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

19/03/2024

Ćwiczenia laboratoryjne
sem. Letni 2023/2024

Modelowanie ekstrakcji danych

Co powinniśmy wziąć pod uwagę podczas tworzenie kodu służącego do ekstrakcji danych?

1. Dobre praktyki wynikające z inżynierii oprogramowania
2. Przemyślane zamodelowanie warstwy danych
3. Przypadki użycia

Dobre praktyki

- Modularyzacja kodu
  - Podział odpowiedzialności
  - Otwartość na rozszerzenia
  - OOP? https://pl.wikipedia.org/wiki/SOLID
- Optymalizacja względem wykorzystywanych zasobów
  - Przechowywanie danych w pamięci tylko wtedy, kiedy jest to niezbędnie potrzebne

Model danych

Model danych powinien być użyteczny oraz dobrze opisywać zbiór danych, który docelowo chcemy uzyskać po scrapingu.

Nie powinien natomiast być bardzo szczegółowy lub skrojony pod jeden rodzaj wyekstrahowanych danych.

Przykład

Jakie dane są nam potrzebne do porównania cen wybranego modelu butów pomiędzy dwoma sklepami internetowymi?

- Marka
- Model
- Numer seryjny/fabryczny
- Rozmiar
- Kolor
- Kategoria producenta
- Kategoria sklepu
- Rodzaj tkaniny
- Cena

Przykład

Jakie dane są nam potrzebne do porównania cen wybranego modelu butów pomiędzy dwoma sklepami internetowymi?

- Marka
- Model
- Numer seryjny/fabryczny
- Rozmiar
- Kolor
- Kategoria producenta
- Kategoria sklepu
- Rodzaj tkaniny
- Cena

Przypadki użycia

Przed rozpoczęciem prac nad pisaniem kodu źródłowego, należy zbadać dziedzinę problemu, który chcemy rozwiązać i zadać sobie następujące pytania:

- Czy znam źródła danych (strony) z których dokonana będzie ekstrakcja?
  - Czy w przyszłości będą potrzebne inne źródła?
- Czy problem jest jednolity i dobrze zdefiniowany?
  - Czy wymagane jest podzielenie go na mniejsze części?
- Jaka będzie częstotliwość wykonywania programu?
  - Automatyzacja + orkiestracja
- O jakiej ilości danych mówimy?
  - Skalowanie

Robots.txt – jak przestrzegać?

```
import urllib.robotparser


rp = urllib.robotparser.RobotFileParser()
rp.set_url("https://example.com/robots.txt")
rp.read()
if rp.can_fetch("MyBot", "https://example.com/page-to-scrape"):
    print("Scraping is allowed for this URL")
else:
    print("Scraping is not allowed for this URL according to robots.txt")
```

BeautifulSoup.select()

Szukanie z uwzględnieniem selectorów CSS:

https://www.educative.io/answers/beautiful-soup-select

KONIEC