

به نام خدا

حامد ثنائی - ۹۸۴۴۳۰۵۱

پروژه شماره ۲

درس بازیابی اطلاعات

توجه:

- ورژن نسخه پایتون ۳.۸.۲ می‌باشد.
 - فایل‌های دیتا برای این دو پروژه در کنار آن‌ها قرار داده شده است.
 - برای فریم ورک الاستیک سرچ از آخرین نسخه ی آن استفاده نشده است. با توجه به استفاده از پارسی آنالیزر از نسخه ۷.۴.۰ الاستیک استفاده شده است.
 - برای دریافت آخرین ویرایش فایل مربوط به پروژه :
- `git clone https://github.com/HamedSanaei/py-elasticsearch-parsianalyzer.git`

در اولین قدم، الاستیک سرچ را با دستورات زیر اجرا میکنیم:

```
cd elasticsearch-7.4.0/bin
./elasticsearch
```

سپس باید اطلاعات مربوط به خبرها را کراال کنیم . کد کراال کردن در فایل webscraper.py وجود دارد. اطلاعات از قبل بازیابی شده‌اند و در پوشه دیتا و همچنین در ریپازیتوری گیت‌هاب پروژه موجود است. لازم به ذکر است فرایند آن حدود دو ساعت و نیم طول می‌کشد. پس پیشنهاد می‌شود پس از تست کارایی کراال اجرای برنامه را متوقف کنید و بدون کراال شروع به ایندکس کردن و سرچ کنید.

فایب خبرها:

لینک گیت‌هاب:

https://github.com/HamedSanaei/py-elasticsearch-parsianalyzer/blob/master/Data/news_data.json

لینک meganz:

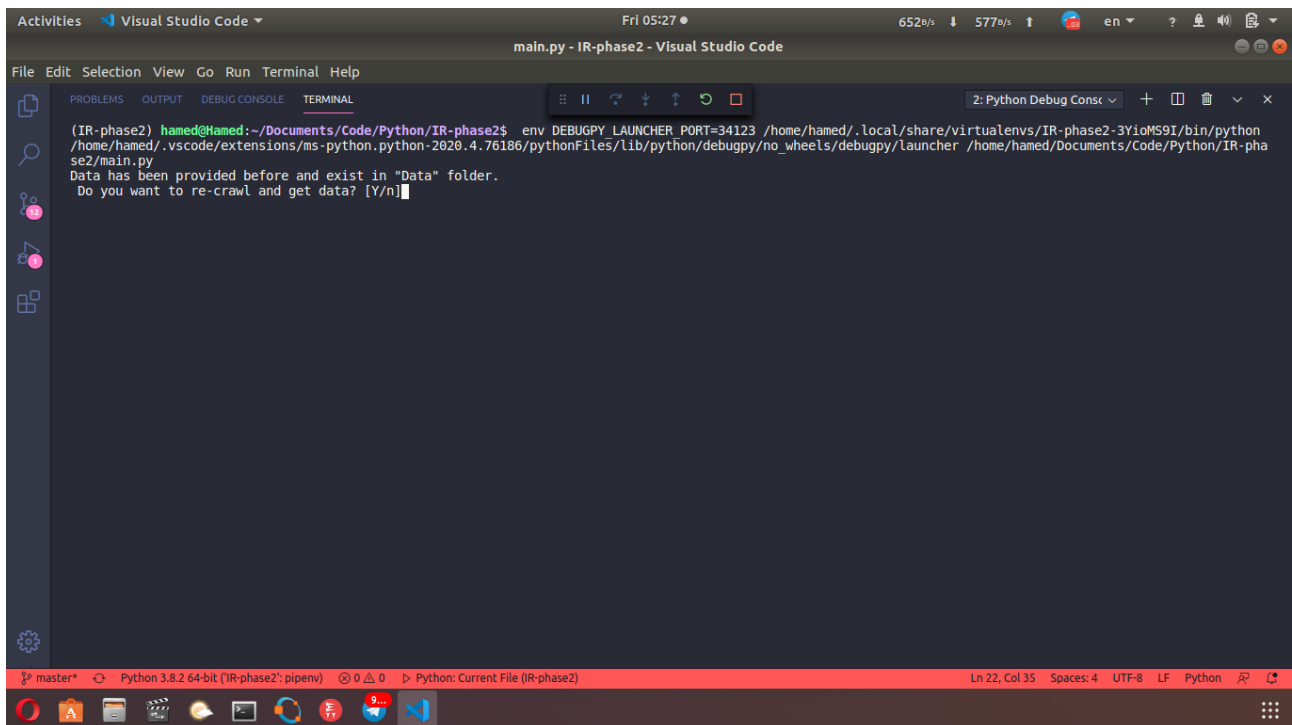
https://mega.nz/file/JVhhATyS#EveU2TTouVpd_KAr9z4-AUhRJPIIebg5AJtzCfiwquA

