

به نام خدا
دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر و فناوری اطلاعات



عنوان: گزارش مرحله اول پروژه

اعضای گروه:
مصطفی معصومی ۹۵۳۱۰۷۹
امیرحسین قندهاری ۹۵۳۱۰۶۹
محمد عساری ۹۳۳۱۷۰۳

استاد درس:
دکتر نیک آبادی

درس:
بازیابی اطلاعات

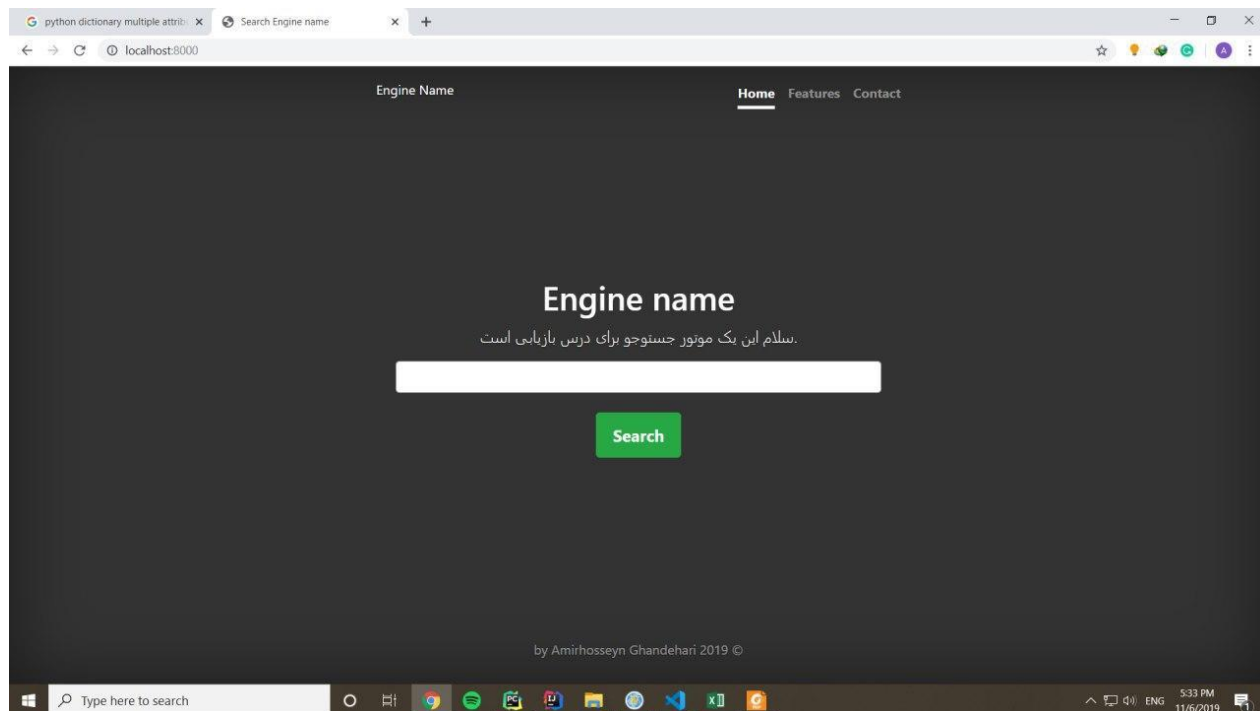
آبان ۱۳۹۸

زبان‌های برنامه نویسی

پروژه ی ما تحت وب می باشد که قسمت front-end پروژه با استفاده از bootstrap پیاده سازی شده و نیز قسمت back-end پروژه با استفاده از فریم ورک django توسعه یافته است.

