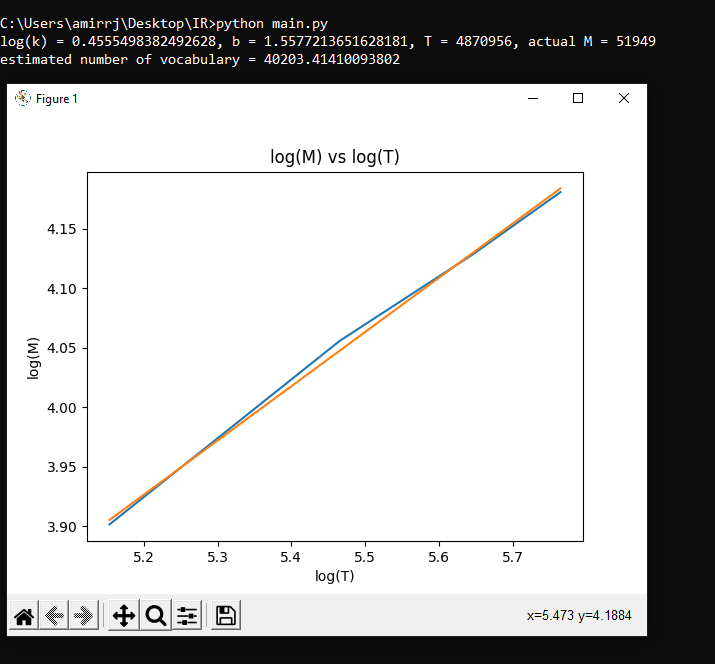
**سوال 3 )**

قبل از ریشه یابی مقادیر مربوطه به مقادیری پارامتری که در (اسلاید) واقعیت باید باشند نزدیک تر شد چرا که ما با عمل ریشه یابی داریم طول دیکشنری را کوچک تر می کنیم.

اما ازلحاظ نزدیک بودن تخمین به مقدار واقعی در حالت قبل از ریشه یابی بهتر است .

****

قبل از ریشه یابی



بعد از ریشه یابی

**سوال 4 )**

**نمونه اول** از مشکلاتی که در ریشه یابی بر خورد کردم عدم تبدیل ایرانی به ایران بود که در سند

چهارم سوال 5 قسمت ب "تیم مذاکره کننده ایرانی" را برنگردانند هم میتوانست بر گرداند که

تعیین مسِله به خودی خود سخت است که آیا در این مورد عملکرد درستی داشته یا نه .

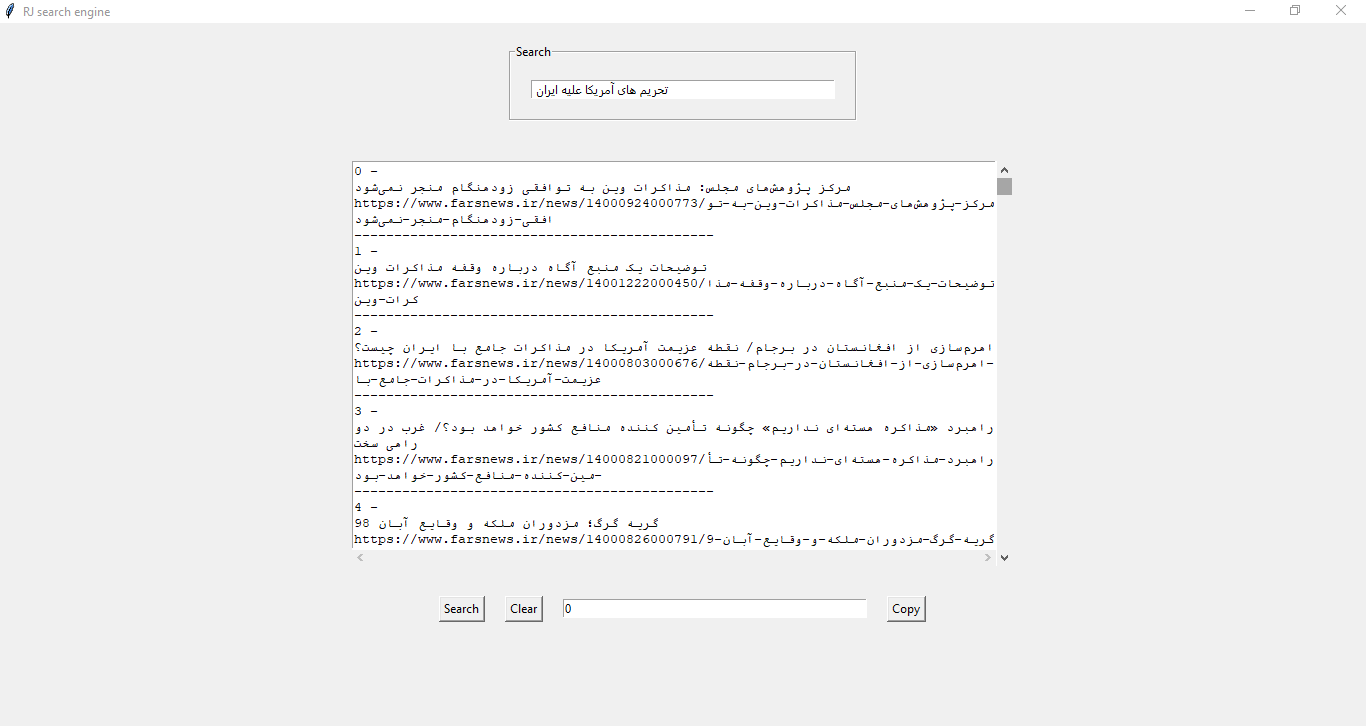
**نمونه دوم** کلمه هسته‌ای اگر به شکل هستهای باشد تبدیل به "هست" می شود .

