



Web Scraper - dokumentacja

1. Wprowadzenie i opis teoretyczny

Czy jest web scraping? Jest to technika wyodrębniania danych ze stron internetowych, która zastępuje ręczne, powtarzalne wpisywanie lub kopiowanie i wklejanie. Dodatkowo pozyskane dane często są przechowywane w ustrukturyzowanym formacie.

Program Currency Scraper pobiera i wyświetla aktualny kurs głównych walut (notowania Narodowego Banku Polskiego - tzw. kursy średnie) oraz informację o ich zmianie w procentach w ciągu ostatnich 24h. Waluty, które są wyświetlane, to:

- USD - dolar amerykański
- EUR - euro
- CHF - frank szwajcarski
- GBP - funt szterling
- NOK - korona norweska

Program wykorzystuje bibliotekę `requests` w celu obsługi żądań HTTP (wysyłanie ich, odbieranie odpowiedzi), bibliotekę `Beautiful Soup` (ang. "piękna zupa"), która wykorzystuje wstępnie zainstalowany parser html / xml i konwertuje stronę internetową do drzewa składającego się z tagów, elementów, atrybutów i wartości.

2. Sposób uruchomienia

Program uruchamiany jest za pomocą polecenia:

```
python3 CurrencyScraper.py
```

Może on zostać uruchomiony w oknie terminala lub w zintegrowanym środowisku programistycznym (ang. IDE - integrated development environment).

3. Uwagi na temat implementacji

a) informacje wstępne

Program pobiera informacje na temat aktualnego kursu walut z portalu finansowego [mybank.pl](#). Poniżej przedstawiony jest zrzut ekranu strony głównej tego serwisu:

AKTUALNE WIADOMOŚCI I KOMENTARZE RYNKOWE

PRZEGŁĄD WYDARZEŃ NASTĘPNEGO TYGODNIA

2021-01-22 Komentarz tygodniowy TMS Brokers

Średnioterminowe nadzieje ścierają się z krótkoterminowymi ryzykami, co utrudnia zawiązanie stabilnego trendu wzrostowego rynkowych aktywów. Optymizm podparty zaszczepieniem ludności i planami fiskalnego wsparcia ożywienia nie powinien zaniknąć, ale z rozwinięciem nowej fali wzrostów inwestorzy czekają na silny impuls – być może od Fed w środę?

INFORMACYJNA PRÓŻNIA

2021-01-22 Raport DM BOŚ z rynku walut

W piątek rano widzimy próbę podbić dolara na szerokim rynku, co można wiązać ze schłodzeniem nastrojów na rynkach akcji. Na Wall Street mieliśmy realizację zysków, podczas kiedy nastroje w Azji zdominały obawy związane z COVID. Według doniesień planowany jest lockdown na części obszaru Hong Kongu, ale i też większe restrykcje mają dotknąć wybrane miasta w Chinach - tamtejsze władze obawiają się transmisji wirusa w kontekście zbliżających się obchodów chińskiego Nowego Roku.

STABILNE DANE Z EUROPY

2021-01-22 Poranny komentarz walutowy XTB

Europejskie gospodarki nadal muszą mierzyć się z serią ograniczeń. Jak na te okoliczności wstępne indeksy PMI z Europy i tak wypadły nieźle. Coraz bardziej rozgrzany jest za to amerykański rynek nieruchomości. Przelom roku nie był szczególnie dobry dla europejskiej gospodarki. Co prawda udało się „rzutem na taśmę” znaleźć porozumienie ws. Brexitu, ale i tak wyjście Wielkiej Brytanii z UE wiąże się z komplikacjami.

Kursy walut - Notowanie z dnia 2021-01-22

Waluta	Kurs (zł)	Zmiana (%)	Zmiana (zł)		
USD	3,7255	-0.15%	-0.0057		
3.956	3.874	3.791	3.708	3.625	
mybank.pl	3.956	3.874	3.791	3.708	3.625
2020.07.21	2021.01.22				
EUR	4,5354	↑ 0.10%	↑ 0.0044		
CHF	4,2102	↑ 0.12%	↑ 0.0049		
GBP	5,0920	-0.63%	-0.0324		
NOK	0,4391	-0.72%	-0.0032		

[Darmowe komponenty na stronę www](#)

Kursy walut na żywo - prosto z rynku Forex

Waluta	Kurs (zł)	Zmiana	Czas
USD	3.72940	↑ 0.10%	22:59:54
EUR	4.53972	↑ 0.10%	22:59:59
CHF	4.21250	↑ 0.05%	22:59:52
GBP	5.10406	↑ 0.24%	22:59:59
NOK	0.43907	-0.01%	22:59:52

Kursy walut są osadzone zawsze w tym samym miejscu w kodzie źródłowym strony, dzięki czemu możliwe jest ich pobranie. Wartości liczbowe w procentach określające zmianę kursu walut w ciągu ostatnich 24 godzin są również umieszczone w kodzie źródłowym strony.