نحوه ی کار با برنامه (واسط کاربری)

با اجرای برنامه، پروژه بر روی localhost و پورت ۸۰۰۰ بالا می‌آید و منتظر گرفتن query می‌باشد. مطابق شکل زیر:




با وارد کردن query مورد نظر متن به قسمت back-end می‌رود و پس از پردازش، آیدی document هایی را که دارای ویژگی‌های query می‌باشد را طبق مکانیزمی که در قسمت back-end توضیح خواهیم داد، بازیابی کرده و بر می‌گرداند.

سپس اطلاعات هر document به کاربر نمایش داده می‌شود که اگر تعداد document ها بیش تر از ۲۰ عدد باشد، اطلاعات صفحه بندی می‌شود. مطابق شکل زیر:

Search

ایران و روسیه | خلیج

Contact Features Home Engine Name




یادگارهای برجلم: تحریم‌های ویزا، کاناسا و اینسا/ کمک ویژه دولت برای مالک شدن اهالی مسکن مهر/ آقای پمپلو! دروغگو کم‌حافظه است

mashreghnews.ir

September 8th 2019, 13:03:00.000

آخرین مواضع وزیر امور خارجه در کنار افزایش تحریم‌های ضد ایرانی توسط آمریکا نشان می‌دهد مذاکرات چند ماه اخیر او در نیویورک و پاریس، اقدامی نابجا و موجب زمان‌سواری و غفلت از نقشه حریف بوده است.




بازگشت پورتوریکو برابر ایران از بهترین های تاریخ جام جهانی بسکتبال

isna.ir

September 1st 2019, 16:05:00.000

تیم ملی بسکتبال پورتوریکو توانست در کوارتر چهارم اختلاف ۱۴ امتیازی را برابر ایران جبران کند و یکی از بهترین بازگشت‌های تاریخ جام جهانی را به نام خود ثبت کند.




خدمه نفتکش توقیف شده انگلیسی آزاد شدند7

borna.news

September 6th 2019, 15:54:47.000

نیروی دریایی مالک نفتکش توقیف‌شده انگلیس در ایران گفت: برای آزادی ۱۶ خدمه دیگر تلاش می‌کند.



رزمایش امنیت و اقتدار پایدار در دریای خزر برگزار می‌شود

ilina.ir


August 31st 2019, 19:17:10.000

فرمانده نیروی دریایی ارتش جمهوری اسلامی ایران گفت: رزمایش امنیت و اقتدار پایدار در دریای خزر بزودی برگزار می‌شود.

Search

ایران و روسیه

Contact Features Home Engine Name




دیدگاه «ایران و روسیه» در تأمین امنیت خلیج فارس بسیار نزدیک است

mehnews.com

September 2nd 2019, 20:54:00.000

وزیر امور خارجه گفت: دیدگاه ایران و روسیه در تأمین امنیت خلیج فارس بسیار به یکدیگر نزدیک است، ما معتقدیم امنیت از طریق همکاری بین کشورهای ساحلی ایجاد خواهد شد به در تقابل با یکدیگر.




یادگارهای برجلم: تحریم‌های ویزا، کاناسا و اینسا/ کمک ویژه دولت برای مالک شدن اهالی مسکن مهر/ آقای پمپلو! دروغگو کم‌حافظه است

mashreghnews.ir

September 8th 2019, 13:03:00.000

آخرین مواضع وزیر امور خارجه در کنار افزایش تحریم‌های ضد ایرانی توسط آمریکا نشان می‌دهد مذاکرات چند ماه اخیر او در نیویورک و پاریس، اقدامی نابجا و موجب زمان‌سواری و غفلت از نقشه حریف بوده است.




بازگشت پورتوریکو برابر ایران از بهترین های تاریخ جام جهانی بسکتبال

isna.ir

September 1st 2019, 16:05:00.000

تیم ملی بسکتبال پورتوریکو توانست در کوارتر چهارم اختلاف ۱۴ امتیازی را برابر ایران جبران کند و یکی از بهترین بازگشت‌های تاریخ جام جهانی را به نام خود ثبت کند.