**نمونه سوم** کلمه جنگلبان را تبدیل به جنگلب می کند .

**نمونه چهارم** کلمه دیوان را تبدیل به دیو می کند .

**سوال 5**

**الف ) تحریم های آمریکا علیه ایران :**



سند اول که به پرسمان کار بر مرتبط است و کلمات در تصویر هایلایت شده اند :



سند دوم که به پرسمان کار بر مرتبط است و کلمات در تصویر هایلایت شده اند :



سند سوم که به پرسمان کار بر مرتبط است و کلمات در تصویر هایلایت شده اند :



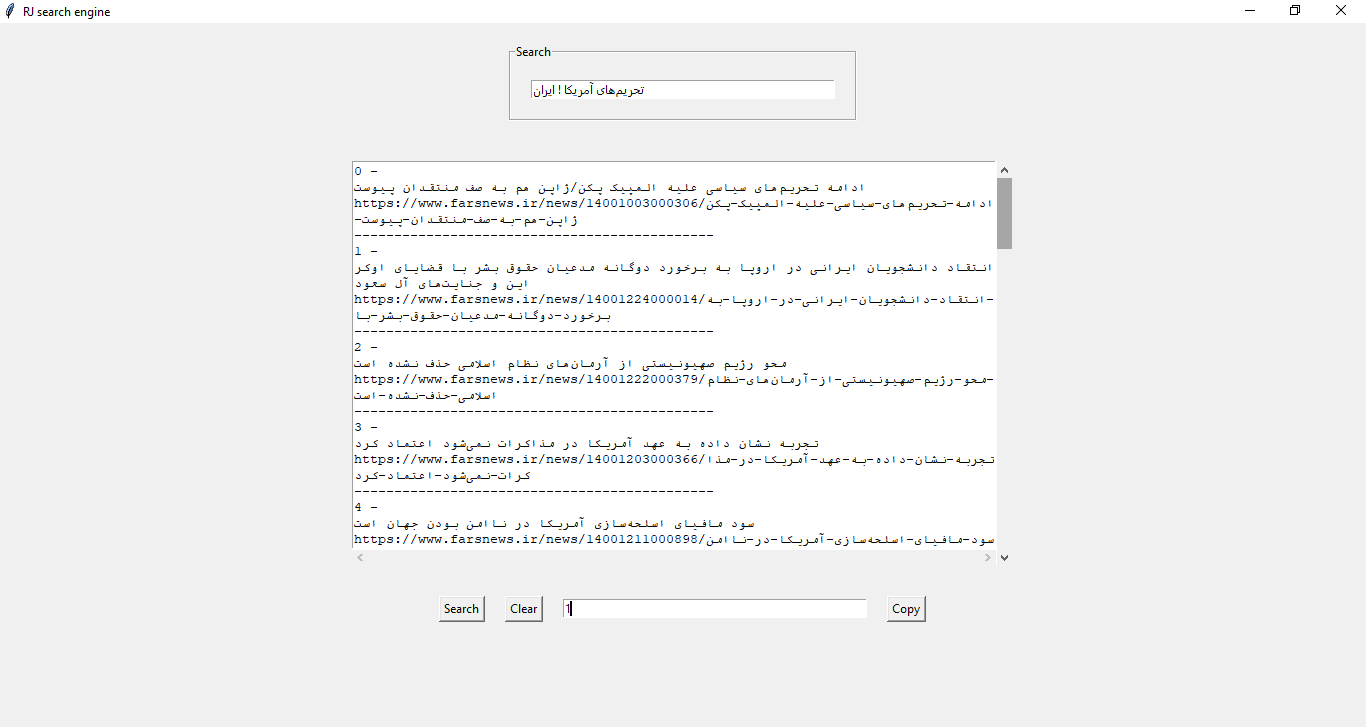
سند چهارم که به پرسمان کار بر مرتبط است و کلمات در تصویر هایلایت شده اند :