کد ایندکس برنامه و همچنین کوئری و نقطه ورود به برنامه به ترتیب در elastic_query.py ، elastic_index.py و main.py وجود دارد.

برای اجرای برنامه دستورات موجود در فایل Readme.md که در ریپازیتوری پروژه وجود دارد را قدم به قدم اجرا کنید. در اجرای این برنامه از pipenv استفاده کردم و نصب پکیج‌های مرتبط استفاده شده توسط شما به راحتی ای که در فایل Readme.md توضیح داده شده قابل انجام است.

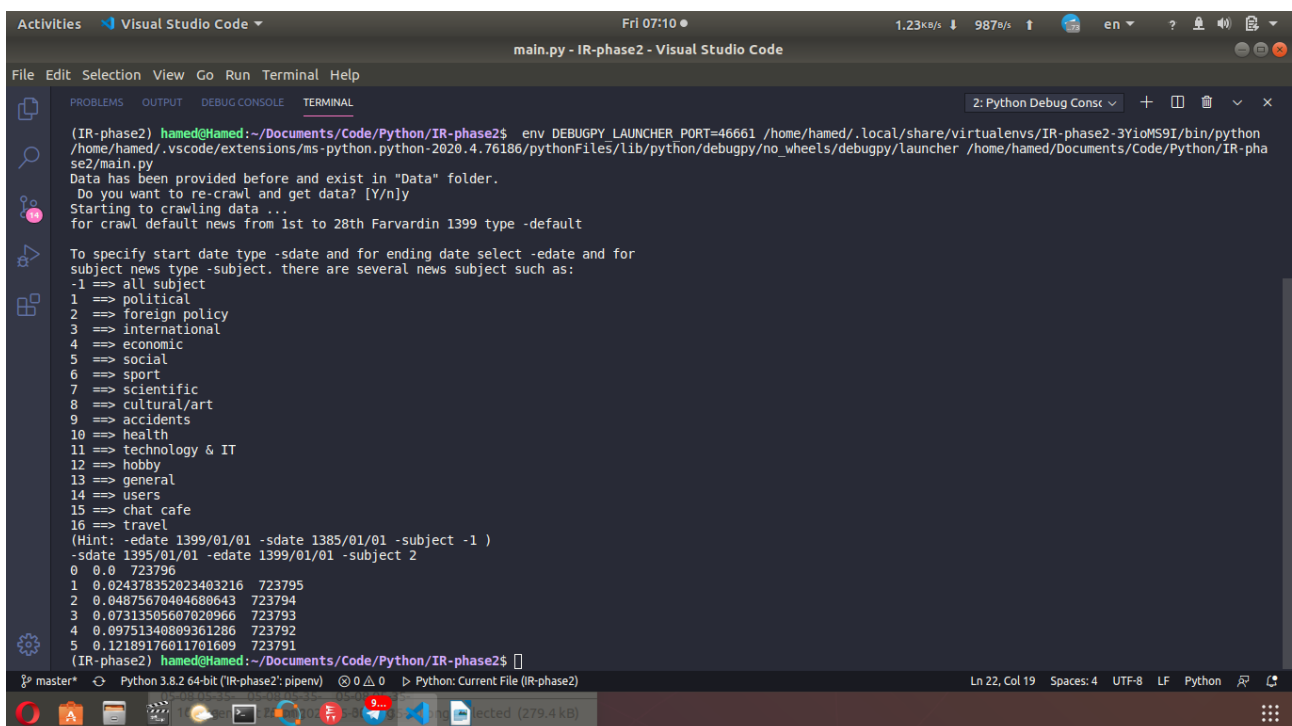
برای هر کوئری تعدادی ترم پیشنهاد شده است و روش مورد استفاده من ساده‌ترین روش پیشنهاد کوئری بود و روش‌های دیگر برای زبان فارسی جوابگو نبودند. هر ترکیب ممکن دیگر انتخاب برای ستاپ الاستیک اجرا شد و نتیجه‌ی مورد نظر گرفته نشد. بیشترین وقت بنده هم صرف همین موضوع شد.