Nie są to jednak wartości ze znakiem „+” lub „-”, dzięki którym można by jednoznacznie określić czy wartość danej waluty wzrosła czy zmalała. Zamiast tego znajdują się one w odpowiednich klasach HTMLa, nazwanych odpowiednio „b3 ziel” - dla kursów, które wzrosły oraz „b3 czer” dla kursów, które zmalały.

Na podstawie tej informacji program w konsoli wypisuje poprawnie zmianę kursu - jeśli wartość liczbową znajdowała się w klasie o nazwie „b3 ziel” - przed zmianą procentową wyświetlany jest znak „+”, a jeśli w „b3 czer” - “-”.

W serwice [mybank.pl](#) w zależności od nazwy klasy, wyświetlany jest przy wykorzystaniu kaskadowych arkuszy stylu graficzny znak zielonej strzałki do góry dla kursów, które wzrosły oraz graficzny znak czerwonej strzałki w dół dla kursów, które spadły.

b) biblioteka requests

Program wykorzystuje bibliotekę `requests`, która pozwala w bardzo prosty sposób wysyłać żądania HTML. Nie ma potrzeby przechowywać całych żądań GET, PUT, POST, PATCH, DELETE w postaci napisów. W celu pobrania całej strony wystarczy wykonać polecenie:

```
import requests
response = requests.get('https://adreswitryny.pl')
```

Następnie na obiekcie przechowującym odpowiedź można wykonać polecenie `response.text` i uzyskać cały kod strony w formie tekstowej.

c) sprawdzanie poprawności pobranej strony

Program sprawdza, czy strona została załadowana poprawnie, odczytując zwrócony kod odpowiedzi HTTP - gdy wszystko przebiegnie pomyślnie, zwracany jest status "200 OK". Gdy kod odpowiedzi będzie inny, program zwróci wyjątek.

Sprawdzany jest również sam kod strony, a dokładniej to, czy został on załadowany w całości, poprzez wyszukanie na początku i końcu kodu fragmentów, które są dla tej strony stałe. Tu również w przypadku błędu zostanie zwrócony wyjątek.

d) biblioteka Beautiful Soup

Biblioteka Beautiful Soup pozwala w bardzo prosty sposób wyszukiwać na stronie interesujące nas informacje i wyciągać je z niej. Posiada wiele intuicyjnych metod pozwalających na ich efektywne wykorzystanie, jak np.

```
soup.findAll('<h1>')
```

Lista jej zastosowań i możliwości jest bardzo długa. W połączeniu z biblioteką `re` służącą do obsługi wyrażeń regularnych można w bardzo łatwy sposób np. znaleźć na stronie fragment zawierający datę oraz sprawdzić jego poprawność przy pomocy wyrażenia regularnego.

4. Podsumowanie

Program został skonstruowany w ten sposób, by był odporny na błędy - był w stanie je wykrywać i zgłaszać odpowiednie wyjątki.

Mimo zwięzłego kodu wykorzystuje on wiele skomplikowanych zagadnień, takich jak obsługa zapytań HTTP z poziomu programu, parsowanie stron, testy przy użyciu asercji, wyrażenia regularne. Całość została wykonana w ten sposób, by była możliwie najbardziej czytelna i zrozumiała dla osób, które miałyby z tym programem styczność po raz pierwszy. Umieszczone w kodzie programu pojedyncze komentarze mają na celu ułatwienie zrozumienia zagadnień mogących wyglądać na niezroz

umiałe na pierwszy rzut oka.

5. Literatura i źródła

- Portal finansowy mybank.pl, z którego pobierane są informacje na temat kursów walut
 - <https://mybank.pl/>

Dokumentacja biblioteki Beautiful Soup

- <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

Dokumentacja biblioteki Requests

- <https://requests.readthedocs.io/en/master/>

Artykuł objaśniający działanie biblioteki Requests

- <https://realpython.com/python-requests/>

Dokumentacja biblioteki Re

- <https://docs.python.org/3/library/re.html>

Poradnik wykorzystania biblioteki Re

- https://www.w3schools.com/python/python_regex.asp

Poradnik wideo przedstawiający na przykładach działanie Beautiful Soup

- <https://www.youtube.com/watch?v=ng2o98k983k>

Artykuł tłumaczący działanie assert w Pythonie

- <https://www.programiz.com/python-programming/assert-statement>

Strona przedmiotu "Język Python"

- <https://ufkapano.github.io/algorytmy/index.html>