**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ**

**“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО”**

Факультет прикладної математики

Кафедра програмного забезпечення комп’ютерних систем

**Лабораторна робота № 1**

з дисципліни “Бази даних”

спеціальність 121 – Програмна інженерія

на тему: **Вивчення базових операцій обробки XML-документів**

**Студент**

**групи** **КП-81** Шапошник Б.І. \_\_\_\_\_\_\_\_\_\_

(підпис)

**Викладач**

**к.т.н, доцент кафедри**

**СПіСКС** Петрашенко А.В. \_\_\_\_\_\_\_\_\_\_

(підпис)

Київ - 2021

*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>

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

1. 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.

1. 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*)

**Варіант 21**

У першому завданні було замінено сайт [www.ostriv.in.ua](http://www.ostriv.in.ua) на [www.kpi.ua](http://www.kpi.ua).

|  |  |  |
| --- | --- | --- |
| [www.kpi.ua](http://www.kpi.ua) | Вивести список гіперпосилань | [www.fishing-mart.com.ua](http://www.fishing-mart.com.ua) |

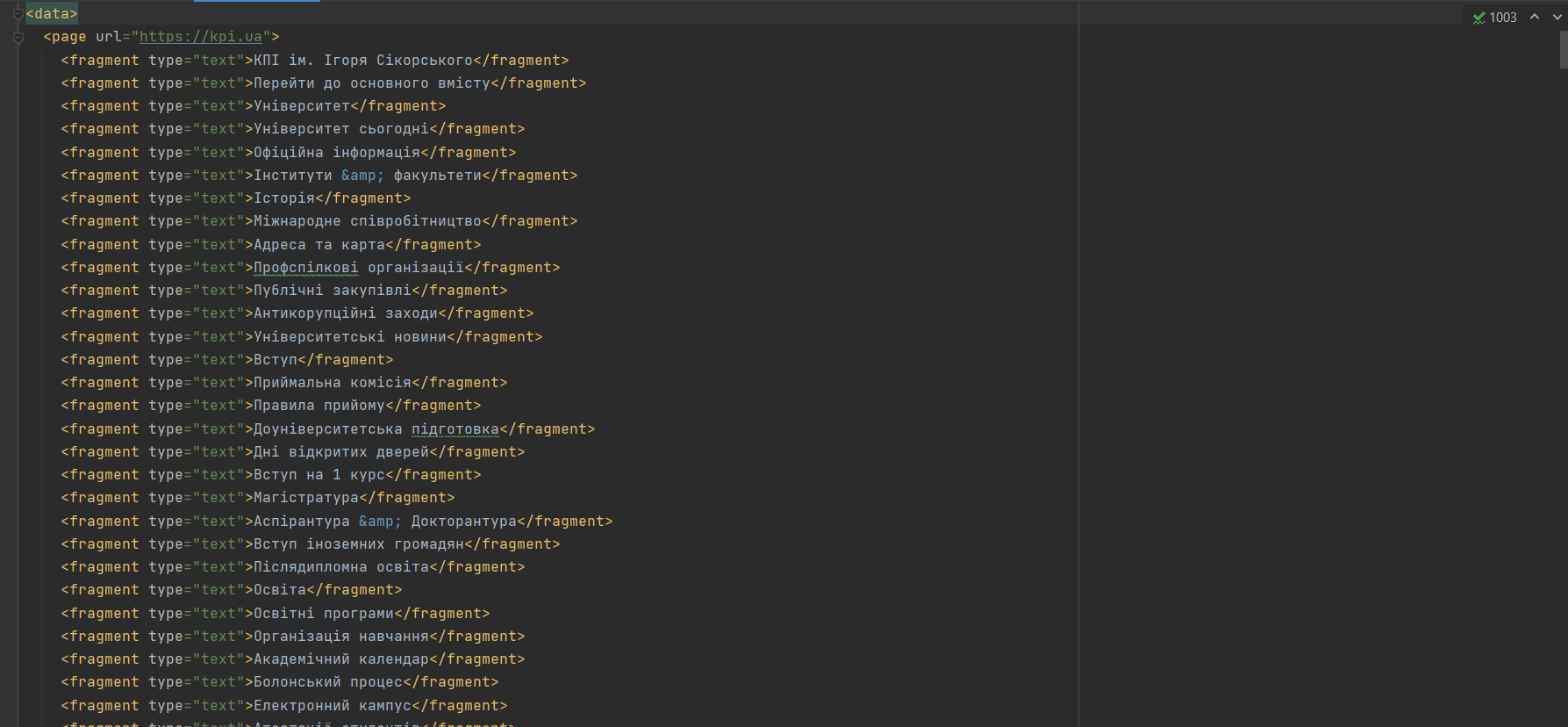
Посилання на гітхаб:

