giganci programowania

Pokedex

Web scraping

- **Web scraping** to proces automatycznego pobierania danych ze stron internetowych przy użyciu skryptów. Python jest popularnym językiem do tego celu ze względu na jego bogaty ekosystem bibliotek ułatwiających pracę z HTML i HTTP.

Etapy web scrapingu

- **Wysyłanie żądań do stron internetowych** – pobieranie zawartości strony za pomocą HTTP.
- **Analizowanie struktury HTML** – przetwarzanie i interpretowanie pobranej treści w celu identyfikacji interesujących elementów.
- **Wyodrębnianie danych** – wybieranie danych z odpowiednich elementów, takich jak nagłówki, tabele, paragrafy itp.
- **Zapis danych** – przechowywanie danych w formacie przydatnym do analizy, np. w plikach CSV czy bazach danych.

Zastosowania web scrapingu

- **Monitorowanie cen w sklepach online** – automatyczne pobieranie danych o cenach produktów w różnych sklepach internetowych.
- **Gromadzenie opinii użytkowników** – zbieranie komentarzy, recenzji i ocen z forów i portali społecznościowych.
- **Tworzenie baz danych produktów lub usług** – aktualizowanie katalogów online poprzez automatyczne zbieranie informacji z różnych źródeł.
- **Analiza rynkowa i trendy** – śledzenie wiadomości, blogów czy artykułów z serwisów branżowych w celu identyfikacji nowych trendów.
- **Zbieranie danych szkoleniowych dla AI** – pozyskiwanie dużych zbiorów tekstów, obrazów czy danych liczbowych do trenowania modeli sztucznej inteligencji.
- **Porównywanie treści** – analizowanie różnic między treściami na różnych stronach, np. w celu badania konkurencji.
- **Zbieranie danych naukowych** – automatyczne gromadzenie informacji z publikacji naukowych i artykułów na potrzeby badań.

HTML

- HTML (HyperText Markup Language) to standardowy język znaczników używany do tworzenia i strukturyzowania treści na stronach internetowych. Kod HTML definiuje elementy strony, takie jak teksty, obrazy, linki, nagłówki, listy, formularze i inne elementy, które przeglądarka internetowa interpretuje i wyświetla użytkownikowi w postaci graficznej.

Cechy HTML

- **Język znaczników** – HTML składa się z **tagów**, które oznaczają różne części dokumentu. Każdy element HTML zaczyna się od tagu otwierającego (np.) i kończy tagiem zamykającym (np.), choć niektóre tagi mogą być samodzielne (np. <img>).
- **Struktura dokumentu** – HTML jest uporządkowany hierarchicznie, co oznacza, że elementy mogą zawierać inne elementy. Cały dokument HTML zazwyczaj zaczyna się od deklaracji <!DOCTYPE html>, po której następuje znacznik <html>.
- **Nagłówki i treść** – HTML zawiera tagi do definiowania nagłówków (np. <h1>, <h2>), paragrafów ( ), list (<ul>, <ol>), linków (<a>) i wielu innych elementów.
- **Atrybuty** – Tagom można przypisywać **atrybuty**, które dodają im dodatkowych informacji lub modyfikują ich działanie. Na przykład, tag <a> może mieć atrybut href, który wskazuje adres URL, do którego prowadzi link (Kliknij tutaj).

Przykład dokumentu HTML

```html
<!DOCTYPE html>
<html lang="pl">
<head>
    <meta charset="UTF-8">
    <title>Moja Pierwsza Strona</title>
</head>
<body>
    <h1>Witaj na mojej stronie!</h1>
    <p>To jest przykładowy paragraf. HTML pozwala tworzyć strony internetowe.</p>
    <a href="https://example.com">Odwiedź tę stronę</a>
</body>
</html>
```

- **<!DOCTYPE html>** – określa typ dokumentu jako HTML5, co jest najnowszą wersją HTML.
- **<html>** – otacza cały dokument HTML.
- **<head>** – zawiera metadane strony, takie jak tytuł (<title>) i informacje o kodowaniu znaków (<meta>).
- **<body>** – główna część dokumentu, w której umieszczane są treści wyświetlane na stronie.
- **<h1>** – nagłówek najwyższego poziomu, używany do tytułów.
- **<p>** – paragraf, używany do wyświetlania bloków tekstu.
- **<a>** – link do innej strony lub zasobu.