بعد از اجرای برنامه از شما برای دوباره کراال کردن اطلاعات سؤال پرسیده می‌شود. مانند تصویر زیر:



```
(IR-phase2) hamed@Hamed:~/Documents/Code/Python/IR-phase2$ env DEBUGPY_LAUNCHER_PORT=34123 /home/hamed/.local/share/virtualenvs/IR-phase2-3YioMS9I/bin/python /home/hamed/.vscode/extensions/ms-python.python-2020.4.76186/pythonFiles/lib/python/debugpy/no_wheels/debugpy/launcher /home/hamed/Documents/Code/Python/IR-phase2/main.py
Data has been provided before and exist in "Data" folder.
Do you want to re-crawl and get data? [Y/n]
```

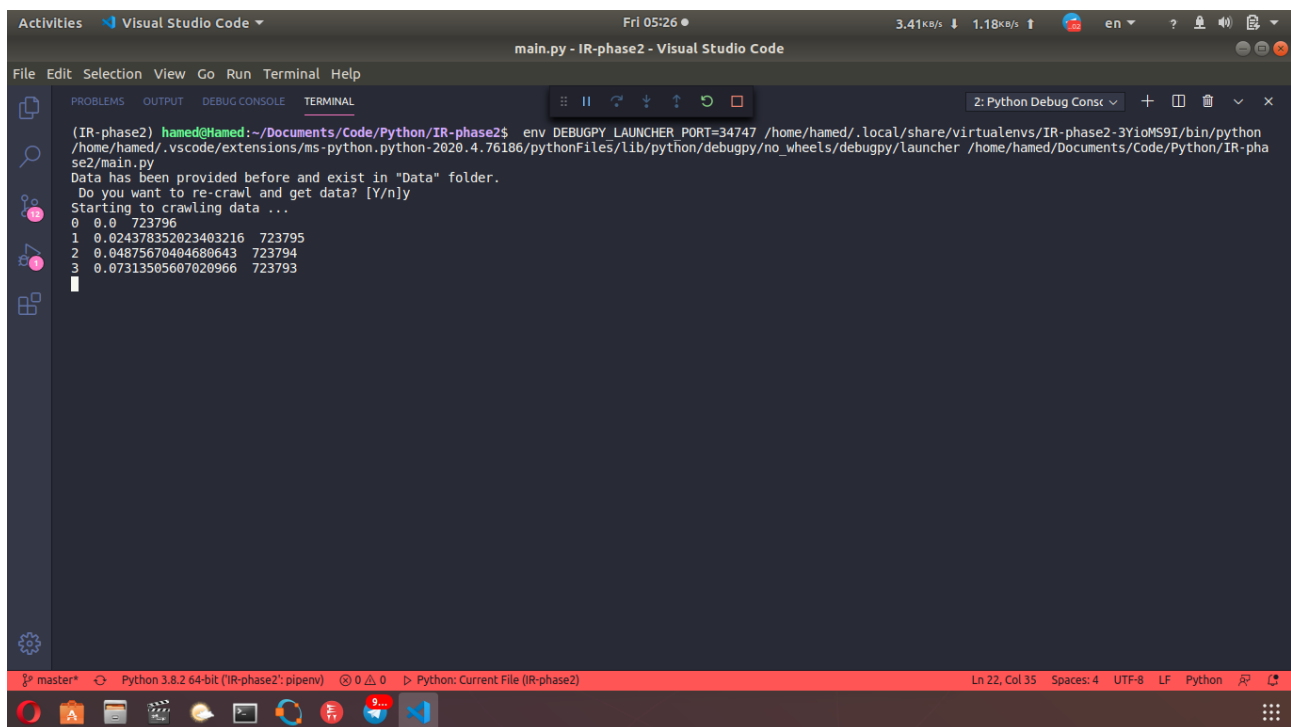
در اینجا موضوع خبر و تاریخ ابتدایی و انتهایی بازه تاریخی مورد جستجو را وارد میکنیم:



```
(IR-phase2) hamed@Hamed:~/Documents/Code/Python/IR-phase2$ env DEBUGPY_LAUNCHER_PORT=46661 /home/hamed/.local/share/virtualenvs/IR-phase2-3YioMS9I/bin/python /home/hamed/.vscode/extensions/ms-python.python-2020.4.76186/pythonFiles/lib/python/debugpy/no_wheels/debugpy/launcher /home/hamed/Documents/Code/Python/IR-phase2/main.py
