Beautiful Soup

Заняття 6 Web Scraping

Лектор
Владислав Абрамов

ПЛАН ЗАНЯТТЯ

- Встановлення бібліотеки
- Знаходження потрібних HTML-тегів за допомогою BS
- Отримання тексту і значень атрибутів
- Написання парсера
- Застосування multiprocessing для парсингу

РОБОТА З БІБЛІОТЕКОЮ BEAUTIFUL SOUP

ВСТАНОВЛЕННЯ БІБЛІОТЕКИ

Beautiful Soup — це модуль Python для аналізу HTML- і XML-документів. Він створює дерево аналізу для розібраних сторінок, яке можна використовувати для вилучення даних.

Документація — https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Встановлення: pip install beautifulsoup4

ЗНАХОДЖЕННЯ ПОТРІБНИХ ТЕГІВ ЗА ДОПОМОГОЮ BS

Основні методи до використання:

- find()
- find_all()

Приклади пошуку:

- find(“h1”) — за тегом
- find(id=”someId”) — за id
- find(class_=”some-class”) — за класом
- find(string=”some text”) — за текстом

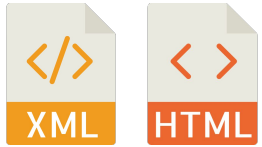
Підтримка regexp:

- find_all(string=re.compile(”<regexp>”)) — за текстом з регуляркою

/ПРАКТИКА:

- **Практика скрапінгу XML-файлу**
- **Практика скрапінгу HTML-сайту**

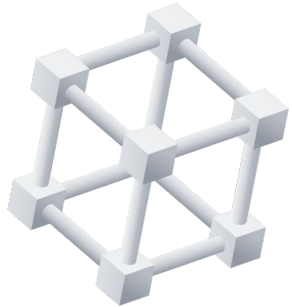
  Сайт — https://job.morion.ua/jobs/



ЯК НАПИСАТИ ПАРСЕР

НАПИСАННЯ ПАРСЕРА (ЗАГАЛЬНИЙ АЛГОРИТМ)

1) Отримання контенту сторінки
2) Створення об’єкта Beautiful Soup
3) Пошук потрібного контенту
4) Збереження інформації (CSV, JSON, Postgres)



ЗНАЙОМСТВО З MULTIPROCESSING

ПЕРЕВАГИ

Скрапінг вебсторінок часто містить затримки через завантаження сторінок або очікування відповідей від серверів.

Використання multiprocessing дає змогу запустити кілька запитів одночасно, що зменшує загальний час виконання завдань.

НЕДОЛІКИ

Кратне навантаження на сервер.
Етичність скрапінгу.
Ризик бану ip.

Документація — https://docs.python.org/3/library/multiprocessing.html#
Встановлення — pip install multiprocess
Built-in function **map**

/ПРАКТИКА:

Практика скрапінгу HTML-сайту паралельно

Сайт — https://job.morion.ua/jobs/
Порівнюємо швидкість скрапінгу

ДОМАШНЯ РОБОТА

Написати парсер, використовуючи Beautiful Soup для сайту https://www.bbc.com/sport
Потрібно для перших **5(!)** новин зібрати Related Topics.

На виході у вас має бути JSON-формат:

```
[
      {
            “Link”: “https://www.bbc.com/sport/tennis/articles/c4nnw5ydlnjo”,
            “Topics”: [“Tennis”]
      },
      {...} x 4
]
```

Q&A

r_d

ЗАВЖДИ Є КУДИ ЗРОСТАТИ