<https://github.com/BogdanShaposhnik/db_2_labs/tree/master/db_lab1>

Приклади роботи програми:

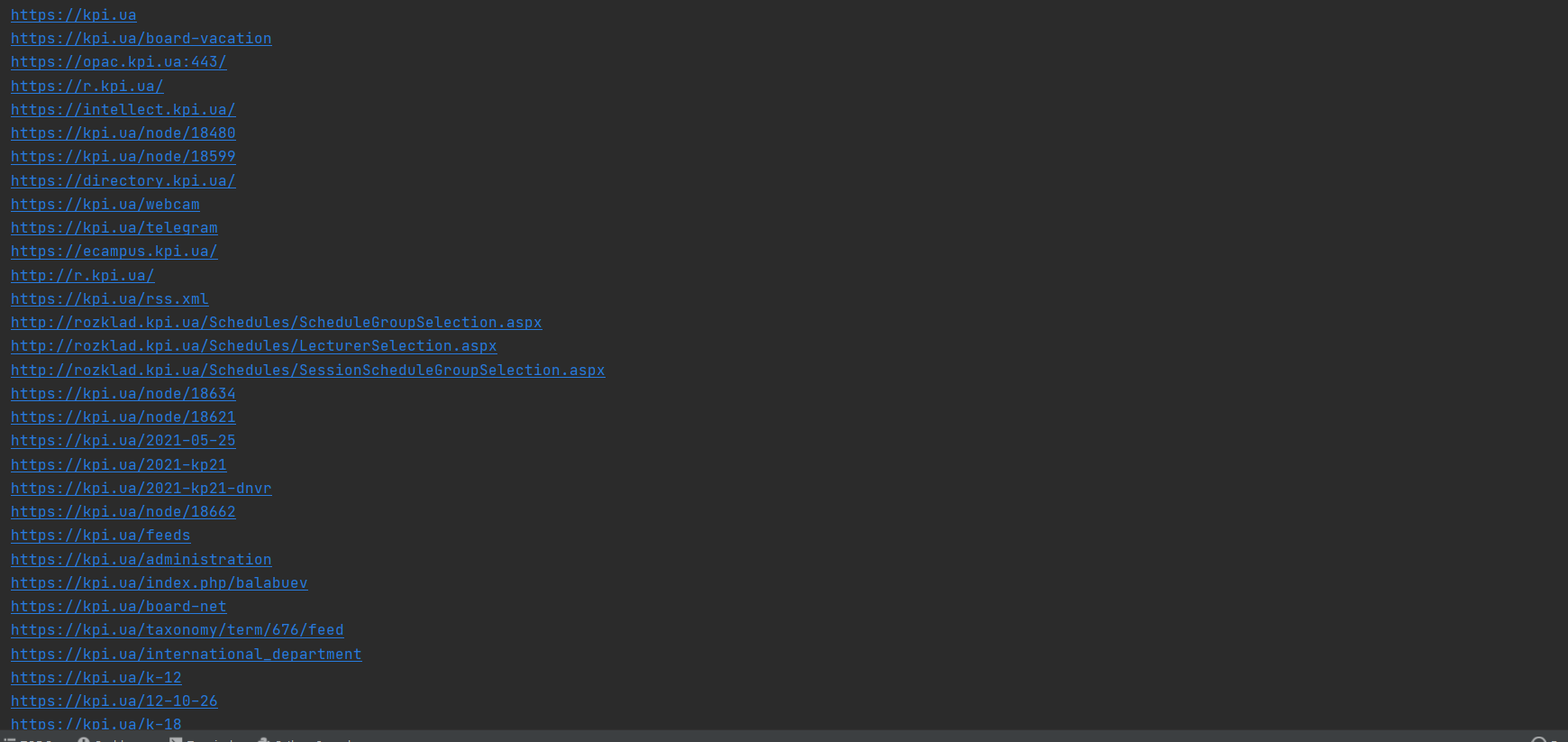
1 завдання:

