Lab 5 – Poszukiwanie bibliotek o określonej funkcjonalności (Web Scraping)

👤 Autor

Tomasz Królikowski, numer albumu: 153790

Zadanie znajduje się w repozytorium GIT pod adresem:

https://github.com/krolikowski80/studia_WSB/tree/main/Python/intro/zad_5

1. Cel zadania

Celem niniejszego laboratorium było zapoznanie się z bibliotekami do automatycznego pobierania danych z internetu (**web scraping**) i ich wykorzystanie w praktycznych przykładach.
Zastosowano dwie różne biblioteki: - **BeautifulSoup** – do prostego parsowania HTML, - **Scrapy** – do budowy wydajnych crawlerów.

Przedstawiono działanie obu rozwiązań na przykładzie rzeczywistych stron WWW, porównano ich możliwości oraz przygotowano działające skrypty demonstracyjne.

🧩 Wykorzystane biblioteki

🔷 BeautifulSoup

Biblioteka do prostego parsowania kodu HTML/XML.
Używana z `requests` do pobierania i analizowania zawartości stron internetowych.

🔷 Scrapy

Pełnoprawny framework do tworzenia crawlerów, obsługujący: - wiele adresów URL, - automatyczne przechodzenie po podstronach, - eksport danych do JSON/CSV/XML.

📁 Struktura katalogów

```
ZAD_5/
├── examples/
│   ├── bs4_example_1.py              # Nagłówki H2 z Wikipedii
│   ├── bs4_example_2.py              # Linki ze strony
│   ├── scrapy_example/
│   │   ├── scrapy_example/
│   │   │   ├── spiders/
│   │   │   │   └── quotes_spider.py # Spider Scrapy – cytaty
│   │   │   └── ...
│   │   └── scrapy.cfg
│   └── scrapy_example_advanced.py   # Uruchamianie spidera z poziomu Pythona
├── raport.md
└── README.md
```

▶️ Jak uruchomić przykłady

🔷 BeautifulSoup

📘 *bs4_example_1.py*

Pobiera nagłówki `<h2>` z Wikipedii:

```
cd examples
python bs4_example_1.py
```

📘 *bs4_example_2.py*

Pobiera pierwsze 10 linków ze strony:

```
python bs4_example_2.py
```

🔷 Scrapy

🕷️ *quotes_spider.py*

Zbiera cytaty i autorów z wielu stron i zapisuje je do pliku `quotes.json`:

```
cd examples/scrapy_example
scrapy crawl quotes -O quotes.json
```

🧪 *scrapy_example_advanced.py*

Uruchamia spidera z poziomu Pythona:

```
cd examples
python scrapy_example_advanced.py
```

💻 Wymagania

Zainstaluj wymagane biblioteki (najlepiej w środowisku wirtualnym):

```
pip install beautifulsoup4 lxml requests scrapy
```

🔗 Linki do dokumentacji

- BeautifulSoup – dokumentacja
- Scrapy – dokumentacja