

Python średnio zaawansowany

Dzień 4



Blok nr 2:

Akwizycja danych

AGENDA

- Web scraping
- Narzędzia

Web scraping

Web scraping

Web scraping to metoda automatycznego pozyskiwania danych ze stron internetowych przez boty (web crawler, harvester). Pozyskane dane mogą być przechowywane w różnych formatach, bazach danych, arkuszach kalkulacyjnych w celu późniejszej analizy.

Proces web scrapingu składa się z dwóch aktywności:

1. Pobranie strony internetowej
2. Wydobycie danych z dokumentu



Web scraping vs API

Dostęp:

- aplikacja internetowa nie oferuje dostępu poprzez API
- API nie udostępnia pełnego zestawu danych
- API może wymagać płatnego dostępu
- aktywność wynikająca z korzystania z API jest monitorowana

Dodatkowo:

- dokument HTML musi zostać sparsowany w celu ekstrakcji danych

Złota zasada web scrapingu

Nie korzystaj z web scrapingu jeżeli aplikacja internetowa oferuje dostęp poprzez API

Charakterystyka web scrapingu

- Kod odpowiedzialny za web scraping generuje requesty (symuluje aktywność użytkownika)
- Serwer aplikacji odpowiada na żądania crawlera
- Jeżeli crawler generuje ruch 1000 zapytań na minutę w niektórych przypadkach może to negatywnie wpłynąć na wydajność strony (nawet spowodować **tymczasową niedostępność zasobu!**)

Utrudnienia w web scrapingu

- Osoba odpowiedzialna za administrację serwerem aplikacji może zablokować ruch przychodzący z maszyny na której został uruchomiony crawler jeżeli zbyt często scrapujemy zawartość (jesteśmy traktowani jako intruz)
- Regulamin korzystania z usług strony internetowej może przewidywać **wyciągnięcie konsekwencji w przypadku korzystania z botów**

Rozwiązanie: jeśli piszesz bota, zapisuj ściągnięte strony i na nich lokalnie udoskonalaj kod żeby minimalizować ryzyko wykrycia

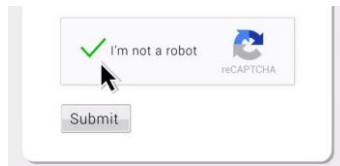
Utrudnienia w web scrapingu

- Scrapowanie stron korzystających z Javascript jest mocno utrudnione (dociągają kolejne zasoby)
- W przypadku zmiany struktury HTML na stronie konieczna jest zmiana kodu crawlera 😞
- Aplikacje sprawdzają nagłówki, zwłaszcza pola **User-Agent**
- Aplikacje sprawdzają obecność i zawartość cookies

Utrudnienia w web scrapingu

- Sprawdź zawartość pliku robots.txt, np. <https://www.lotto.pl/robots.txt>
- Kod HTML nie musi być poprawnie zbudowany

- CAPTCHA



- Badaj timeout
- Jesteś świadom różnic w aktywności przeglądarki i skryptu, np. czy myślisz o ściąganiu obrazków 1x1 piksel?

Narzędzia

Narzędzia deweloperskie

Przeglądarki internetowe oferują funkcjonalność inspektora elementów HTML aktualnie przeglądanej, kompletnej, strony internetowej. Pozwala on analizować i edytować widoczny kod.

Kiedyś: Firebug



Dziś:

- **Firefox:**
 - F12 -> zakładka "Inspektor,,
- **Chrome/Internet Explorer:**
 - F12 -> zakładka "Elementy"

Wyszukiwanie elementu HTML

The screenshot shows the Lotto website interface. At the top, there are logos for various lottery games: LOTTO, Lotto, EURO JACKPOT, Ekstra Pensja, Multi Multi, Mini Lotto, KASKADA, Super Szansa, Zdrapki, and KENO. A search bar and a newsletter sign-up link are also visible.

The main banner features the Lotto logo and the text: "A CO JEŚLI TO TY TERAZ WYGRASZ? KUMULACJA 3 000 000". Below this, there are several promotional tiles for different lottery games and events, including "GRAJ W ZAKŁADY SPECJALNE!", "Gdzie odebrać wygraną od 2280 zł?", "GRAJ W ZAKŁADY SPECJALNE!", "Gramy dla sportu i kultury", "Mini Lotto", "Zagraj i poznaj swoje szczęśliwe liczby!", "Miliony", and "Sprzedaż gry trwa tylko do 10 listopada".