Текстова та графічна інформація:



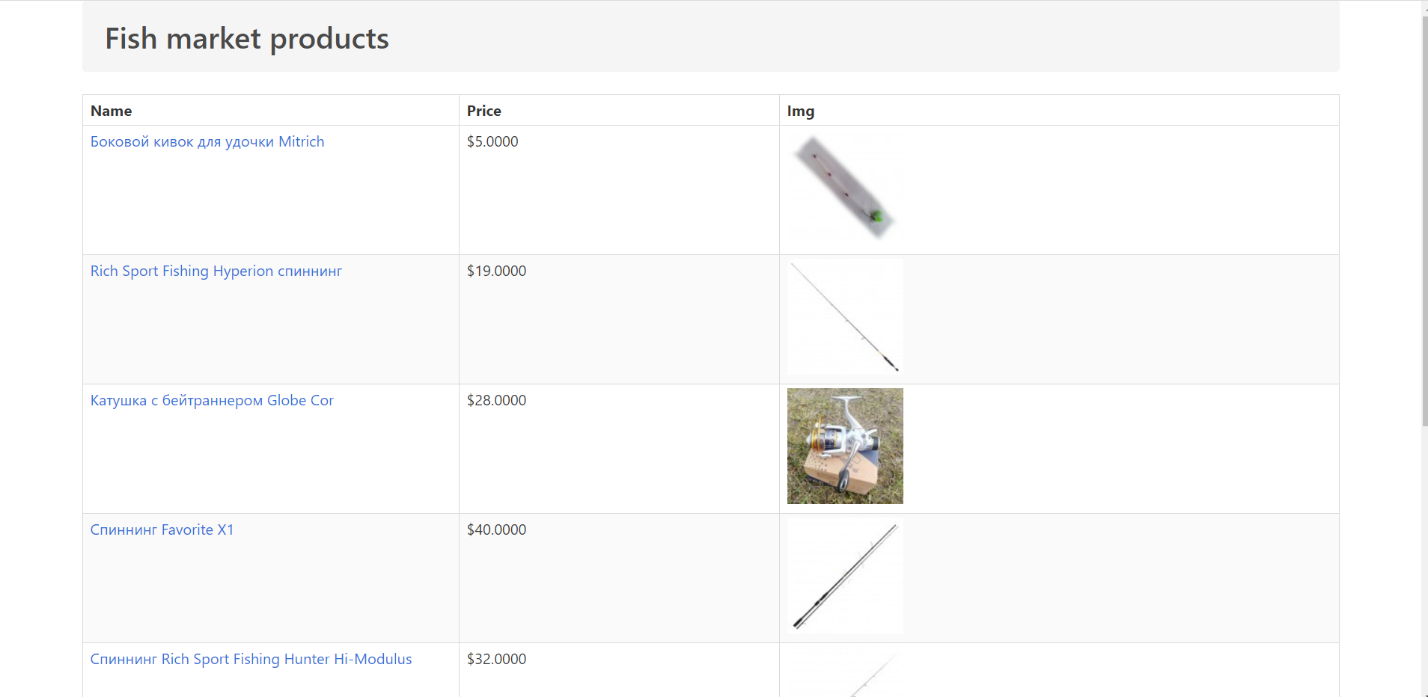
2 завданння:

Список посилань:



3 завдання:

Html file з продуктами



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.
5. Коректний документ ([англ.](https://uk.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D1%96%D0%B9%D1%81%D1%8C%D0%BA%D0%B0_%D0%BC%D0%BE%D0%B2%D0%B0" \o ") *well-formed document*) відповідає всім синтаксичним правилам XML.

* Лише один елемент у корені.
* Непорожні елементи розмічено початковим та кінцевим тегами (наприклад, <пункт>Пункт 1</пункт>). Порожні елементи можуть позначатися «закритим» тегом, наприклад <IAmEmpty />. Така пара еквівалентна <IAmEmpty></IAmEmpty>.
* Один елемент не може мати декілька атрибутів з однаковою назвою. Значення атрибутів перебувають або в одинарних ('), або у подвійних (") лапках.
* Теги можуть бути вкладені, але не можуть перекриватись. Кожен некореневий елемент мусить повністю перебувати в іншому елементі.
* Документ має складатися тільки з правильно закодованих дозволених символів Юнікоду. Єдиними кодуваннями, які обов'язково має розуміти XML-процесор, є UTF-16 та UTF-8. Фактичне та задеклароване кодування ([англ.](https://uk.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D1%96%D0%B9%D1%81%D1%8C%D0%BA%D0%B0_%D0%BC%D0%BE%D0%B2%D0%B0" \o "Англійська мова) *character encoding*) документа мають збігатись. Кодування може бути задекларовано ззовні, як у заголовку «Content-Type» при передачі за протоколом [HTTP](https://uk.wikipedia.org/wiki/HTTP), або в самому документі використанням явної розмітки на самому початку документа. У разі відсутності інформації про кодування документ має бути в кодуванні UTF-8 (або його підмножині ASCII).

1. До XML-обробки відносяться форматування, синтаксичний аналіз, редагування, перевірка коректності і перетворення в інші формати.

До традиційних технологій обробки XML-документів належать такі три технології:

* Написання програм на мові програмування із використанням API [SAX](https://uk.wikipedia.org/wiki/SAX).
* Написання програм на мові програмування із використанням API [DOM](https://uk.wikipedia.org/wiki/DOM).
* Застосування механізму перетворення та фільтра.

До новіших технологій, що почали здобувати поширення останнім часом належать:

* Активний аналіз.
* Зв'язування даних

1. **XPath**, або **XML Path Language** — це [мова запитів](https://uk.wikipedia.org/wiki/%D0%9C%D0%BE%D0%B2%D0%B0_%D0%B7%D0%B0%D0%BF%D0%B8%D1%82%D1%96%D0%B2), для вибору вузлів з [XML](https://uk.wikipedia.org/wiki/XML) документів, або для обчислення величин (наприклад, [рядкових](https://uk.wikipedia.org/wiki/%D0%A0%D1%8F%D0%B4%D0%BE%D0%BA_(%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D1%83%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F)), числових або [булевих](https://uk.wikipedia.org/wiki/%D0%9B%D0%BE%D0%B3%D1%96%D1%87%D0%BD%D0%B8%D0%B9_%D1%82%D0%B8%D0%BF_%D0%B4%D0%B0%D0%BD%D0%B8%D1%85)) на основі вмісту [XML](https://uk.wikipedia.org/wiki/XML) документа.
2. **XSLT** (*e****X****tensible****S****tylesheet****L****anguage****T****ransformations*) — мова перетворення [XML](https://ru.wikipedia.org/wiki/XML)-документів.