Data has been provided before and exist in "Data" folder.
Do you want to re-crawl and get data? [Y/n]
Starting to crawling data ...
for crawl default news from 1st to 28th Farvardin 1399 type -default

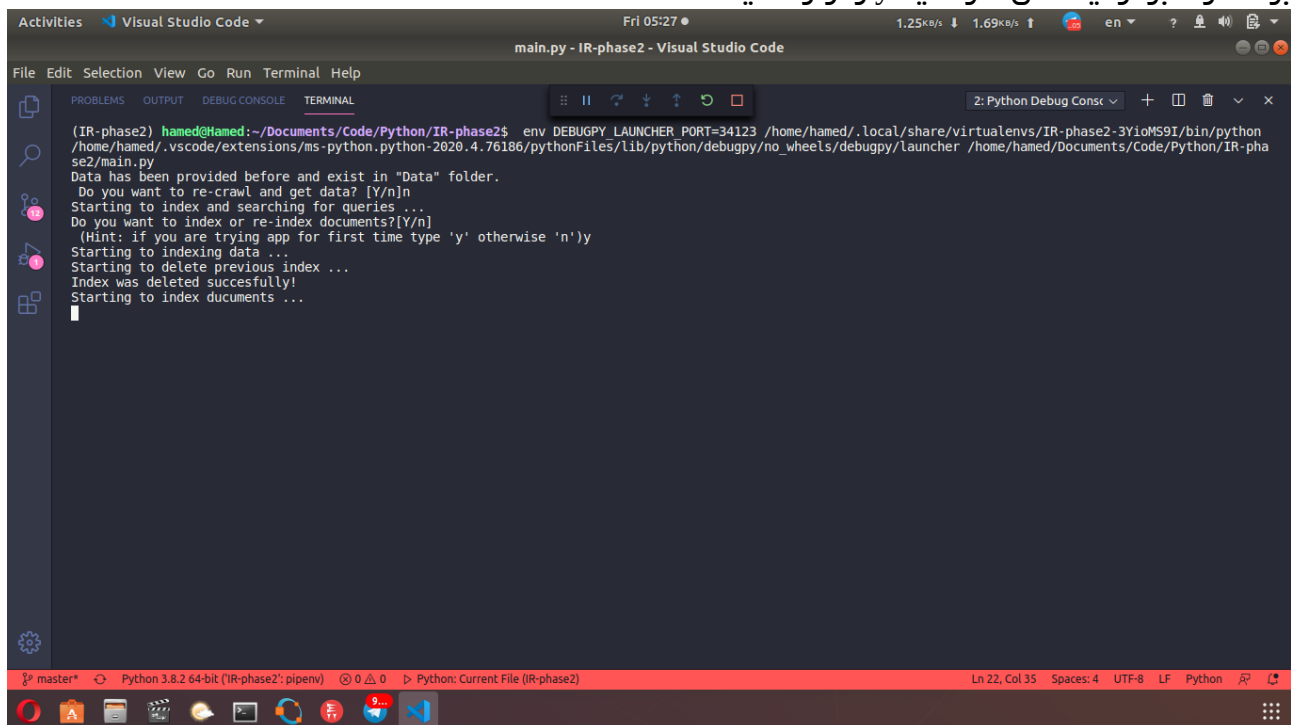
To specify start date type -sdate and for ending date select -edate and for
subject news type -subject. there are several news subject such as:
-1 ==> all subject
1 ==> political
2 ==> foreign policy
3 ==> international
4 ==> economic
5 ==> social
6 ==> sport
7 ==> scientific
8 ==> cultural/art
9 ==> accidents
10 ==> health
11 ==> technology & IT
12 ==> hobby
13 ==> general
14 ==> users
15 ==> chat cafe
16 ==> travel
(Hint: -edate 1399/01/01 -sdate 1385/01/01 -subject -1 )
-sdate 1395/01/01 -edate 1399/01/01 -subject 2
0 0.0 723796
1 0.024378352023403216 723795
2 0.04875670404690643 723794
3 0.07213505607020966 723793
4 0.09751340809361286 723792
5 0.12189176011761609 723791
(IR-phase2) hamed@Hamed:~/Documents/Code/Python/IR-phase2$
```