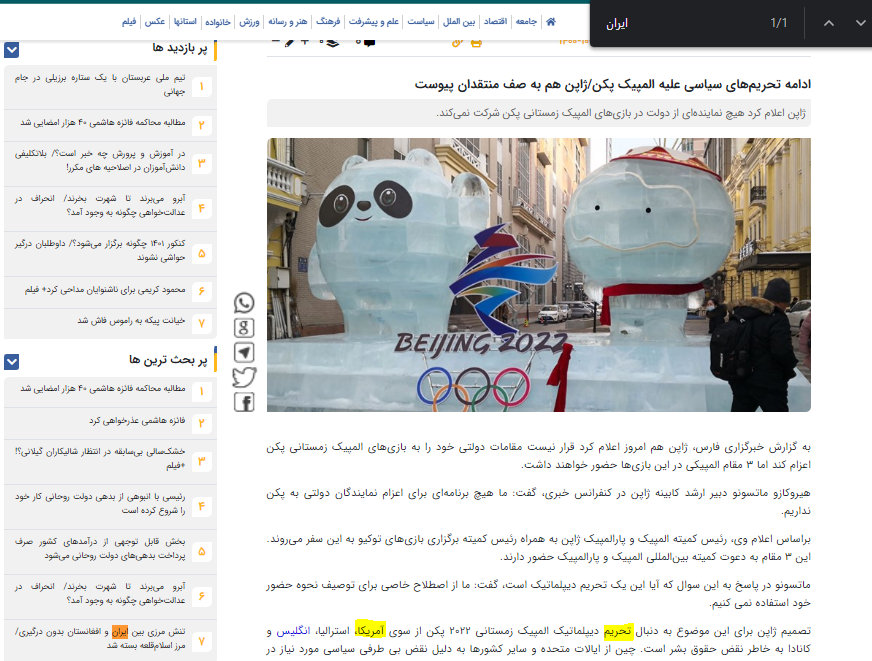
سند پنجم که به پرسمان کار بر مرتبط است و کلمات در تصویر هایلایت شده اند :



**ب ) تحریم‌های آمریکا ! ایران :**



سند اول که به پرسمان کار بر مرتبط است و کلمات در تصویر هایلایت شده اند و ایران در متن نیست :



سند دوم که به پرسمان کار بر مرتبط است و کلمات در تصویر هایلایت شده اند و ایران در متن نیست :



سند سوم که به پرسمان کار بر مرتبط است و کلمات در تصویر هایلایت شده اند و ایران در متن نیست : متن شرایط لازم را دارد اما در عمل (تحریم آمریکا بر علیه ایران است) .



