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

Ekstrakcja danych z internetu

Ćwiczenia laboratoryjne
sem. Letni 2023/2024

05/03/2024

Organizacja zajęć

Ostatnie zajęcia : 18/06/2024

Zgodnie z Sylabusem AGH:
Ocena końcowa = ocena z ćwiczeń laboratoryjnych =
kolokwium zaliczeniowe z materiału zajęć (40%) +
oceny z projektów (40%) +
ocena z aktywności (20%)

Warunki zaliczenia

Kolokwium zaliczeniowe z materiału wykładów i ćwiczeń w formie testu jednokrotnego wyboru (ABCD).

W ramach ćwiczeń laboratoryjnych realizowane będą dwa projekty – oba o tej samej wadze oceny (20%) w stosunku do oceny końcowej.

Warunkiem uzyskania zaliczenia jest uzyskanie oceny pozytywnej z KAŻDEJ składowej oceny końcowej tzn. zarówno kolokwium, ocena z aktywności oraz oba projekty, muszą zostać ocenione na minimum 3.0 (50%).

Warunki zaliczenia

Ocena z aktywności:

Kompletność zadań laboratoryjnych = 70%
Każde brakujące zadanie : -10% (50% jest wymagane do zaliczenia ćwiczeń)

Zadania dodatkowe = 30%

Warunki zaliczenia

Ćwiczenia 1-3 : 05/03-19/03/2024
Wprowadzenie do Projektu 1 : 26/03/2024
Realizacja Projektu 1 : 26/03-16/04/2024
Ćwiczenia 4-6: 23/04-07/05/2024
Wprowadzenie do Projektu 2 : 14/05/2024
Realizacja Projektu 2 : 14/05-04/06/2024
Ćwiczenia 7-8: 11-18/06/2024

Powyższy ”terminarz” jest przykładowy i orientacyjny – będzie dostosowywany na bieżąco

Czas na realizację każdego projektu : 3-4 zajęcia
Tematy i zakres zostanie podany w trakcie zajęć

Tematyka zajęć

1. Podstawy ekstrakcji danych
    1. Narzędzia
    2. Modele ekstrakcji danych
    3. Tworzenie automatów
2. Przetwarzanie danych
    1. Oczyszczanie i normalizacja danych
    2. Odczytywanie dokumentów
    3. Przetwarzanie obrazów na tekst
3. Odczyt i zapisu języka naturalnego (modele Markowa, Natural Language Toolkit)
4. Korzystanie z API
    1. Ogólnodostępne źródła danych
5. Etyka i kwestie prawne

Zakres Projektu 1
Zakres Projektu 2

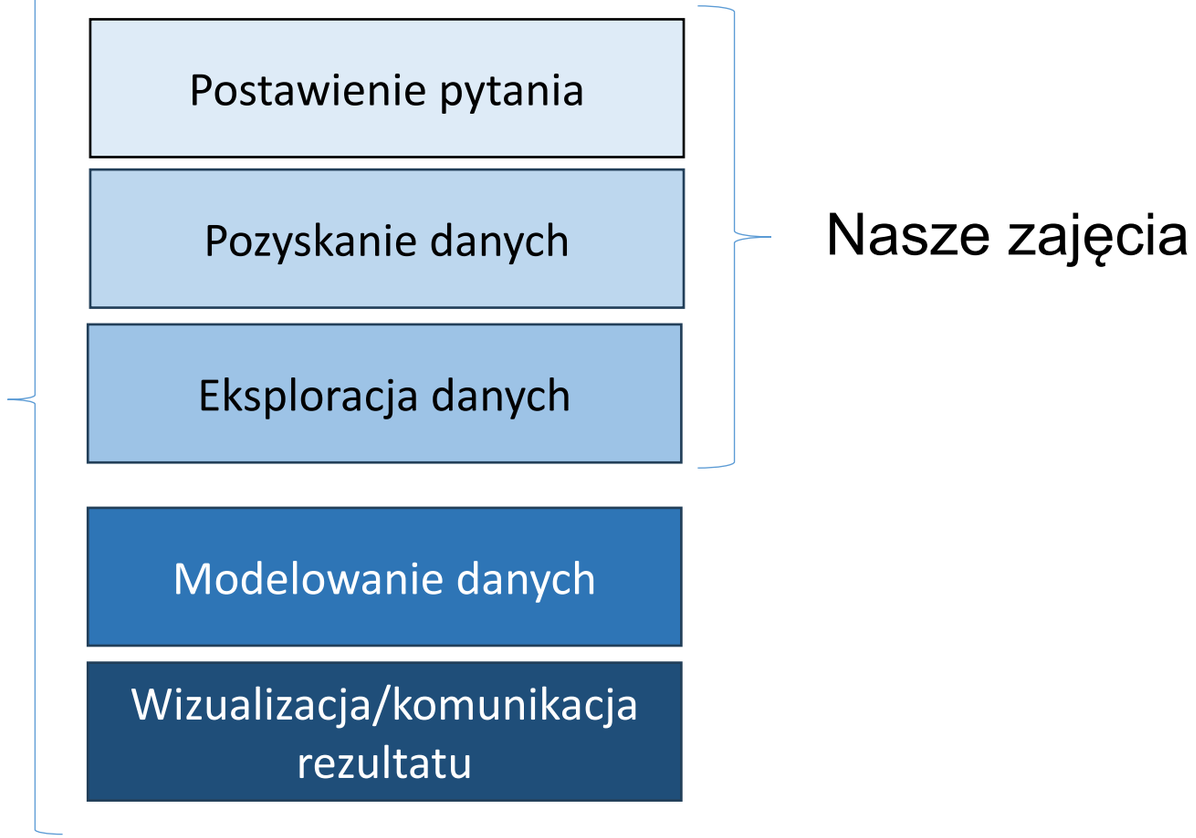
Projekt 2 będzie kontyunujacą działań z Projektu 1

