دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر

درس

تحلیل داده و مصورسازی

دکتر محمدامین صادقی - دکتر محمدرضا ابوالقاسمی

طراح تمرین: مرصاد اصالتی

زمان بارگزاری تمرین: یکشنبه ۲۴ مهر ماه ۱۴۰۱

تمرین شماره ۲

موضوع تمرین: آشنایی با ابزار های crawling

نیمسال اول سال تحصیلی ۱۴۰۱ - ۱۴۰۲

بخش اول – Crawl اخبار

کتابخانه beautifulsoup یک کتابخانه python است که به منظور استخراج داده از فایل های html و xml مورد استفاده قرار میگیرد. در این بخش از تمرین به دنبال جمع آوری لیستی از اخبار مربوط به طلا و سکه و ارزی از سایت https://www.tgju.org با استفاده از کتابخانه beautifulsoup هستیم.

آرشیو اخبار طلا و سکه:

https://www.tgju.org/news/category/93964/%D8%A7%D8%AE%D8%A8%D8%A7%D8%B1-%D8%B7%D9%84%D8%A7-%D9%88-%D8%B3%DA%A9%D9%87/page/1

آرشیو اخبار ارزی:

https://www.tgju.org/news/category/93965/%D8%A7%D8%AE%D8%A8%D8%A7%D8%B1-%D8%A7%D8%B1%D8%B2%DB%8C/page/1

در این بخش از تمرین می بایست از هر یک از آرشیو های خبری به تعداد حداقل ۵۰۰ خبر crawl کرده و در قالب یک فایل csv ذخیره نمایید. اطلاعاتی که می بایست در فرایند crawling جمع آوری نمایید به شرح زیر می باشد:

- title: عنوان
- description: توضیح مختصر خبر
- datetime: تاریخ و زمان ثبت
- category: دسته بندی
- agency: خبرگزاری

نکته: تمامی اطلاعات خواسته شده در صفحه لیست اخبار موجود می باشد و نیازی به crawl کردن صفحه خبر نمی باشد.

نکته: اطلاعات datetime می بایست در قالب استاندار و به صورت تاریخ میلادی ذخیره شود. نیازی به تبدیل تاریخ شمسی به میلادی وجود ندارد و اطلاعات تاریخ میلادی در محتویات سایت وجود دارد. به طور مثال:

2022-10-10 10:47:07

نکته: زمان مورد نیاز برای crawl کردن اخبار متناسب با سرعت اینترنت شما متغیر می باشد. طبق تجربه صورت گرفته توسط ما زمانی معادل ۲-۳ دقیقه برای دریافت ۵۰۰ خبر مورد نیاز می باشد.

پس از دریافت اخبار عملیات های زیر را انجام دهید:

۱. پنج خبرگزاری که بیشترین تعداد خبر را منتشر کرده اند را بیابید. (به تفکیک دسته بندی)
۲. نمودار trend مربوط به تعداد اخبار منتشر شده در روز های مختلف را رسم نمایید.

بخش دوم – Crawl اطلاعات آب و هوا

ابزار selenium یک ابزار قدرتمند در زمینه web scraping و web automation می باشد. با selenium میتوان از طریق کدنویسی به صورت تعاملی با مرورگر کار کرد و محتوای سایت را crawl کرد. از این ابزار برای نوشتن ربات نیز استفاده می شود که در این تمرین مد نظر ما نمی باشد. در این تمرین به دنبال جمع آوری اطلاعات آب و هوای شهر Washington در سال ۲۰۲۱ از سایت https://www.wunderground.com هستیم.

راهنمایی: برای راحتی کار بهتر است گزارش های ماهیانه را crawl کنید. به طور مثال اطلاعات آب و هوای ۰۱-۲۰۲۱ در آدرس زیر یافت می شود:

https://www.wunderground.com/history/monthly/us/va/arlington/KDCA/date/2021-01

اطلاعاتی که می بایست در فرایند crawling جمع آوری نمایید به شرح زیر می باشد:

- Temperature (°F) - only Avg metric
- Dew Point (°F) - only Avg metric
- Humidity (%) - only Avg metric
- Wind Speed (mph) - only Avg metric
- Pressure (in) - only Avg metric
- Precipitation (in) - only Avg metric
- date

نکته: اطلاعات date می بایست در قالب استاندار و به صورت تاریخ میلادی ذخیره شود. به طور مثال:
2021-02-01

نکته: اطلاعات مربوط به آب و هوا هر ماه را در یک فایل csvبه صورت مجزا ذخیره نمایید.

اطلاعات مربوط به آمار استفاده از دوچرخه های اشتراکی شهر Washington در سال ۲۰۲۱ را از سایت https://s3.amazonaws.com/capitalbikeshare-data/index.html دانلود نمایید.

پس از جمع آوری اطلاعات آب و هوا و دانلود اطلاعات دوچرخه های اشتراکی عملیات های زیر را به ترتیب انجام دهید:

۱. جداول مربوط به اطلاعات آب و هوا در ماه های مختلف را با یکدیگر ادغام نمایید.
۲. جداول مربوط به دوچرخه های اشتراکی در ماه های مختلف را با یکدیگر ادغام نمایید.
۳. عملیات data cleaning را برای دوچرخه های اشتراکی انجام دهید.

a. حذف ردیف هایی که حداقل یک مقدار nan دارند
b. حذف ردیف هایی که اطلاعات تاریخ شروع و پایان فرمت نامناسب دارند
c. حذف ردیف هایی که مدت زمان سفر در آن ها بیشتر از یک روز میباشد

۴. جدول اطلاعات آب و هوا را با دوچرخه های اشتراکی join نمایید. (ستون مشترک را date در جدول آب و هوا و بخش date از ستون started_at از جدول دوچرخه های اشتراکی در نظر بگیرید)

۵. نمودار هیستوگرام مربوط به Humidity و Temperature را رسم نمایید.

نمودار مربوط به trend تغییرات Temperature در روز های مختلف سال و تغییرات مربوط به تعداد سفر های انجام شده در روز های مختلف سال را در یک نمودار رسم نمایید.

نکات پیاده سازی و تحویل

- مهلت ارسال این تمرین تا پایان روز جمعه ۶ آبان ماه خواهد بود.
- انجام این تمرین به صورت یک نفره می باشد.
- خروجي مورد انتظار تمرین فایل jupyter ضمیمه شده و فایل zen_of_python.txt که بر روی آن تغییرات گفته شده را اعمال کرده‌اید، می‌باشد.
- هر گونه توضیحات و گزارشنویسي را به صورت Markdown داخل کتابچه jupyter انجام دهید.
- هر گونه تشابه میان تمرین‌های تحویل داده شده به عنوان تقلب در نظر گرفته می‌شود.
- در صورت استفاده از کدهای آماده، لینک مورد استفاده حتما ذکر شود.
- لطفا گزارش، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمائید.

HW2_[Lastname]_[StudentNumber].zip

برای مثال: HW2_esalati_12345678.zip

- در صورت وجود سوال و یا ابهام میتوانید از طریق رایانامه زیر با دستیار آموزشی در ارتباط باشید:

مرصاد اصالتی - mersad.esalati@gmail.com

شاد و سلامت باشید ☺