

Python średnio zaawansowany

Dzień 6



Blok nr 3:

Przetwarzanie danych

AGENDA

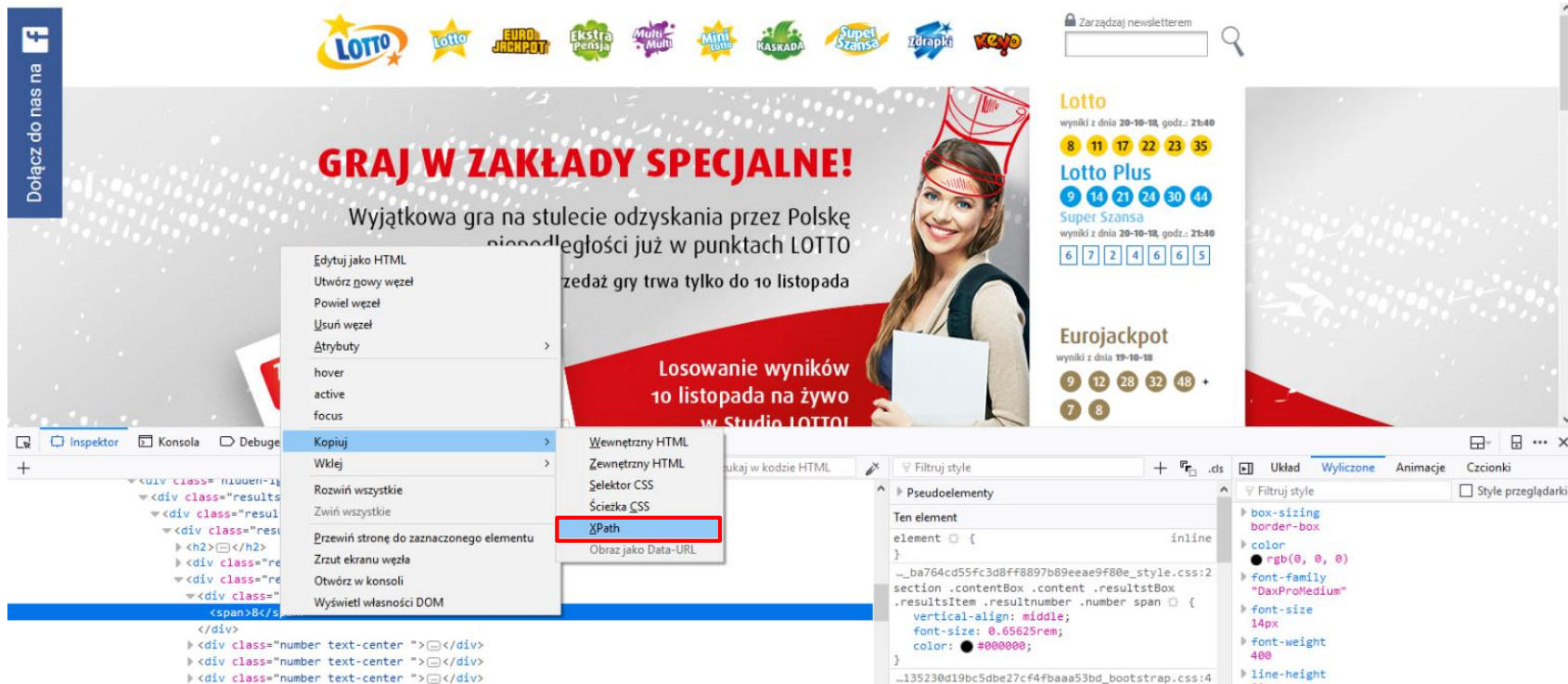
- XPATH
- Selektory CSS
- Moduł Parsel

XPATH

XPATH

XPATH (XML Path Language) to język do przeszukiwania ścieżek XML. Pozwala w elastyczny sposób wskazywać różne części dokumentu XML a ponadto może być używany do sprawdzania, czy wskazane węzły dokumentu pasują do wzorca.

XPATH w przeglądarce



Zaznaczanie elementów

nodename	Zaznacz wszystkie węzły o nazwie "nodename"
/	Zaznaczanie od węzła początkowego (root node)
//	Zaznaczanie węzłów w dokumencie od obecnego, który został dopasowany
.	Zaznaczanie aktualnego węzła
..	Zaznaczanie rodzica aktualnego węzła
@	Zaznaczanie atrybutów

Więcej na https://www.w3schools.com/xml/xpath_intro.asp

Przykład

```
<bookstore>
  <book>
    <title lang="en">Harry Potter</title>
    <price>29.99</price>
  </book>

  <book>
    <title lang="en">Learning XML</title>
    <price>39.95</price>
  </book>
</bookstore>
```


bookstore	Zaznaczenie węzłów o nazwie <i>bookstore</i>
/bookstore	Zaznaczenie najmniej zagnieżdżonego węzła <i>bookstore</i> (root)
bookstore/book	Zaznaczenie węzłów <i>book</i> , które są dziećmi węzła <i>bookstore</i>
//book	Zaznaczenie węzłów <i>book</i> niezależnie od ich położenia w dokumencie
bookstore//book	Zaznaczenie węzłów <i>book</i> , które są potomkami węzła <i>bookstore</i> niezależnie od ich położenia w węźle <i>bookstore</i>
//@lang	Zaznaczenie węzłów, które mają atrybut o nazwie <i>lang</i>











