LL5. Colectarea automată a datelor.

Informații necesare: Web scraping cu Python

https://realpython.com/python-web-scraping-practical-introduction/

https://realpython.com/beautiful-soup-web-scraper-python/

BeautifulSoup poate fi folosit pentru analiza html.

Colectarea datelor de pe site-uri web folosind un proces automatizat este cunoscută sub numele de web scraping. Unele site-uri web interzic în mod explicit utilizatorilor să accese datele cu instrumente automate precum cele pe care le veți crea în acest tutorial. Site-urile web fac acest lucru din două motive posibile:

- Site-ul are un motiv întemeiat pentru a-și proteja datele. De exemplu, Google Maps nu vă permite să solicitați prea multe rezultate într-un interval scurt de timp.
- Efectuarea multor solicitări repetate către serverul unui site web mărește consumul de trafic, încetinind site-ul web pentru alți utilizatori și poate supraîncărca serverul astfel încât să nu mai poată răspunde deloc.

Înainte de a vă folosi abilitățile Python pentru web scraping, ar trebui să verificați întotdeauna politica de utilizare acceptabilă a site-ului țintă pentru a vedea dacă accesarea site-ului web cu instrumente automate reprezintă sau nu o încălcare a termenilor de utilizare.

Important: rețineți că următoarele tehnici pot fi ilegale atunci când sunt utilizate pe site-uri web care interzic scrapingul web.

Obținerea codului html al unei pagini web

Există câteva biblioteci de care veți avea nevoie, puteți accesa linia de comandă și le puteți instala.

```
pip install requests
pip install lxml
pip install bs4 # beautifulsoup4
```

```
import requests
import bs4
```

```
from bs4 import BeautifulSoup
from urllib.request import urlopen
```

```
url = http://... # url of web page
```

```
page = urlopen(url)
html = page.read().decode("utf-8")
soup = BeautifulSoup(html, "html.parser")
```

```
html_page = requests.get(URL, headers={"User-Agent":"Mozilla/5.0"})
# html_page.text - păstrează codul html al paginii web
```

Exerciții:

1. Alegeți un subiect ce vă interesează pe *wikipedia.org* și îndepliniți următoarele sarcini:
    - capturați titlul paginii;
    - capturați toate titlurile secțiunilor;
    - obțineți minim o imagine de pe acel site.
2. Accesați site-ul web: *http://books.toscrape.com/index.html* care este conceput special pentru testarea web scraping. Obțineți titlul fiecărei cărți care are o evaluare de 2 stele și, la sfârșit, să aveți doar o listă Python cu toate titlurile lor.
    - găsiți structura URL-ului pentru a parcurge fiecare pagină ;
    - parsați fiecare pagină din catalog;
    - găsiți ce etichetă/clasă reprezintă evaluarea cu stele ;
    - filtrați cu *if* evaluarea cu stele;
    - stocați rezultatele într-o listă.
3. Faceți cereri către minim 3 site-uri pentru a obține informația dorită (la alegere: date meteo, curs valutar, produse la reducere, rating etc.). Salvați rezultatele în fișier .csv.
   Notă: asigurați-vă că site-urile permit web scraping.