ذخیره دیتا

در فاز اول پروژه، توانستیم با موفقیت سایت را هک کرده و وارد آن شویم.

کسانی که فاز اول پروژه را نزده اند می‌توانند نام کاربری و رمز عبور خود را از PDF داخل کانال ببینند. همچنین به دلیل هک شدن های عمده کپچا توسط دانشجویان علوم کامپیوتر 1400 دانشگاه تهران، مسئول سایت فعلا کپچا را حذف کرده است و نیازی به وارد کردن کپچا نیست.

در ادامه مصائب فاز اول پروژه، پس از وارد شدن در سایت با لیست عظیمی از دیتای خودروهای مختلف مواجه می‌شویم. نزدیک به 500 صفحه دیتا از ماشین های مختلف گران یا ارزان قیمت با کارکردهای مختلف و از اقصی نقاط ایران.(نزدیک به 17هزار آگهی!) می‌خواهیم تمام دیتای این ماشین‌ها را از سایت گرفته و در یک ماتریس ذخیره کنیم. نحوه ذخیره را با یک مثال توضیح می‌دهیم:

در عکس بالا به عنوان مثال 3 آگهی مختلف آمده است. این آگهی‌ها را به عنوان نمونه در ماتریس ذخیره میکنیم که در سطر اول آن نام ستون آمده و زیر هر ستون و در هر سطر اطلاعات هر کدام از ماشین ها. یک راه ذخیره کردن اطلاعات به صورت فارسی است:

اما در این روش ذخیره سازی در برخی IDEاها ممکن است به مشکل بخورید. همچنین انتقال دیتا به صورت فارسی کمی مشکل دار خواهد شد. پس کمی بیشتر HTML دیتا را نگاه کنید تا به اطلاعات انگلیسی ماشین دسترسی انجام دهید:

همچنین اگر دیتایی را به صورت انگلیسی نداشتید(مثلا توافقی بودن قیمت) خودتان دستی تعیین کنید که به جای این عبارت فارسی چه عبارت انگلیسی ای قرار بگیرد.(یا حتی می‌توانید قرارداد کنید که هرجا قیمت توافقی بود قیمت 0 یا -1 تعیین شود)

برای اینکار ابتدا باید متن HTML سایت را در قالب یک استرینگ دریافت کنید(که با دستور get به سادگی قابل انجام است) سپس با استفاده از لایبرری bs4 یا beautiful soup به تفکیک بخش های مختلف ماشین‌ها بپردازید.

فرض کنید سطر اول ماتریس(اطلاعات اولین آگهی) را استخراج کرده اید. این آگهی را در یک دیکشنری ذخیره می‌کنید. برای مثال:

```
{'company':'peugeot',
 'car':'206',
 'tream':'ir-type2',
 'kilometer':0,
 'year':1399,
 'price':'agreement'
}
```

پس در کل به تعداد 17هزار تا از دیکشنری به فرم بالا خواهید داشت که دیتای هر آگهی در آن آمده است.

در اینجا می‌توانیم ماتریس را به صورت مستقیم از دیکشنری‌ها بسازید. اما برای آنکه از صحت کامپایلر زبان Pyxcel که در فاز دوم زدید مطمئن بشید، ترجیح می‌دهید که ماتریس را یکبار هم در زبان Pyxcel بسازید. به همین خاطر دیتای هر دیکشنری به فرم بالا که به وجود آمد را به فرم دستور زبان Pyxcel مینویسید(در ابتدا باید هدر را ست کنید و سپس دیتارا وارد کنید):

```
$ Set Header for the first row
A1 = 'company'
B1 = 'car'
C1 = 'tream'
D1 = 'kilometer'
E1 = 'year'
F1 = 'price'

$ set data of first car
A2 = 'peugeot'
B2 = '206'
C2 = 'ir-type2'
D2 = 0
E2 = 1399
F2 = 'agreement'

$ set data of third car
A3 = ...
...
```

پس کافیست دستورات فوق را به کامپایلر Pyxcel ورودی بدهید. اما از آنجایی که کد کامپایلر زبان پیکسل خیلی طولانی است، ترجیح می‌دهید که...

Online Pyxcel!

از آنجایی که کد کامپایلر زبان پیکسل خیلی طولانی است، ترجیح می‌دهید که آن را در یک سرور گذاشته و به آن سرور خطوط زبان پیکسل را ارسال کنید. در سرور خطوط دستورات شما به کامپایلر داده شده و ماتریس ایجاد شده به سمت شما(کلاینت) ارسال خواهد شد.

