AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE
AGH UNIVERSITY OF KRAKOW

# 12/03/2024

Ćwiczenia laboratoryjne
sem. Letni 2023/2024

# Agenda

- Uwarunkowania prawne i etyczne web scrapingu
- Plik robots.txt
- Stworzenie prostego crawlera

# Uwarunkowania prawne

- Web scraping nie jest wprost regulowany w polskim prawie
- Pozyskiwanie danych osobowych za pomocą web scrapingu to rodzaj ich przetwarzania – wymaga posiadania podstawy prawnej i spełnienia obowiązków informacyjnych
- Web scraping może wiązać się z kopiowaniem elementów objętych prawami autorskimi i osobistymi, takimi jak np. grafiki oraz teksty

# Co z regulaminem strony?

- Regulamin strony internetowej może być wiążący nawet jeżeli nie został przez nas zaakceptowany
- Ekstrakcja danych ze strony, które jawnie w regulaminie zabraniają web scrapingu, może skończyć się zablokowaniem konta użytkownika i/lub zablokowaniem adresu IP

# Przykład regulaminu

## 2. Postanowienia Ogólne

do góry

1. Warunki korzystania z Serwisu, w tym zasady Rejestracji, publikacji Ogłoszeń i nabywania Usług Odpłatnych, a także kwestie dotyczące płatności i postępowania reklamacyjnego określa Regulamin. Każdy korzystający z Serwisu jest zobowiązany do zapoznania się z treścią Regulaminu.
2. Goście mogą korzystać jedynie z ograniczonych funkcji Serwisu na zasadach określonych w niniejszym Regulaminie, z poszanowaniem przepisów prawa i zasad uczciwości.
3. Treści publikowane w Serwisie, w tym w szczególności Ogłoszenia, niezależnie od ich formy, tj. materiały tekstowe, graficzne oraz wideo, są przedmiotem ochrony praw własności intelektualnej, w tym prawa autorskiego oraz praw własności przemysłowej, Grupy OLX, Sprzedawców lub osób trzecich. Zabrania się jakiegokolwiek wykorzystywania tych treści bez pisemnej zgody uprawnionych. Zabrania się jakiegokolwiek agregowania i przetwarzania danych oraz innych informacji dostępnych w Serwisie w celu ich dalszego udostępniania osobom trzecim w ramach innych serwisów internetowych jak i poza Internetem. Zabrania się również wykorzystywania oznaczeń Serwisu oraz Grupy OLX, w tym charakterystycznych elementów grafiki bez zgody Grupy OLX.
4. Z zastrzeżeniem licencji udzielonej na rzecz Grupy OLX zgodnie z punktem 4.3, żadne z postanowień niniejszego Regulaminu nie stanowi udzielenia zgody na wykorzystywanie praw Grupy OLX lub praw osób trzecich, o których mowa w punkcie 2.3 ani też nie powinno być interpretowane jako zrzeczenie się tych praw.
5. Grupa OLX nie jest stroną Transakcji. Do Transakcji zawartych pomiędzy Konsumentami, tj. osobami fizycznymi niebędącymi przedsiębiorcami, w tym przy skorzystaniu z usługi Przesyłki OLX i Płatności OLX, nie znajdą zastosowania przepisy o ochronie praw konsumentów.

# Odstępstwa

1. W charakterze ilustracji, w celach dydaktycznych lub badawczych (ze wskazaniem źródła)
2. Cele administracyjno-sądowe
3. Dla dobra osób niepełnosprawnych lub będących beneficjentami w rozumieniu przepisów o prawie autorskim i prawach pokrewnych

Powyższy zakres odstępstw od generalnego zakazu korzystania z cudzych baz danych może zostać rozszerzony na mocy nowej unijnej dyrektywy prawnoautorskiej

# Odstępstwa

Projekt ustawy implementującej wspomnianą dyrektywę zakłada:

Wykorzystanie cudzych baz danych m.in. W celu eksploracji tekstów i danych, chyba, że uprawniony do nich podmiot zastrzegł inaczej i uczynił to w odpowiedni (zgodny z prawem) sposób.

Więcej informacji (data publikacji: 05/02/2024)
https://rynek-ksiazki.pl/aktualnosci/rzad-zajmie-sie-ustawa-ws-implementacji-dyrektywy-dotyczacej-praw-pokrewnych/

# Plik robots.txt

Plik tekstowy posiadający w swoim zapisie wytyczne dla robotów indeksujących/szukających treści (crawlerów).
Stosowany, aby ograniczyć indeksowanie zasobów zbędnych lub wyznaczyć zasoby, których automaty nie powinny przeszukiwać

# Przykłady pliku robots.txt

http://olx.pl/robots.txt
https://www.onet.pl/robots.txt
https://pl.wikipedia.org/robots.txt

# Dodatkowe materiały

Poradnik tworzenia i przesyłania pliku robots.txt od Google:
https://developers.google.com/search/docs/crawling-indexing/robots/create-robots-txt?hl=pl

Walidator pliku robots.txt
https://technicalseo.com/tools/robots-txt/

# Co to jest web crawler?

Inaczej robot indeksujący, spider website crawler, web spider, web robot – **jest programem lub zautomatyzowanym skryptem, którego zadanie poleca na przeszukiwaniu stron internetowych.**

Można ich używać w celu ekstrakcji treści oraz jej indeksacji, sprawdzania poprawności linków lub zgodności kodu HTML itp itd.

# Przykład prostego crawlera

```python
import requests
from bs4 import BeautifulSoup
from urllib.parse import urljoin

def get_links(url):

    response = requests.get(url)
    bs = BeautifulSoup(response.text, 'html.parser')
    links = bs.find_all('a', href=True)
    absolutes = [urljoin(url, link['href']) for link in links]

    return absolutes

def main():
    website_url = input("Enter the URL of the source to crawl: ")

    links = get_links(website_url)

    print("Links across the website:")
    for link in links:
        print(link)

if __name__ == "__main__":
    main()
```

# Bs4 find_all()

https://beautiful-soup-4.readthedocs.io/en/latest/index.html#find-all

## `find_all()`

Signature: find_all(name, attrs, recursive, string, limit, **kwargs)

The `find_all()` method looks through a tag's descendants and retrieves *all* descendants that match your filters. I gave several examples in Kinds of filters, but here are a few more:

```
soup.find_all("title")
# [<title>The Dormouse's story</title>]

soup.find_all("p", "title")
# [<p class="title"><b>The Dormouse's story</b></p>]

soup.find_all("a")
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
#  <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
#  <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>]

soup.find_all(id="link2")
# [<a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>]

import re
soup.find(string=re.compile("sisters"))
# u'Once upon a time there were three little sisters; and their names were\n'
```

# KONIEC