# ORW Ćwiczenie 3 – Ekstrakcja danych z sieci WWW

Autor: Łukasz Pawłowski

Grupa: 125NCI\_B

Link do projektu: <https://github.com/mojzesz/orw-3>

Język programowania: Ruby

**Zadania do zrealizowania:**

**ORW ćw. 3**

1. Za pomocą wybranego przez siebie języka programowania (wraz z frameworkiem) zademonstruj ekstrakcję wybranej strony WWW.
2. Bądź gotowy do prezentacji możliwości i modyfikacji napisanego programu.

**Przykładowe pytania ORW ćw. 3**

* Co to jest JSON? Omów podstawowe struktury.

JSON – JavaScript Object Notation – jest sposobem zapisu obiektów w ciągu znaków. Jest stosunkowo prosty do opanowania zarówno przez ludzi, jak i interpretowany i generowany przez komputery.

* Co to jest XML? Omów strukturę i zastosowanie.

Jest to język znaczników do prezentowania danych w strukturalny sposób, wykorzystywany zarówno do przechowywania danych do interpretacji jak i przechowywania konfiguracji oraz uporządkowanych struktur danych. Przydatny przy programowaniu oraz używaniu SOAP. Powinien zawierać dokładnie jeden element główny.

* Co to jest XPath? Jak XPath jest wykorzystywany we współpracy z XML?

To język służący do określania adresu części dokumentu XML. We współpracy z XML pozwala na określenie adresu danego elementu, pobranie jego danych oraz atrybutów.

* Kiedy dokument XML jest *well-formed*?

Kiedy podąża za wyznaczonymi wytycznymi standardu (podany encoding, standalone, parametry, wartości elementów, zagnieżdżenia, zamykanie elementów)

* Co to jest DTD? Podaj cel stosowania i przykład.

Rodzaj dokumentu definiujący formalną strukturę dokumentów, którego spełnienie oznacza, że dokument powstały na jego definicjach jest poprawny (*ang. valid*). Dotyczy najczęściej formatów XML, HTML lub XHTML i pochodnych.

* Omów na przykładzie ideę web-scrapingu.

Web-scraping jest wtedy, gdy przy pomocy zautomatyzowanego programu ze strony internetowej, która nie udostępnia np. API do tego celu, pobierane są dane, normalnie przeznaczone tylko do wyświetlania dla odwiedzającego. Z web scrapingu mogą korzystać m.in. wyszukiwarki internetowe – w celu indeksowania zawartości oraz oceny jakości stron.

# Realizacja zadań

Zadanie zostało zrealizowane przy pomocy Ruby oraz gemu Mechanize. Gem Mechanize jest przeznaczony stricte do celów udawania przeglądarki i wyciągania z niej danych. Przedstawiony przykład najpierw: podmienia user-agent użytkownika, następnie wyszukuje formy logowania, loguje się podanymi danymi, przechodzi do konkretnego adresu i stamtąd pobiera plik PDF dostępny dla zalogowanych użytkowników. Po wszystkim wylogowuje się z systemu. Cały proces jest zautomatyzowany i wystarczy wydać komendę uruchomienia lub dodać zadanie do CRON.