مفهوم کلاینت و سرور را در فاز1 دیدیم. در پایتون امکان ایجاد ارتباط بین دو فایل مختلف با فرمت .py به وسیله socket فراهم شده است. حال ممکن است این دو فایل روی دو کامپیوتر مختلف باشد، و یا هر دو روی کامپیوتر خودتان قرار گرفته باشد. در هر صورت روش کار مشابه است.

برای بررسی Socket میتوانید سرچ انحام دهید و در مورد این لایبرری تحقیق کنید تا در جلسه توضیح پروژه به طور دقیق براتون شرح داده بشه.

پس کار شما ساده است:

۱. ارسال خط به خط کدی که در بخش قبل اماده کرده بودید از کلاینت به سرور
۲. اجرای تمام دستورات ورودی به سرور توسط کامپایلری که در سرور قرار دارد
۳. پس از انتهای ارسال خطوط برای آنکه سرور بداند کار شما تمام شده است، دستور 'get result' را به سرور ارسال کنید تا سرور برایتان ماتریس را ارسال کند.

برای انتقالات بین سرور، برای آنکه درگیر اینکود و دیکود های مزاحم نباشید، می‌توانید از لایبرری JSON استفاده کنید. این لایبرری به صورت خودکار عملیات اینکود و دیکود دیتا را انجام میدهد.

پس از دریافت ماتریس از سرور، مطمئن شده اید که کامپایلر pyxcel به خوبی کار می‌کند! پس حالا به سراغ تحلیل دیتا می‌رویم...

دلال ماشین

کتابخانه pandas کتابخانه ای بسیار بزرگ و محبوب برای زبان پایتون است. این کتابخانه مهم ترین ابزار تحلیل big data و data science می باشد.

از کاربرد های این کتابخانه می‌توان به موارد زیر اشاره کرد:

- انجام محاسبات آماری روی داده ها مانند محاسبه بزرگترین و کوچکترین مقدار، محاسبه میانگین داده ها، محاسبه صدک ها و...
- نگاهی به نحوه توزیع داده ها در یک ستون
- بررسی احتمال وابسته بودن ستون ها به یکدیگر
- پاک سازی داده ها: مثلا حذف کردن ردیف هایی که مقادیر ناقص دارند و یا حذف کامل بخش هایی که مقداری ندارند و خالی هستند، یا مرتب سازی یا فیلتر کردن ستون های خاص بر اساس شرط هایی خاص و الی آخر.
- همکاری با پکیج های بزرگ دیگر مانند Matplotlib برای بصری سازی داده ها: تولید نمودار های مختلف، هیستوگرام ها و الی آخر.
- ذخیره سازی داده های پاک سازی شده در یک فایل (CSV یا Excel و ...)

حال ماتریس گفته شده را مطابق شکل بخش اول در یک دیتافریم Pandas ذخیره کنید.

از روی دیتافریم بدست آمده و تنها با کمک کوئری های پانداز به این سوالات از بین تمام آگهی های سایت جواب بدید:

- انواع مدل های خودروی شرکت هیوندا و شرکت پژو را نام ببرید.
- چند خودرو در حداکثر 5 سال پیش تولید شده اند؟
- خودرو پژو206SD چند مدل مختلف دارد؟ بیشترین مدل آن کدام است؟
- چه آگهی دهنده‌ای بیشتر اهل سفر بوده و کار زیادی از ماشین کشیده است؟
- میانگین قیمت های آگهی های کدام شرکت خودروسازی از بقیه بیشتر است؟
- بیشترین آگهی مربوط به کدام ماشین بوده است؟
- خودرو های 206 مدل 85 تا 92 به صورت میانگین چقدر با خودروهای مدل 93 تا 1400 اختلاف قیمت دارند؟

بارم فاز سوم پروژه

بخش اول و سوم هرکدام 100 نمره و بخش دوم پروژه 200 نمره خواهند داشت.

در روز های 24 و 25 بهمن ماه ارائه از هر سه فاز پروژه انجام خواهد شد و نمره این فاز، و بخش آخر فاز2 وارد خواهد شد.

آپلود کوئری پانداز

در این بخش کوئری های پانداز خود را در قالب .py آپلود کنید