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. Частина 2*”
на тему “*Вивчення базових операцій обробки XML-документів*”

Виконав
студент III курсу
групи КП-81
Бухаленков Дмитро Олександрович
Варіант 2

Зарахована: Петрашенко А. В.

Київ 2021

Мета роботи: 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 за варіантом.

Варіант 2

(зайт змінено на kpi.ua тому що на ukr.net немає зображень і всі гіперпосилання ведуть на інші сайти)

Перший сайт	Завдання 2	Інтернет-магазин
kpi.ua	Середня кількість текстових фрагментів	repka.ua

Результати:

Код для “павука”, що вибирає дані з kpi.ua:

```
class KPISpider(Spider):
    name = 'kpi'
    start_urls = ['https://kpi.ua/']
    allowed_domains = ['kpi.ua']
    max_pages = 20
    filename = 'output/kpi.xml'

    def __init__(self, **kwargs):
        super().__init__(**kwargs)
        self.visited_pages = []

    def parse(self, response):
        if len(self.visited_pages) >= self.max_pages:
            raise exceptions.CloseSpider('Page limit exceeded.')
        if response.url not in self.visited_pages:
            self.visited_pages.append(response.url)
            yield self.parse_page(response)
        urls = Selector(response=response).xpath('//a/@href').getall()
        for url in urls:
            yield Request(url=urljoin(response.url, url),
callback=self.parse)
```

```
    def parse_page(self, response):
        selector = Selector(response=response)
        text_data = selector.xpath('//h1[contains(@class,
\'page-title\')]/span/text() | //div[contains(@class, '
                                   '\'node__content\')]/div/p').getall()
        images = selector.xpath('//img/@src').getall()
        return {
            'url': response.url,
            'text': [t.strip() for t in text_data],
            'images': [response.urljoin(src) for src in images]
        }

    @classmethod
    def get_average_texts_count(cls):
        return
etree.parse(cls.filename).xpath("count(//page/fragment[@type='text']) div
count(//page)")
```

Код для павука інтернет магазиину repka.ua:

```
class RepkaSpider(Spider):
    name = 'repka'
    start_urls = ['https://repka.ua/products/noutbuki/?brands=73,']
    max_pages = 20
    filename = 'output/repka.xml'

    def parse(self, response):
        links =
Selector(response=response).xpath("//div[@class='product-item-name']/a/@hr
ef").getall()[:RepkaSpider.max_pages]
        for link in links:
            yield Request(url=urljoin(response.url, link),
callback=self.parse_laptops)

    def parse_laptops(self, response):
        selector = Selector(response=response)
        yield {
            'name':
selector.xpath("//h1[@class='page-title']/span//text()").get(),
            'price':
selector.xpath("//span[@class='price-wrapper']/@data-price-amount").get(),
            'image':
selector.xpath("//img[@class='fotorama__img']/@src").get(),
            'description':
selector.xpath("normalize-space(//div[@id='product_description']/div[@clas
s='box'])").get()
        }

    @staticmethod
    def create_xhtml():
        parsed = etree.parse(RepkaSpider.filename)
        template = etree.parse('template.xsl')
        transform = etree.XSLT(template)
```

```
        table = transform(parsed)
        with open('output/table.xhtml', 'w') as f:
            f.write(etree.tostring(table, pretty_print=True).decode())
```

XSLT для трансформації:

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:template match="/">

 <html xmlns="http://www.w3.org/1999/xhtml">
  <head>
   <title>Laptops</title>
   <link rel="stylesheet"
href="https://stackpath.bootstrapcdn.com/bootstrap/4.5.2/css/bootstrap.min
.css"
         crossorigin="anonymous"/>
  </head>
  <body>
   <div class="container mt-3">
    <table class="table table-bordered">
     <tbody>
      <tr>
       <th>Зображення</th>
       <th>Назва</th>
       <th>Ціна</th>
       <th>Опис</th>
      </tr>
      <xsl:for-each select="//laptop">
       <tr>
        <td>
         <img src="{image}" width="250px"/>
        </td>
        <td>
         <strong>
          <xsl:value-of select="@name"/>
         </strong>
        </td>
        <td>
         <xsl:value-of select="price"/>
         грн
        </td>
        <td>
         <xsl:value-of select="description"/>
        </td>
       </tr>
      </xsl:for-each>
     </tbody>
    </table>
   </div>
  </body>
 </html>