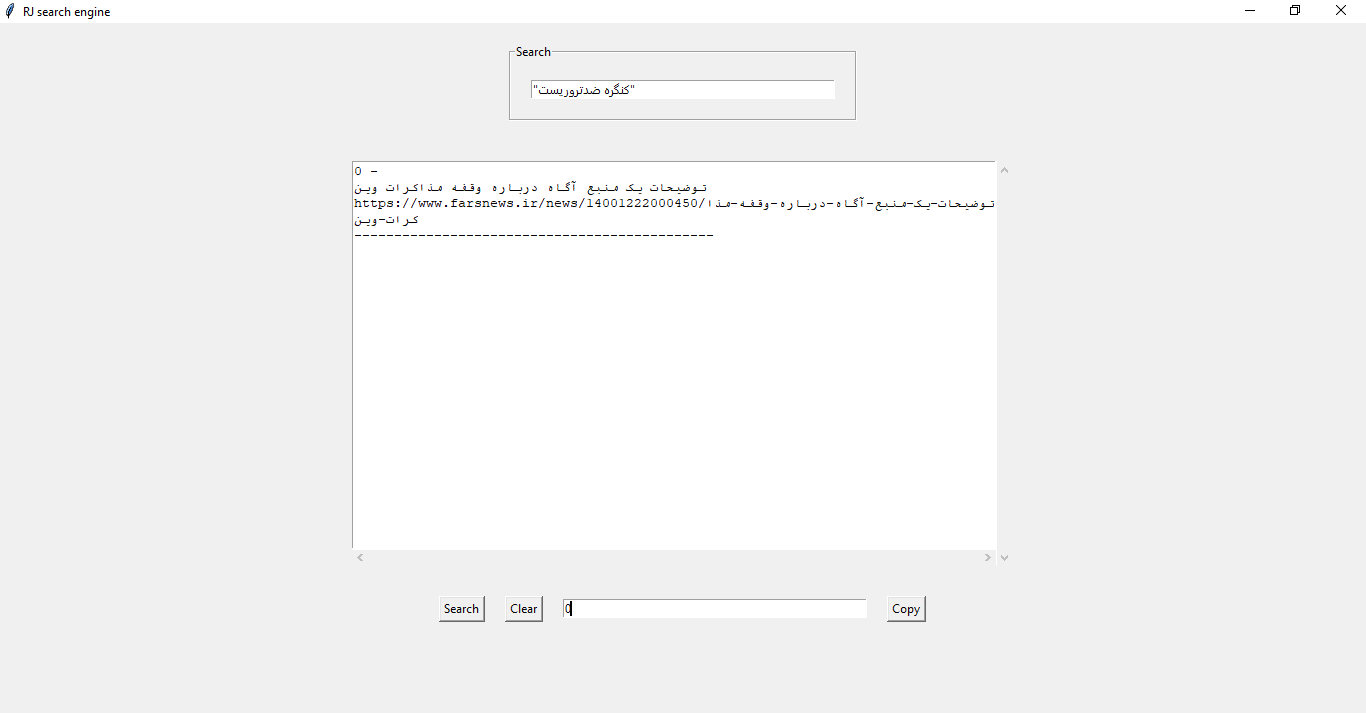
سند چهارم که به پرسمان کار بر مرتبط است و کلمات در تصویر هایلایت شده اند و ایران در متن نیست : متن شرایط لازم را دارد اما در عمل (تحریم آمریکا بر علیه ایران است) . توضیحات ستاره در سوال 4 مورد اول آمده است.



سند پنجم که به پرسمان کار بر مرتبط است و کلمات در تصویر هایلایت شده اند و ایران در متن نیست : متن شرایط لازم را دارد اما در عمل (تحریم آمریکا بر علیه ایران است) .