Predykaty

Wyrażenie	Wynik
/bookstore/book[1]	Wybiera pierwszy węzeł (element) book , który jest dzieckiem węzła bookstore .
/bookstore/book[last()]	Wybiera ostatni węzeł book , który jest dzieckiem węzła bookstore .
/bookstore/book[last()-1]	Wybiera przedostatni węzeł book , który jest dzieckiem węzła bookstore .
/bookstore/book[position()<3]	Wybiera dwa pierwsze węzły book , które są węzłami potomnymi węzła bookstore
//title[@lang]	Wybiera wszystkie węzły title , które mają atrybut o nazwie lang
//title[@lang='en']	Wybiera wszystkie elementy title , które mają atrybut o nazwie lang o wartości "en"
/bookstore/book[price>35.00]	Wybiera wszystkie elementy book elementu bookstore , które mają element price o wartości większej niż 35,00
/bookstore/book[price>35.00]/title	Wybiera wszystkie elementy title będące elementami book elementu bookstore , które mają element price o wartości większej niż 35,00

Selektory CSS

Selektory CSS w przeglądarce

Dołącz do nas na 


         

Zarządzaj newsletterem

KONKURS
Podróże z Eurojackpot

Lotto
wyniki z dnia 20-10-18, godz.: 21:40
8 11 17 22 23 35
Lotto Plus
9 14 21 24 30 44
Super Szansa
wyniki z dnia 20-10-18, godz.: 21:40
6 7 2 4 6 6 5

Eurojackpot
wyniki z dnia 19-10-18
9 12 28 32 48 +
7 8



Edytuj jako HTML
Utwórz nowy węzeł
Powiel węzeł
Usuń węzeł
Atrybuty
hover
active
focus

Kopiuji
Wklej
Rozwiń wszystkie
Zwiń wszystkie
Przełącz stronę do zaznaczonego elementu
Zrzut ekranu węzła
Otwórz w konsoli
Wyświetl właściwości DOM

Wewnętrzny HTML
Zewnętrzny HTML
Selektor CSS
Ścieżka CSS
XPath
Obraz jako Data-URL

Filtruj style
Pseudoelementy
Ten element
element { inline
}
_28195b2fad71b3f9e6fb0920ec135761_style.css:2
section .contentBox .content .resultstBox
.resultsItem .resultnumber .number span {
vertical-align: middle;
font-size: 0.65625rem;
color: #000000;
}
_7e8e409hd67a47726777a612907e_hootstran.css:4

Filtruj style
box-sizing
border-box
color
● rgb(0, 0, 0)
font-family
"DaxProMedium"
font-size
14px
font-weight
400
line-height

```

<div class="number-ig-down">
  <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>8</span>
          </div>
        </div>
        <div class="number text-center"></div>
        <div class="number text-center"></div>
        <div class="number text-center"></div>
      </div>
    </div>
  </div>

```

Selektory CSS

Selektor	Przykład	Opis
.class	.intro	Wszystkie elementy o klasie <i>intro</i>
#id	#firstname	Element o id <i>firstname</i>
element	p	Wszystkie paragrafy <p>
element element	div p	Wszystkie paragrafy w elemencie <i>div</i>
element > element	div > p	Wszystkie paragrafy, których rodzicem jest element <i>div</i>

Selektory CSS

Selektor	Przykład	Opis
element + element	div + p	Wszystkie paragrafy następujące bezpośrednio po elemencie <i>div</i>
element ~ element	p ~ ul	Wszystkie elementy <i>ul</i> poprzedzone przez paragraf
[atrybut]	[target]	Wszystkie elementy posiadające atrybut o nazwie <i>target</i>
[atrybut=wartosc]	[target=_blank]	Wszystkie elementy posiadające atrybut o nazwie <i>target</i> o wartości “_blank”

Moduł parsel

Parsel

Moduł służący do wyodrębniania danych z dokumentów XML/HTML przy pomocy XPath i selektorów CSS.

Dodatkowo możliwe jest wykorzystanie wyrażeń regularnych.

<https://parsel.readthedocs.io/en/latest/>

Parsel - selektory XPATH

```
from parsel import Selector
```

```
(...)
```

```
text = requests.get(url).text
```

```
selector = Selector(text=text)
```

```
selector.xpath('//title')
```

```
selector.xpath('//title/text()').get() # pojedynczy, pierwszy rezultat
```

```
selector.xpath('//title/text()').getall() # lista elementów
```


Parsel - selektory CSS

```
from parsel import Selector
```

```
(...)
```

```
text = requests.get(url).text
```

```
selector = Selector(text=text)
```

```
selector.css('h1')
```

```
selector.css('h1::text').get()
```

```
selector.css('h1::text').getall()
```

Dzięki!