Ekstrakcja danych: Motywacja

Zbiór aktywności służacy do uzyskania informacji/danych w celu dalszej obróbki.

Proces pozyskiwania informacji (wartości) z danych

Data Science Process



Ekstrakcja danych: Zastosowania

- Dalsza analiza pozyskanych danych w celu stworzenia systemu informacyjnego np. systemy rekomendacyjne
- Badania naukowe:
  - Ekonomia
    - Zachowanie konsumentów na rynku nieruchomości
  - Ekonometria
    - Porównanie cen usług hotelowych
  - Farmacja
    - Wprowadzenie nowego leku na rynek
  - Powiązanie wykluczenia cyfrowego ze stopniem urbanizacji
  - Analiza ofert pracy

Jakie narzędzia?

- Python3
  - BeautifulSoup
  - Requests
  - Scrapy
  - NumPy
  - Pandas
- SQL/NoSQL
- Git + GitHub
- RegEx

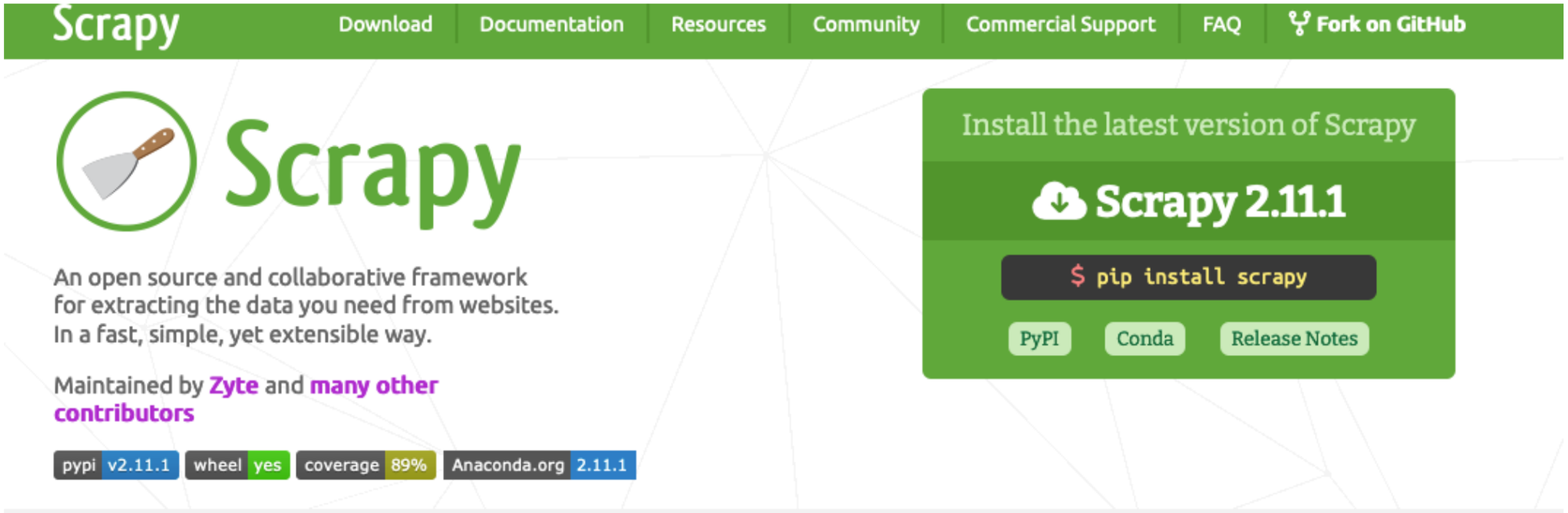
BeautifulSoup

https://pypi.org/project/beautifulsoup4/

Beautiful Soup is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree.

Scrapy

https://scrapy.org/



KONIEC