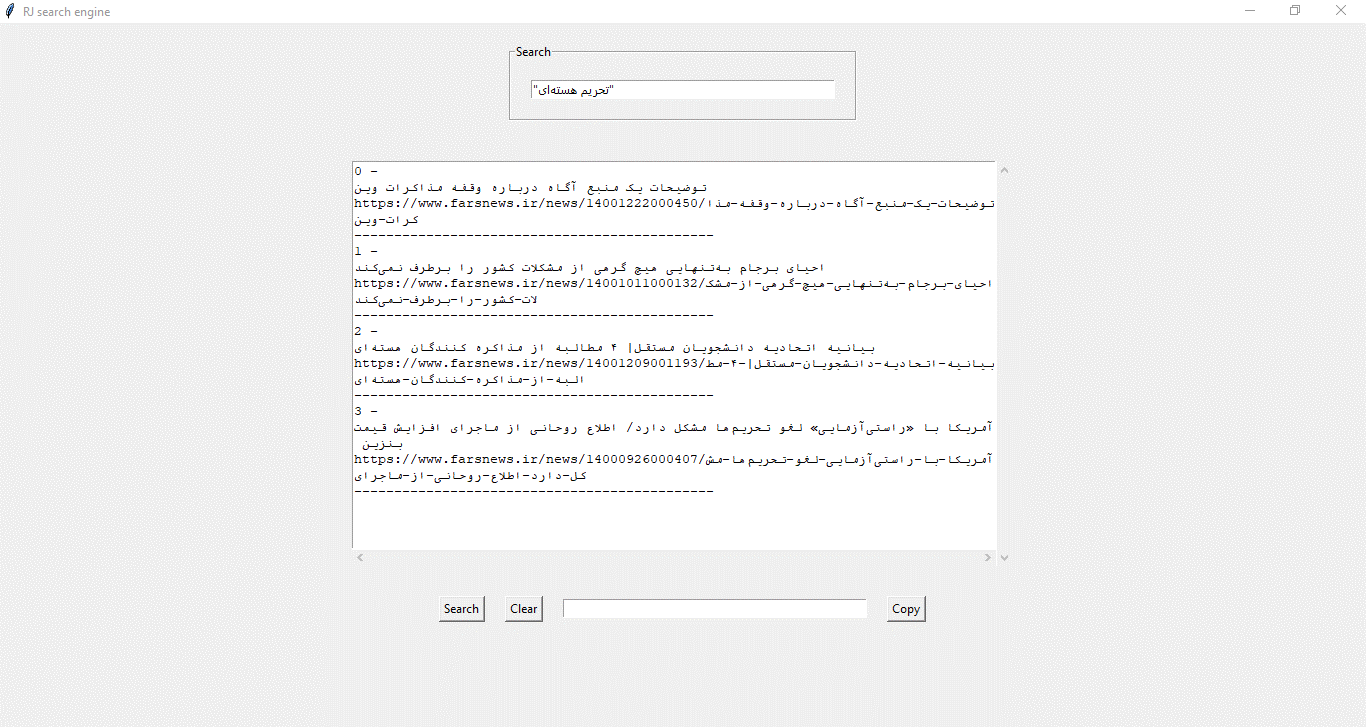
**پ ) "کنگره ضدتروریست" :**

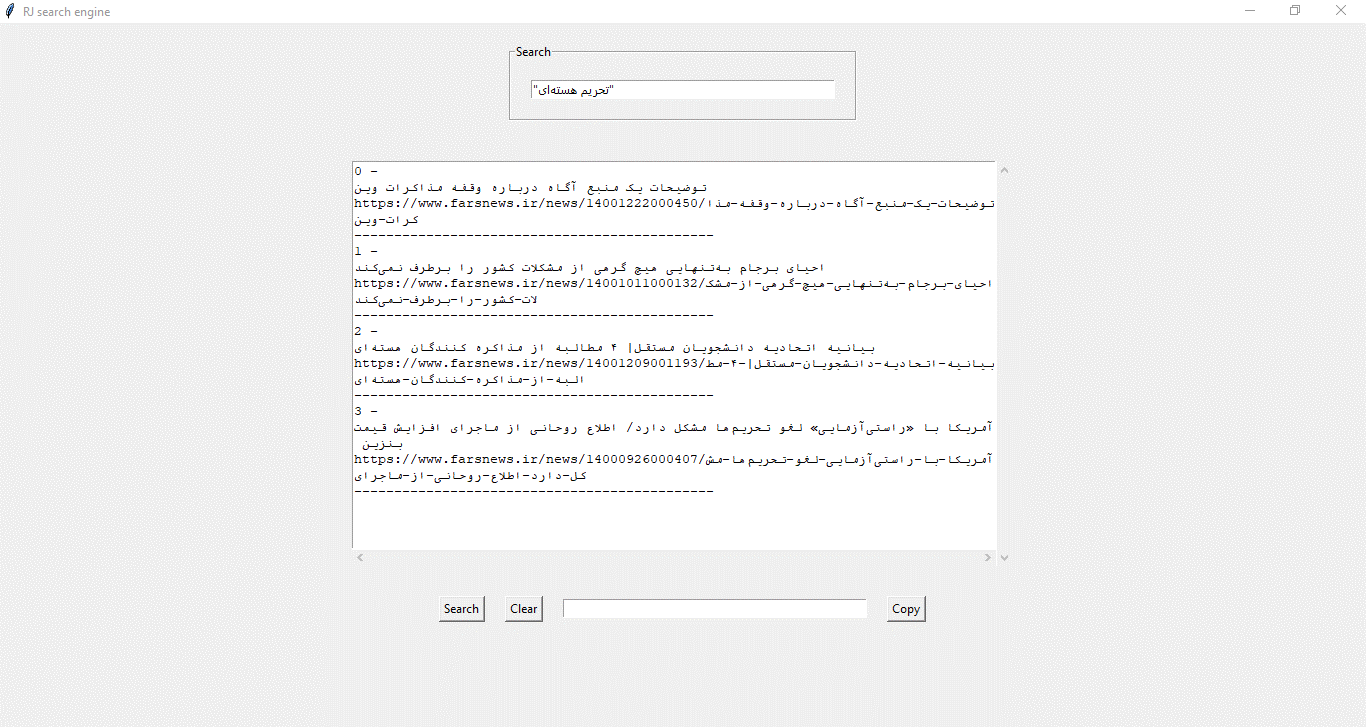


شامل یک سند است که :

