Лабораторна робота № 1.

Вивчення базових операцій обробки XML-документів

Метою роботи є здобуття практичних навичок створення програм, орієнтованих на обробку XML-документів засобами мови Python.

Завдання роботи полягає у наступному:

1. Виконати збір інформації зі сторінок Web-сайту за варіантом.
2. Виконати аналіз сторінок Web-сайту для подальшої обробки текстової та графічної інформації, розміщеної на ньому.
3. Реалізувати функціональні можливості згідно вимог, наведених нижче.

Функціональні вимоги

1. На основі базової адреси Web-сайту виконати обхід наявних сторінок сайту, відокремлюючи текстову та графічну інформацію від тегів HTML. Пошук вузлів виконувати засобами XPath. Наступну сторінку для аналізу **цього ж сайту** обрати як одне із гіперпосилань на даній сторінці (тег <a href=”url”/>). Обмежитись аналізом 20 сторінок сайту. Зберегти XML у вигляді файлу. Формат XML-документу:

```
<data>
  <page url=”wwww.server.com/index.hml”>
  <fragment type=”text”>
…. знайдений текст
 </fragment>
  <fragment type=”image”>
…. url зображення
 </fragment>
 </page>
 <page url=”wwww.server.com/index1.hml”>
  <fragment type=”text”>
…. знайдений текст
 </fragment>
  <fragment type=”image”>
…. url зображення
 </fragment>
 </page>
…
</data>
```

2. Виконати аналіз отриманих даних засобами XML згідно варіанту та вивести результати у консольне вікно. Відбір вузлів та розрахунки за варіантом **виконувати засобами XPath**.

3. Проаналізувати вміст Web-сторінок інтернет-магазину (див. варіант). Отримати ціну, опис та зображення для 20 товарів з нього за допомогою DOM-парсеру та мови XPath для пошуку відповідних вузлів. Результат записати в XML-файл.

4. Перетворити отриманий XML-файл у XHTML-сторінку за допомогою мови XSLT. Дані подати у вигляді XHTML-таблиці та записати його у файл.

Вимоги до інтерфейсу користувача

Використовувати консольний (текстовий) інтерфейс користувача.

Вимоги до інструментарію

1. Мова програмування – Python 3
2. Бібліотека сканування та отримання даних - пошуковий робот Scrapy.
3. Бібліотека для генерації та перетворення XML – libxml2 або (xml.dom, xml.xpath.. - вбудовані пакети), lxml – на вибір студента.
4. Середовище розробки програмного забезпечення – PyCharm Community Edition (*опціонально*)

Вибір варіанту

Робота виконується індивідуально. Варіант обирається шляхом вибору останніх двох цифр номеру залікової книжки студента.

Варіанти

№ вар.	Базова сторінка (завдання 1)	Зміст завдання 2	Адреса інтернет-магазину (завдання 3)
1.	www.kpi.ua	Максимальна кількість текстових фрагментів	www.rozetka.ua
2.	www.ukr.net	Середня кількість текстових фрагментів	www.repka.ua
3.	www.bigmir.net	Мінімальна кількість графічних фрагментів	www.sokol.ua

4.	www.korrespondent.net	Кількість текстових фрагментів по кожному документу	www.hotline.ua
5.	www.football.ua	Кількість графічних фрагментів по кожному документу	www.moyo.ua
6.	www.isport.ua	Вивести список гіперпосилань	www.portativ.ua
7.	www.tsn.ua	Мінімальна кількість графічних фрагментів	www.wallet.ua
8.	www.golos.ua	Середня кількість текстових фрагментів	www.petmarket.ua
9.	www.xsport.ua	Вивести список гіперпосилань	www.meblium.com.ua
10.	www.ridna.ua	Кількість графічних фрагментів по кожному документу	www.veliki.com.ua
11.	www.ukraine-is.com	Вивести список гіперпосилань	www.instrument.in.ua
12.	www.uartlib.org	Середня кількість текстових фрагментів	www.hozmart.com.ua
13.	www.doroga.ua	Мінімальна кількість графічних фрагментів	www.freedelivery.in.ua
14.	www.stejka.com	Вивести список гіперпосилань	www.mebli-lviv.com.ua
15.	www.ua.igotoworld.com	Середня кількість графічних фрагментів	https://allo.ua
16.	www.posolstva.org.ua	Максимальна кількість текстових фрагментів	www.odissey.kiev.ua
17.	www.uahotels.info	Середня кількість текстових фрагментів	www.zvetsad.com.ua
18.	www.osvita.ua	Мінімальна кількість графічних фрагментів	www.auto-store.kiev.ua
19.	www.shkola.ua	Кількість текстових фрагментів по кожному документу	www.tennismag.com.ua

20.	www.ostriv.in.ua	Вивести список гіперпосилань	www.fishing-mart.com.ua

Вимоги до оформлення лабораторної роботи у електронному вигляді

Опис лабораторної роботи у репозиторії включає: назву лабораторної роботи, варіант студента, 2-3 копії екранних форм (screenshots).

Контрольні запитання

1. Дати визначення well-formed та коректному XML-документу.
2. Назвати способи обробки XML-документів.
3. Охарактеризувати мову XPath.
4. Охарактеризувати мову XSLT.