به نام خدا

تمرین جمع‌آوری داده از وب:

برای استخراج داده‌ها از سایت ره‌آورد۳۵۶ از کتابخانه‌های BeautifulSoup، selenium، unidecode و pandas استفاده شده است. کدهای مورد استفاده در این تمرین به شرح زیر می‌باشد:

```
class stock():
    Name = ''
    Url = ''
    Buy = ''
    Sell = ''
    Neutral = ''
url = 'https://rahavard365.com/stock'
base = 'https://rahavard365.com'
driver = webdriver.Chrome(executable_path=r'C:\Users\ASUS\Downloads\Compressed\chromedriver')
driver.get(url)
htmlSource = driver.page_source
df = pandas.DataFrame(columns=["Name", "Buy", "Neutral", "Sell"])
try:
    soup = BeautifulSoup(htmlSource, "html.parser")
    for i in soup.findAll('a', {'class': 'symbol'}):
        item = stock
        item.Name = i.text
        item.Url = base + i['href']
        driver.get(item.Url)
        htmlSource = driver.page_source
        soup = BeautifulSoup(htmlSource, "html.parser")
        item.Buy = unidecode(soup.find('div', {'id': 'indc_buy'}).text)
        item.Sell = unidecode(soup.find('div', {'id': 'indc_sell'}).text)
        item.Neutral = unidecode(soup.find('div', {'id': 'indc_neutral'}).text)
        df.loc[len(df)] = [item.Name, item.Buy, item.Neutral, item.Sell]
    driver.quit()
except:
    pass
df.to_csv("data.csv", sep=',', index=False)
```

۱

۲

۳ معادل HTML آن : ### <a class="symbol">

۴ ###

۵

در بخش اول برای اسامی نمادها، لینک هر کدام از نمادها و مقادیر خرید و فروش و خنثی گیج اندیکاتور کلاس‌ها و ظروف خالی تعریف می‌شود.

در بخش دوم آدرس سایت و درایور را تعریف کرده و با استفاده از درایور یک صفحه وب جدید را در پنجره مرورگرِ موجود به این ترتیب بارگذاری می‌کند، سپس منبع HTML صفحه فعلی را دریافت می‌کند.

در بخش سوم یک چارچوب اطلاعاتی برای ۴ متغیر خواسته شده به وجود می‌آید، ابتدا تابع stock را معادل با item در نظر می‌گیرد و برای هر نماد پیدا شده در سایت https://rahavard365.com/stock، نام آن نماد را به صورت متنی درآورده و در متغیر item.Name می‌ریزد.

همچنین لینک سهام مختلف که طبق کد HTML آن، در کلاس href قرار دارند را برای هر سهمی به صورت i['href'] بدست آورده و در ادامه لینک اصلی (https://rahavard365.com) قرار می‌دهد و متغیر item.Url به وجود می‌آید. به این ترتیب هربار وارد صفحه مربوط به هر سهم شده و اطلاعات گیج اندیکاتور را به این صورت از آنجا بدست می‌آورد (که این مطلب در بخش بعدی توضیح داده می‌شود).

در بخش چهارم، ۳ متغیر مربوط به گیج اندیکاتور را آدرس‌دهی و تعریف می‌کند و پس از تکمیل اطلاعات از درایو خارج می‌شود.

در بخش آخر نیز چارچوب اطلاعاتی را به فایل CSV تبدیل کرده است.

به دلیل عدم درست خواندن فونت فارسی، اسامی ذخیره شده در فایل CSV ناخوانا می‌باشند بنابراین برای رفع این مشکل ابتدا فایل CSV را باز کرده و در قسمت Data بخش From text را انتخاب کرده و پس از انتخاب فایل مذکور با تغییر گزینه فعلی در بخش File origin به گزینه UTF-8 و ذخیره دوباره فایل در قالب Excel Workbook داده‌های خوانا به وجود می‌آیند که بخشی از خروجی حاصل به شرح جدول۱ می‌باشد.

(برای از دست نرفتن داده‌ها هنگام تبدیل آن، فایل را به جای CSV به Excel Workbook (xlsx) تغییر فرمت می‌دهیم.)

Name	Buy	Neutral	Sell
آ س پ	9	7	11
آبادا	11	5	8
آبین	6	6	15
آپ	8	8	11
اپرداز	18	5	4
اتکام	18	5	4
اخابر	10	7	10
ارفع	9	6	12
آرمان	19	3	5
آریا	13	9	3
آریان	9	7	11
آسیا	13	7	7
اعتلا	7	6	14
افرا	11	6	10
افق	14	9	4
امید	8	7	12
امین	9	9	4
انرژی3	7	6	9
آینده	10	6	11
بالاس	9	5	13
بالبر	6	7	14
بایکا	6	11	10
بپاس	18	4	5
بپیوند	8	6	1
⋮			
وکادو		2	16
وگردش	10	8	9
وگستر	9	5	13
ولانا	8	3	16
ولبهمن	8	6	13
ولپارس	15	5	5
ولتجار	7	5	15
ولساپا	7	8	12
ولشرق	7	6	14
ولصنم	8	7	12
ولغدر	10	4	13
ولملت	8	3	16
ولیز	8	4	15
ومشان	14	8	5
ومعلم	10	7	10
وملل	16	7	4
وملی	10	4	13
ومهان	5	11	11
ونفت	11	8	8
ونوین	14	7	6
ونیرو	8	4	15
ونیکی	11	7	9
وهنر	7	6	14
وهور	10	9	8

جدول۱