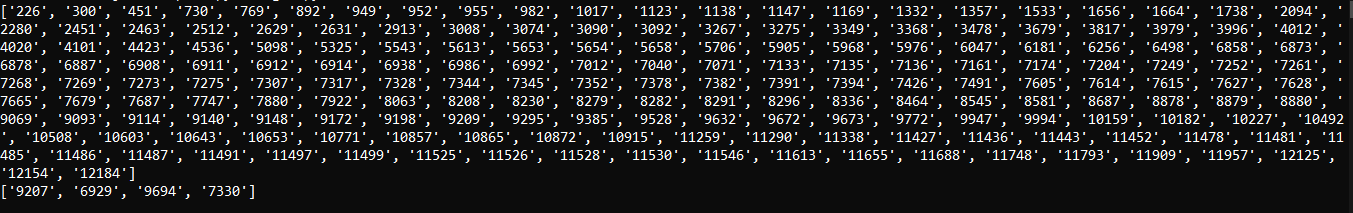


**ت ) "تحریم هستهای" آمریکا ! ایران :**

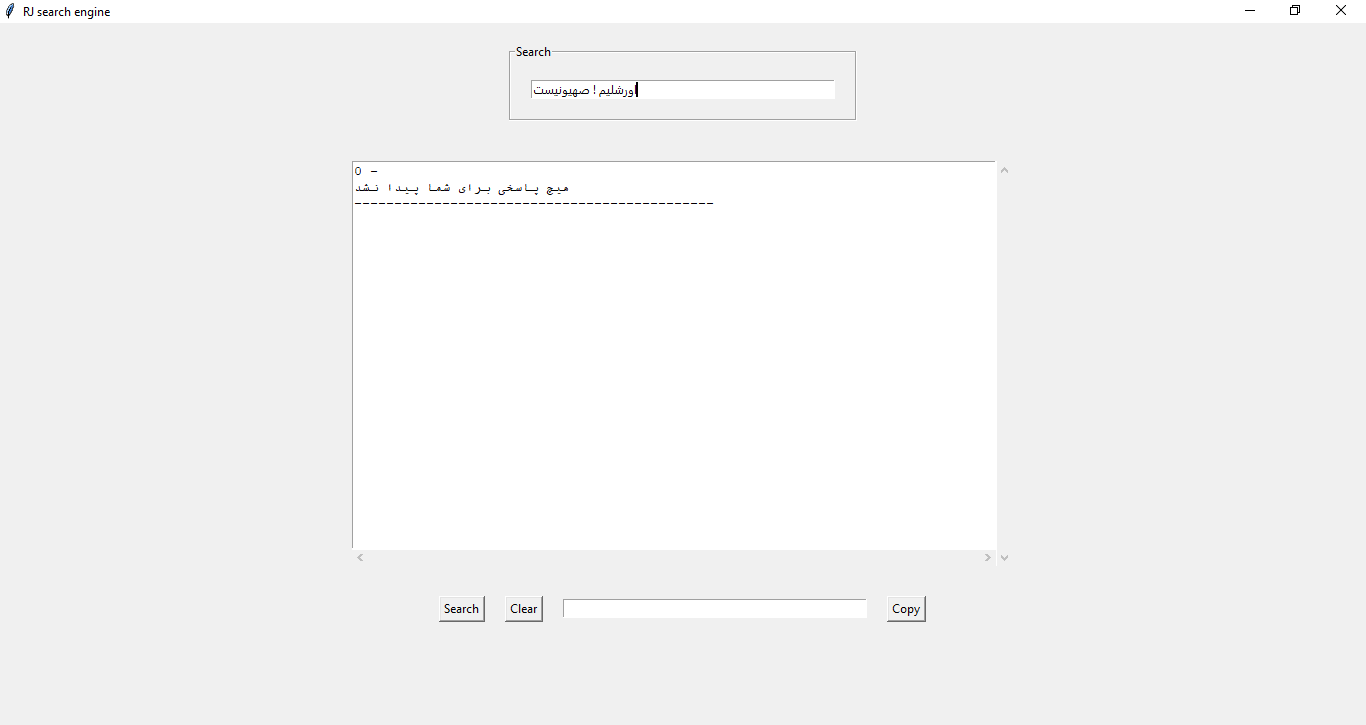
 

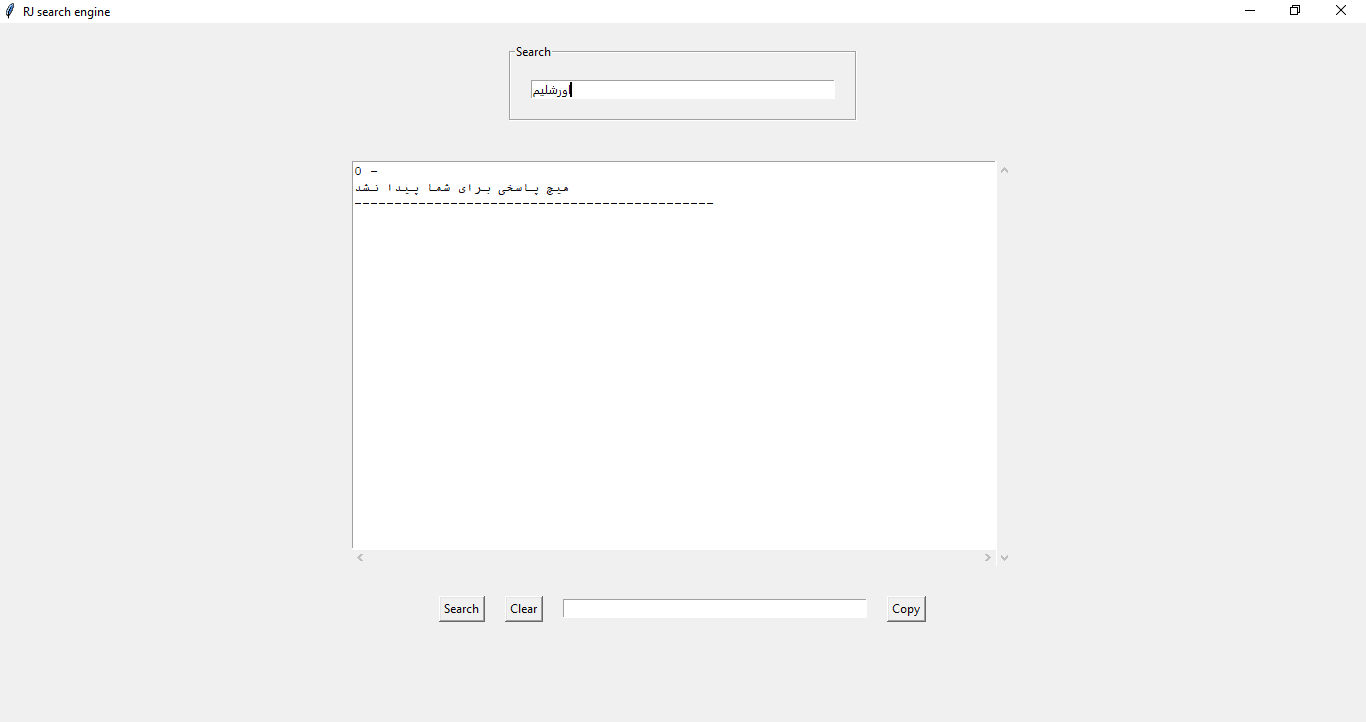


همانطور که میبینم 187 سند با این ویژگی که آمریکا دارند ولی ایران ندارند یافت شد و 4 سند هم با phrase query تحریمه هسته ای یافت شد اما اشتراکی بین آنها نبود



**ث ) اورشلیم ! صهیونیست :**





به طور کلی کلمه اورشلیم وجود ندارد .