خدمه نفتکش توقیف شده انگلیسی آزاد شدند7

borna.news

September 6th 2019, 15:54:47.000


نیروی دریایی مالک نفتکش توقیف‌شده انگلیس در ایران گفت: برای آزادی ۱۶ خدمه دیگر تلاش می‌کند.

python dictionary multiple attri...

Search Engine name

+

localhost:8000/search/?search=




شبهستان

shabestan.ir

September 1st 2019, 02:46:00.000

اینکار با بیان اینکه از کمن های به به پلاستیک، ردیای آب و کاهش پسماند در نوانند پروژه و طرح های خودهای مردم نهاد میکنم، گفت: ننگامور باتوان و خانواده حمایت می را در این خصوص ارائه




درصد مطالبات شهرداری اردستان از دستگاه‌های دولتی است ۷۰

mehnews.com

September 1st 2019, 20:47:00.000

اردستان- سرپرست شهرداری اردستان گفت: مطالبات شهرداری اردستان از دستگاه‌ها و انتخاص حدود سه میلیارد ریال است که بیش از ۷۰ درصد این مطالبات از طریق نهادهای اداری است.




سازمانهای نروزیستی مسئولیت کیفری دارند

mehnews.com

September 1st 2019, 00:55:00.000

رئیس قوه قضایه گفت: سازمانهای نروزیستی مسئولیت کیفری دارند و پیگیری حقوقی اقدامات نروزیستها در دستور کار است.



برتری ترابوژان در حضور 90 دقیقه‌ای مدافع ایرانی

khabaronline.ir

August 26th 2019, 04:35:00.000

سید مجید حسینی در شب برد تیمش در لیگ ترکیه در ترکیب تیمش به میدان رفت.


Page 1 of 5.

next

last »

by Amirhosseyn Ghandehari 2019 ©

Type here to search



5:34 PM

11/6/2019

3

مدل بازیابی اطلاعات (back-end)

در ابتدا فایل اکسل را تبدیل به فایل csv کردیم زیرا کار کردن با آن با استفاده از کتابخانه‌هایی که در پروژه استفاده کردیم راحت‌تر است. سپس در فایل main.py با استفاده از کتابخانه HTMLParser قسمت متنی HTML مستنداتمان را استخراج کردیم. سپس در تابع handle_data با استفاده از لیستی از علائم نگارشی که خودمان تهیه کردیم، علائم نگارشی را از این متون پاک کردیم.

پس از آن با استفاده از کتابخانه ی parsivar بخش همسان سازی و ریشه‌یابی را انجام دادیم. برای ساختن دیکشنری نیز از دیکشنری پایتون (dict) استفاده کردیم به گونه‌ای که با خواندن فایل های HTML یک دیکشنری می‌سازد که key آن لغت می‌باشند و value آن یک دیکشنری دیگری است که به ازای هر المان key آن شماره ی id آن document است و نیز به عنوان value یک لیست داریم که به position های آن لغت در آن document را اشاره می کند.

پس از ساخته شدن این dictionary آن را با استفاده از کتابخانه ی pickle در یک فایل ذخیره می کنیم. برای واکنشی خبر نیز در فایل search.py که به قسمت front-end متصل است در ابتدا query را از کاربر می‌گیریم و آن را پارس می‌کنیم و سپس به stem لغات آن تبدیل می کنیم. در اینجا لغات query می‌توانند چهار حالت داشته باشند: یا یک لغت هستند، یا یک عبارت هستند، یا نقیض یک لغت هستند و یا نقیض یک عبارت می باشند. سپس با خواندن فایل مربوط به دیکشنری با استفاده از کتابخانه ی pickle برای هر کدام از این چهار مجموعه id های مربوط به document ها را پیدا می‌کنیم و اشتراک این مجموعه ها document هایی را می‌دهد که مطابق query می باشد. در انتها نیز لیست id این مجموعه به قسمت front-end داده می‌شود تا همانطور که در قسمت front-end توضیح دادیم خبرها ی مورد نظر نمایش داده شود.