در تصویر زیر برنامه شروع به کرال کردن دیتا میکند:



```
(IR-phase2) hamed@Hamed:~/Documents/Code/Python/IR-phase2$ env DEBUGPY_LAUNCHER_PORT=34747 /home/hamed/.local/share/virtualenvs/IR-phase2-3YioMS9I/bin/python /home/hamed/.vscode/extensions/ms-python.python-2020.4.76186/pythonFiles/lib/python/debugpy/no_wheels/debugpy/launcher /home/hamed/Documents/Code/Python/IR-phase2/main.py
Data has been provided before and exist in "Data" folder.
Do you want to re-crawl and get data? [Y/n]y
Starting to crawling data ...
0 0.0 723796
1 0.024378352023403216 723795
2 0.04875670404680643 723794
3 0.07313505607020966 723793
```

بعد از اتمام یا توقف این کار از شما سؤال پرسیده می‌شود که می‌خواهید دیتا را ایندکس کنید یا نه؟ اگر قبلاً برنامه را اجرا و ایندکس نکرده‌اید y را وارد کنید.



```
(IR-phase2) hamed@Hamed:~/Documents/Code/Python/IR-phase2$ env DEBUGPY_LAUNCHER_PORT=34123 /home/hamed/.local/share/virtualenvs/IR-phase2-3YioMS9I/bin/python /home/hamed/.vscode/extensions/ms-python.python-2020.4.76186/pythonFiles/lib/python/debugpy/no_wheels/debugpy/launcher /home/hamed/Documents/Code/Python/IR-phase2/main.py
Data has been provided before and exist in "Data" folder.
Do you want to re-crawl and get data? [Y/n]n
Starting to index and searching for queries ...
Do you want to index or re-index documents?[Y/n]y
(Hint: if you are trying app for first time type 'y' otherwise 'n')y
Starting to indexing data ...
Starting to delete previous index ...
Index was deleted successfully!
Starting to index documents ...
```

سپس شروع به جستجو میکنیم و در تصاویر زیر پاسخ سؤالات را به ترتیب میبینیم:

The image displays a Visual Studio Code editor window with a Python script named 'main.py' open. The script is designed to search for news on the Asriran website using BeautifulSoup and requests libraries. The terminal output shows the results of two searches: one for 'کرونا' (Corona) and another for 'محدودیت‌ها' (Restrictions). The interface includes the standard VS Code menu bar, toolbar, and sidebar with Explorer, Search, and Run and Debug views. The status bar at the bottom indicates the file is 'main.py - IR-phase2 - Visual Studio Code' and shows various settings like line length and encoding.