A context menu is open over the "Lotto" text in the banner. The menu options are:

- Zapisać stronę jako...
- Wyślij stronę do Pocket
- Prześlij stronę do
- Pokaż obraz tła
- Zaznacz wszystko
- Pokaż źródło strony
- Pokaż informacje o stronie
- Zbadaj element** (highlighted with a red box)
- Wykonaj zrzut ekranu

Wyszukiwanie elementu HTML

The screenshot shows a web browser displaying the Lotto website. The main banner features the Lotto logo and the text "A CO JEŚLI TO TY TERAZ WYGRASZ? KUMULACJA 3 000 000". On the right, there are sections for "Lotto" results (5, 15, 23, 33, 35, 45), "Lotto Plus" results (2, 10, 11, 26, 33, 48), and "Eurojackpot" results (6, 26, 31, 42, 50, 2, 9). The browser's developer tools are open at the bottom, with the "Inspektor" (Inspector) tab selected. The DOM tree shows the following structure:

```
<div class="resultsOpen">  
  <div class="resultLotto">  
    <div class="resultsItem lotto">  
      <h2></h2>  
      <div class="resultsTime"></div>  
      <div class="resultnumber">  
        <div class="number text-center">  
          <span>5</span>  
        </div>  
        <div class="number text-center"></div>  
        <div class="number text-center"></div>  
      </div>  
    </div>  
  </div>  
</div>
```

A red arrow points to the "Inspektor" tab in the developer tools. Another red arrow points to the selected HTML element, `5`, in the DOM tree. The "Wybrany element HTML" (Selected HTML element) label is placed next to this element.

Wyszukiwanie elementu HTML

The screenshot shows a web browser displaying a lottery results page. The main content area has a blue background with a large yellow star and the text "16 października padła wygrana 3 539 917 zł w Poznaniu". To the right, there are sections for "Lotto", "Lotto Plus", and "Eurojackpot" with their respective winning numbers and dates. A context menu is open over the main content area, listing various actions like "Edytuj jako HTML", "Utwórz nowy węzeł", "Kopij", "Wklej", etc. A red box highlights the "Wewnętrzny HTML" (Internal HTML) option, which is further expanded to show search methods: "Zewnętrzny HTML", "Selektor CSS", "Ścieżka CSS", "XPath", and "Obraz jako Data-URL". The bottom of the browser shows the "Dane" (Data) panel with the "Filtruj style" (Filter styles) tab selected, displaying a list of CSS rules and their corresponding elements.

Dołącz do nas na

LOTTO **lotto** **EURO JACKPOT** **Ekstra Pensja** **Multi Multi** **Mini Lotto** **KASKADA** **Super Szansa** **zdrapki** **Weryo**

Zarządzaj newsletterem

Lotto
wyniki z dnia 10-10-18, godz.: 21:40
5 15 23 33 35 45
Lotto Plus
2 10 11 26 33 48
Super Szansa
wyniki z dnia 10-10-18, godz.: 21:40
9 1 8 2 5 6 7

Eurojackpot
wyniki z dnia 12-10-18
6 26 31 42 50 +
2 9

16 października padła wygrana
3 539 917 zł
w Poznaniu

Edytuj jako HTML
Utwórz nowy węzeł
Powiel węzeł
Usuń węzeł
Atrybuty
hover
active
focus
Kopij
Wklej
Rozwiń wszystkie
Zwiń wszystkie
Przeświń stronę do zaznaczonego elementu
Zrzut ekranu węzła
Otwórz w konsoli
Wyświetl własności DOM
Wewnętrzny HTML
Zewnętrzny HTML
Selektor CSS
Ścieżka CSS
XPath
Obraz jako Data-URL

Dane
Szukaj w kodzie HTML
Filtruj style
Pseudoelementy
Ten element
element { inline
_8ab08d4d733eca64ba6069863da9a3f6_style.css:2
section .contentBox .content .resultstBox
.resultsItem .resultnumber .number span {
vertical-align: middle;
font-size: 0.65625rem;
color: #000000;
_9b4a2de9937541db5d18a5cf7a573_bootstrap.css:4
Filtruj style
box-sizing
border-box
color
rgb(0, 0, 0)
Font-family
"DaxProMedium"
Font-size
14px
Font-weight
400
line-height

Narzędzia



Scrapy

Scrapy:

- rozbudowany kombajn do tworzenia botów ściągających dane i dodatkowo:
 - umożliwia eksport danych w różnych formatach
 - loguje czynności
 - kompleksowo konfigurowalny (liczba botów, obsługa cookies, opóźnienia)
- średnio wysoki próg wejścia

BeautifulSoup + Requests



Dzięki!