</xsl:template>
</xsl:stylesheet>
```

Повний код доступний в репозиторію.

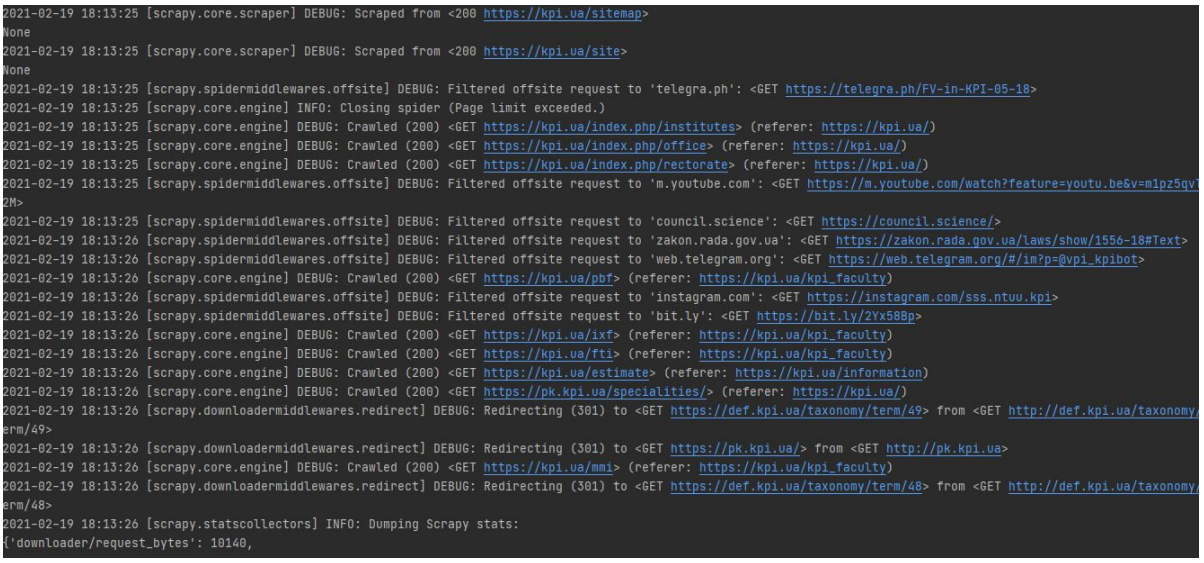

Рис. 1 Scrapy парсить kpi.ua

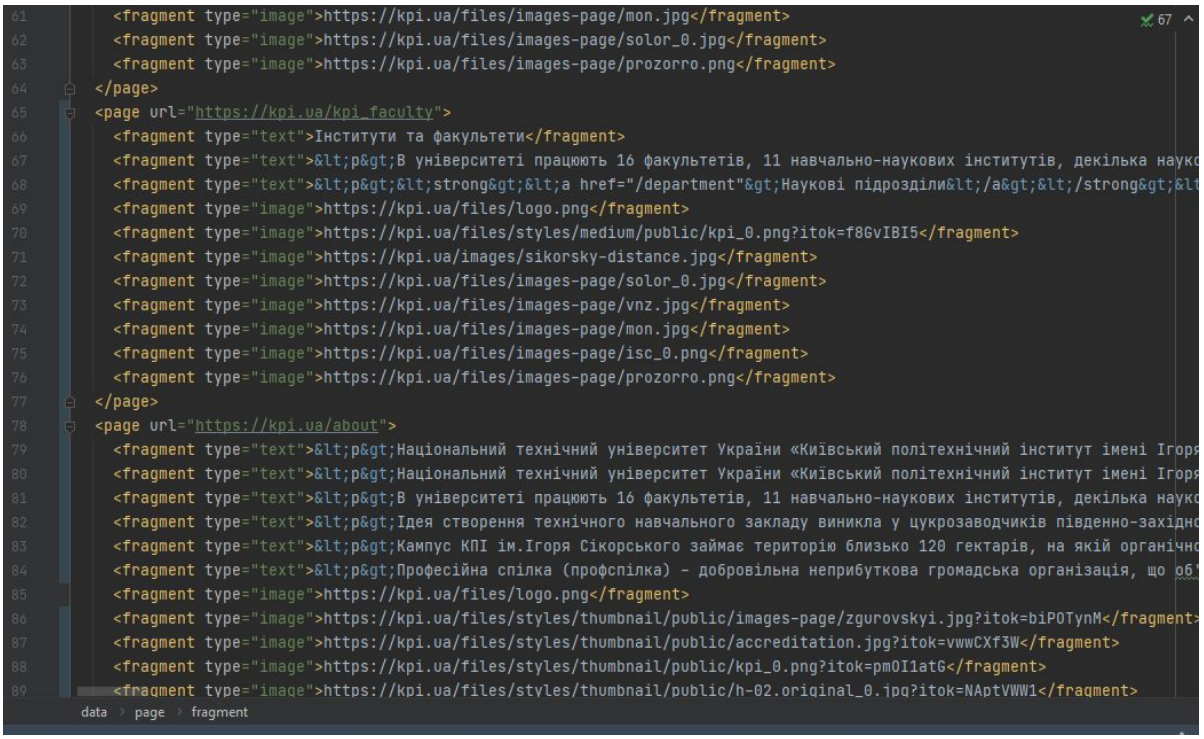

Рис. 2 результуючий XML файл

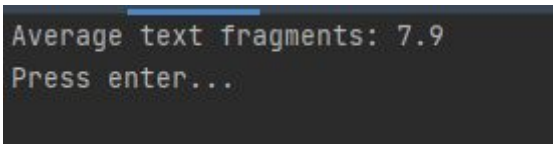

Рис. 3 за допомогою XPath підрахумо середню кількість текстових фрагментів по сторінках

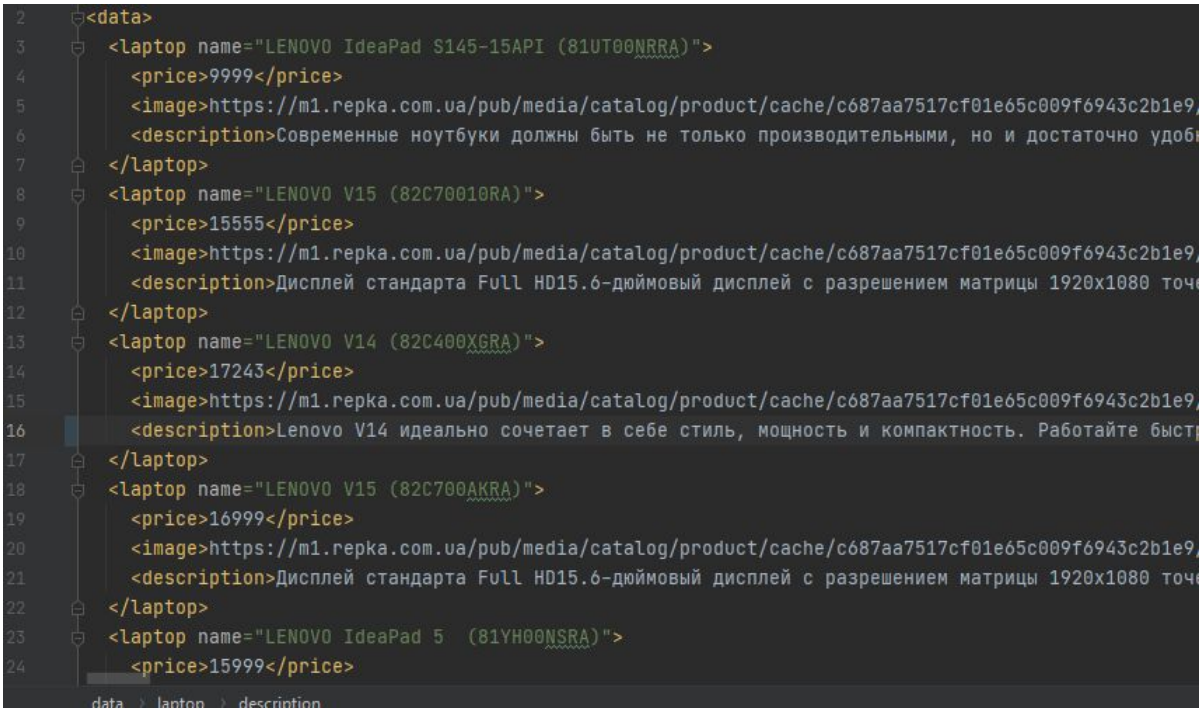

Рис. 4 XML файл з даними про ноутбуки на сайті repka.ua

	**LENOVO V15 (82C700AKRA)**	16999 грн	Дисплей стандарта Full HD15.6-дюймовый дисплей с разрешением матрицы 1920x1080 точек и LED-подсветкой обеспечивает высокую яркость и четкость изображения. ПроизводительностьНаслаждайтесь любимыми играми и приложениями благодаря высокопроизводительному процессору Intel Core 10-го поколения и быстрому накопителю. Поддержка графики высокого разрешенияДобейтесь высочайшего качества изображения, подключив устройство к телевизору или к другому монитору с помощью разъема HDMI. Высокая скорость передачи данныхНоутбук оборудован USB 3.1, что позволяет обмениваться данными с другими устройствами со скоростью, которая в 10 раз