The screenshot shows the Visual Studio Code interface with the main editor displaying search results from a web crawler. The status bar at the bottom indicates the file is named "master*", the Python version is 3.8.2 64-bit, and the current file is "IR-phase2". The search results are organized into three sections, each starting with a search query and followed by a list of hits.

Visual Studio Code interface showing search results for COVID-19 related news in Persian.

The status bar at the top displays:

- Activities
- Visual Studio Code
- Fri 05:34
- 825B/s ↓ 768B/s ↑
- en ?

The main editor area shows the following search results:

Search for: بیش از 118 میلیارد دلار (67 میلیارد پوند)
This is answer to question 5
Suggestion per term:
Term: پوند Suggestions: ['پسوند', 'پرند', 'بسنده', 'پونه', 'پسونده']
Got 478 Hits:
Title: تازه‌ترین رتبه بندی ثروتمندان جهان News_url: asriran.com/0031al Score: 16.259216
Title: فشار کرونا بر صنعت مد و پوشاک آسیا / آيا لباس گران موفود؟ News_url: asriran.com/00322n Score: 15.782657
Title: بولداترین افراد دنيا در کدام کشورها هستند؟ News_url: asriran.com/003liB Score: 15.658424
Title: آمريکا یک ميليارد دلار از کمکهایش به نبروهای امنیتی افغانستان را کاهش میدهد News_url: asriran.com/003lqr Score: 15.304822
Title: (عکس) خبره اقتصادی کرونا به جن شکوفه‌های گلبي در رایي News_url: asriran.com/003IDL Score: 15.280392
Title: مقامهای آمریکاين: ضمانت گرفته ایم ۱.۶ ميليارد دلار قابل انتقال به ايران نباشد News_url: asriran.com/0032Cs Score: 14.6871395
Title: بسته 2 تریلیون دلاری ترامپ برای نجات آمریکا از بحران کرونا News_url: asriran.com/0031QN Score: 14.386492
Title: رئیس دیوان محاسبات: حقوق ۵۳ میلیون تومان يی یک مدیر/تحلفات در عملکرد ارز ۲۲۰۰ تومان يی اساس News_url: asriran.com/0032BF Score: 14.305501
Title: حقوق ۵۳ میلیون تومان يی یک مدیر/تحلفات در عملکرد ارز ۲۲۰۰ تومان يی News_url: asriran.com/0032B2 Score: 14.305501
Title: کاهش درآمد ۷۶ ميليارد دلاری ايران‌های اروپايي News_url: asriran.com/0031V5 Score: 14.244057

Search for: آمايت مهتابی
This is answer to question 6
Suggestion per term:
Term: مهتابی Suggestions: ['مهتابي', 'مهتابم', 'مهتابي', 'مهتابي', 'مهتابي']
Got 1 Hits:
Title: فشار کرونا بر صنعت مد و پوشاک آسیا / آيا لباس گران موفود؟ News_url: asriran.com/00322n Score: 11.713867

Search for: سربانكا
This is answer to question 7
Suggestion per term:
Term: سربانكا Suggestions: ['سرانكا\۲00c\سر']
Got 10 Hits:
Title: (ورود حیوانات به شهرهای تعطیل شده از کرونا (فیلم News_url: asriran.com/00329P Score: 8.108499
Title: دیدگاه‌های جدید: از قیظنه‌های اجاره تا فایله‌گذاری اجتماعی News_url: asriran.com/0031S3 Score: 6.6953998

The status bar at the bottom displays:

- master*
- Python 3.8.2 64-bit (IR-phase2: pipenv)
- Python: Current File (IR-phase2)
- Ln 72, Col 1
- Spaces: 4
- UTF-8
- LF
- Python

Activities Visual Studio Code

main.py - IR-phase2 - Visual Studio Code

File Edit Selection View Go Run Terminal Help

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL

Code

Search for: سرلانکا

This is answer to question 7

Suggestion per term:

Term: سرلانکا Suggestions: ['لانکا u200cاسرى']

Got 10 Hits:

Title: (ورود حیوانات به شهرهای تعطیل شده از کرونا (فیلم) News_url: asriran.com/00329P Score: 8.108499

