

1. Wstęp do przetwarzania XML

XML, co oznacza eXtensible Markup Language, to język znaczników zaprojektowany do przechowywania i przekazywania danych. XML został opracowany przez World Wide Web Consortium (W3C) i jest powszechnie stosowany w dziedzinach takich jak serwisy internetowe, edytorstwo, informatyka naukowa, a także wszędzie tam, gdzie potrzebne jest uniwersalne formatowanie danych.

Podstawowe cechy XML:

1. **Czytelność dla ludzi i maszyn:** XML jest prosty do odczytania zarówno dla ludzi, jak i dla maszyn. Jego struktura jest hierarchiczna i dobrze zorganizowana, co sprawia, że jest łatwy do zrozumienia.
2. **Rozszerzalność:** XML jest "rozszerzalny", co oznacza, że możemy definiować własne elementy i atrybuty zgodnie z naszymi potrzebami.
3. **Samoopisujący:** Dokumenty XML są samoopisujące - struktura i znaczenie danych są zawarte w samych danych.
4. **Uniwersalność:** XML jest platformowo niezależny, co oznacza, że może być czytany przez różne systemy i aplikacje.

```
<?xml version="1.0" encoding="UTF-8"?>
<pracownicy>
  <pracownik id="1">
    <imie>Jan</imie>
    <nazwisko>Kowalski</nazwisko>
    <stanowisko>Manager</stanowisko>
  </pracownik>
  <pracownik id="2">
    <imie>Anna</imie>
    <nazwisko>Nowak</nazwisko>
    <stanowisko>Developer</stanowisko>
  </pracownik>
</pracownicy>
```

2. Parseery DOM i SAX:

Parseery DOM i SAX to dwa różne podejścia do interpretowania dokumentów XML w Pythonie. DOM (Document Object Model) jest techniką, która wczytuje cały dokument XML do pamięci i umożliwia jego manipulację jako zestaw obiektów zagnieżdżonych w drzewie. SAX (Simple API for XML) to podejście sekwencyjne, które interpretuje dokument XML jako serię zdarzeń, takich jak "początek tagu", "koniec tagu", "znaki" itp.

3. Biblioteka xml.etree.ElementTree:

Moduł xml.etree.ElementTree w Pythonie implementuje lekkie i efektywne API dla parsowania i tworzenia dokumentów XML. Jest to jeden z najpopularniejszych modułów do manipulacji XML w Pythonie. Za jego pomocą można tworzyć, modyfikować i nawigować po drzewach XML, a także serializować drzewa do łańcuchów.

4. Document Object Model API – mindom (implementacja minimal):

"mindom", jest częścią standardowej biblioteki Pythona. Mindom to minimalna implementacja API DOM dla Pythona. Pozwala na prostą manipulację dokumentami XML za pomocą struktury drzewa obiektów.

5. Parseery SAX2:

SAX2 to druga wersja interfejsu SAX dla Pythona, która dostarcza bardziej rozbudowany zestaw funkcji w porównaniu do oryginalnego SAX. SAX2 obsługuje takie funkcje jak filtrowanie i obsługa przestrzeni nazw.

6. Handler w SAX

Handler w SAX to obiekt, który jest wywoływany przez parser SAX, kiedy natrafi na różne elementy dokumentu XML (np. początek tagu, koniec tagu, tekst itd.). Można zdefiniować własne handlersy, które wykonują konkretną logikę dla różnych zdarzeń.

7. Użycie expat do szybkiego parsowania XML:

Expat to parser XML napisany w języku C, który jest dostępny w Pythonie za pośrednictwem modułu `xml.parsers.expat`. Expat jest bardzo szybkim parserem, który korzysta z modelu zdarzeń podobnego do SAX. Za jego pomocą można zdefiniować funkcje zwrotne (callback functions), które zostaną wywołane przy różnych zdarzeniach, takich jak początek tagu, koniec tagu, itp. Dzięki temu, że Expat jest napisany w C, jest zazwyczaj szybszy od innych parserów napisanych w Pythonie.