превышает скорость интерфейсов USB более ранних версий. Встроенная веб-камераВстроенная веб-камера обеспечивает четкое видеоизображение, создавая эффект полного присутствия на веб-конференциях и в интерактивных видеочатах.
	**LENOVO IdeaPad 5 (81YH00NSRA)**	15999 грн	Забудьте о стикерах для заклеивания камеры. Защитная шторка веб-камеры ноутбука IdeaPad 5i (14) оградит вас от вездесущих хакеров. Кроме того, IdeaPad 5i (14) оснащен устройством распознавания отпечатков пальцев на кнопке питания. Ноутбук можно настроить так, что он не будет реагировать на взаимодействия, пока не считает вашу биометрическую подпись. При разработке дисплея для ноутбука IdeaPad 5i (14) мы стремились максимально увеличить площадь экрана для удобства пользователей, уменьшая его рамки и расширяя область просмотра. Мы называем это коэффициентом активной площади, или соотношением размеров экрана и окружающей его рамки. Ноутбук IdeaPad 5i (14) имеет невероятный коэффициент активной площади — 90 %. льзования.
	**LENOVO IdeaPad S145-15API (81UT00MFRA)**	17599 грн	Всегда готовБлагодаря процессору AMD ноутбук IdeaPad S145 поможет вам решать любые задачи. Откройте в себе страсть к путешествиямМодель IdeaPad S145 весит всего 1.85 кг и поэтому идеально подходит для путешествий. Узкая рамка делает дизайн более лаконичным и увеличивает видимую область экрана. Ноутбук представлен в разных цветах с фактурной или глянцевой отделкой и при этом продается по привлекательной цене. Превосходный звук, отличное изображениеНезависимо от того, смотрите ли вы видео, слушаете потоковую музыку или общаетесь в видеочате, вам непременно понравится звучание, которое обеспечивают динамики IdeaPad S145 с поддержкой технологии Dolby Audio. Кроме того, вас также приятно удивит изображение на 15.6-дюймовом дисплее с матовым покрытием.
	**LENOVO ThinkBook 14 G2**	16399 грн	Максимальная производительность Максимально реализуйте вычислительную мощь мобильного процессора AMD Ryzen™ серии 4000 для решения самых ресурсоемких задач. Независимо от того, анализируете ли вы данные или

Рис. 5 Трансформований XHTML файл

Висновки: при виконанні данії лабораторної роботи я дізнався про XPath, XSLT та XHTML, здобув навички створення програм для обробки XML документів.