Title: دیدنیهای امروز: از فرستندههای اخبار تا فاصله‌گذاری اجتماعی News_url: asriran.com/003153 Score: 6.695398

Title: 54 هزار میلیارد ریال کالای قاچاق در سال 98 در گمرک News_url: asriran.com/0031Nm Score: 6.0890236

Title: آخرین آمار ابتلا به کرونا در جهان News_url: asriran.com/0031oS Score: 4.964896

Title: دیپلماسی ماسک: چین به دنبال بهبود تصویر خود News_url: asriran.com/0031G1 Score: 3.9452112

Title: آخرین آمار/ شمار تلفات کروناویروس در جهان ۱۱ هزار و ۸۲۲ نفر کشته News_url: asriran.com/0031G2 Score: 3.5308344

Title: آخرین آمار/ شمار تلفات کروناویروس در جهان ۱۱ هزار و ۸۲۲ نفر کشته News_url: asriran.com/0032Ff Score: 3.3546598

Title: آخرین آمار ابتلا به کرونا در جهان News_url: asriran.com/0032I5 Score: 3.3546598

Title: جدیدترین آمارها از کرونا در جهان/ 192 کشور درگیر کرونا News_url: asriran.com/0031L5 Score: 3.3546598

Title: آخرین آمار ۱۸۸ کشور و منطقه درگیر با کروناویروس/ ۳۱۶ هزار و ۶۵ مبتلا News_url: asriran.com/0031IX Score: 3.3546598

Search for: ۱۲

This is answer to question 8

Suggestion per term:

Got 479 Hits:

Title: پیشنهاد استفاده اخباری از ماسک در مترو در شد News_url: asriran.com/0032B7 Score: 2.8963633

Title: (سکانس) سانسور شده از تونج : شوخی با صمد آقا (فیلم) News_url: asriran.com/0032B5 Score: 2.8963633

Title: روحانی: لغو ممنوعیت تردد بین شهرستانها از امروز/ آزادی تردد میان استان ها از اول اردیبهشت News_url: asriran.com/00325q Score: 2.8963633

Title: به بوش «بامحاطالمن حوبه» ببینید: News_url: asriran.com/0031ea Score: 2.8963633

Title: توجیه انگلیس برای مخالفت با لغو تحریمهای ایران در بنجوتو News_url: asriran.com/0031e5 Score: 2.8963633

Title: سخیوی وزارت خارجه: بریلیونها دلار که خرج دجاله در خارمینان شد منبواسه جرح مردم آمریکا شود News_url: asriran.com/0031e3 Score: 2.8963633

Title: وقتی ما هم یک سندرم برقرار داریم News_url: asriran.com/0031e2 Score: 2.8963633

Title: کمک فوری برای زنده ماندن سیده زهرا که به بنجوت کلیه نیاز دارد News_url: asriran.com/0031e1 Score: 2.8963633

Title: هشدار سازمان بهداشت جهانی نسبت به شیوع مجدد کرونا در آسیا News_url: asriran.com/0031e0 Score: 2.8963633

Python 3.8.2 64-bit (IR-phase2: pipenv) Python: Current File (IR-phase2)

Ln 72, Col 1 Spaces: 4 UTF-8 LF Python

Visual Studio Code interface showing a search for "Search for: 13". The results list news articles from asriran.com, including titles like "ایران پیشرفت چشمگیری در مبارزه با کرونا داشته است" and "احتمال آغاز لیگ برتر فوتبال از هفته دوم خرداد". The interface also shows the status bar with "Python 3.8.2 64-bit (IR-phase2: pipenv)" and "Ln 72, Col 1".

Visual Studio Code interface showing a search for "Search for: 11". The results list news articles from asriran.com, including titles like "تعلیلی مدارس نیویورک تا پایان سال تحصیلی" and "تعلیلی مدارس نیویورک تا پایان سال تحصیلی". The interface also shows the status bar with "Python 3.8.2 64-bit (IR-phase2: pipenv)" and "Ln 72, Col 1".

