



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

Длубак Родіон  
Романович

Варіант 6

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 6

Перший сайт	Завдання 2	Інтернет-магазин
<a href="http://www.isport.ua">www.isport.ua</a>	Вивести список гіперпосилань	<a href="http://www.portativ.ua">www.portativ.ua</a>

#### Результати:

Код для “павука”, для сайту isport.ua:

```
from urllib.parse import urljoin

import scrapy

def isEmptyString(str):
    return len(str) > 0

class GolosUaSpider(scrapy.Spider):
    name = "isport"
    custom_settings = {
        'ITEM_PIPELINES': {
            'lab1.pipelines.NewsXmlPipeline': 300,
        }
    }
    fields = {
        'img': '//img/@src',
        'text': '//*[not(self::script)]/text()',
        'link': '//a/@href'
    }
    start_urls = [
        'https://isport.ua/'
    ]
    allowed_domains = [
        'isport.ua'
    ]
    max_pages = 20

    def __init__(self, **kwargs):
        super().__init__(**kwargs)
        self.visited_pages = ['https://isport.ua/']

    def parse(self, response):
        links = response.xpath('//body//a/@href').extract()
        self.visited_pages.append('mailto:info@isport.ua')
        for l in links:
            if l not in self.visited_pages:
```

```

        print(1)
        self.visited_pages.append(1)
        url = urljoin(response.url, 1)
        yield scrapy.Request(url, callback=self.parse_page)

def parse_page(self, response):
    text = filter(isNotEmptyString,
                  map(lambda str: str.strip(),
                      [text.extract() for text in response.xpath(self.fields["text"])]))
    images = map(lambda url: ((response.url + url) if url.startswith('/') else url),
                  [img_url.extract() for img_url in response.xpath(self.fields["img"])]))
    return {
        'text': text,
        'images': images,
        'url': response.url
    }

```

## Код для паука для магазину [portativ.ua](http://portativ.ua):

```

import scrapy
from scrapy import Selector

class PetMarketSpider(scrapy.Spider):
    name = "portativ"
    fields = {
        'price': '//span[@class="price-value UAH"]/text()',
        'name': '//div[@class="cataloggrid-item-name-block"]/a/text()',
        'product': '//div[@class="cataloglist-item-container"]',
        'img': '//a[@class="product-image"]/img/@src',
        'product_link': '//a[@class="product-image"]/@href'
    }
    start_urls = [
        'https://portativ.ua/category_841832.html?tip_podkljuchenija_fe9f=173597'
    ]
    allowed_domains = [
        'portativ.ua'
    ]
    number_of_items = 20

    def parse(self, response):
        for product in response.xpath(self.fields["product"]).getall()[0:self.number_of_items]:
            selector = Selector(text=product)
            yield {
                'link': selector.xpath(self.fields['product_link']).extract(),
                'price': selector.xpath(self.fields['price']).get().strip(),
                'img': selector.xpath(self.fields['img']).extract(),
                'name': ''.join(selector.xpath(self.fields['name']).extract())
            }

```

## XSLT для трансформації:

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <xsl:output method="xml"
    doctype-system="http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd"
    doctype-public="-//W3C//DTD XHTML 1.1//EN" indent="yes"/>

  <xsl:template match="/">
    <html>
      <head>
        <meta charset="utf-8"/>
        <meta name="viewport" content="width=device-width, initial-scale=1"/>
        <link rel="stylesheet"
href="https://cdnjs.cloudflare.com/ajax/libs/bulma/0.7.4/css/bulma.min.css"/>
        <title>
          Portativ electronics
        </title>
      </head>
      <body>
        <div class="container">
          <div class="notification">
            <h1 class="title">
              The Portativ electronics
            </h1>
          </div>

          <table class="table is-bordered is-striped is-narrow is-hoverable is-fullwidth">
            <thead>
              <tr>
                <th>Name</th>
                <th>Price</th>
                <th>Img</th>
              </tr>
            </thead>
            <tbody>
              <xsl:apply-templates/>
            </tbody>
          </table>
        </div>
      </body>
    </html>
  </xsl:template>

  <xsl:template match="item">
    <tr>
      <td style="width:30%">
        <a>
          <xsl:attribute name="href">
            <xsl:value-of select="link"/>
          </xsl:attribute>
          <xsl:value-of select="name"/>
        </a>
        <br/>
      </td>
      <td>
        <xsl:value-of select="price"/>
      </td>
      <td>
        <xsl:element name="img">
          <xsl:attribute name="src">
            <xsl:value-of select="img"/>
          </xsl:attribute>
        </xsl:element>
      </td>
    </tr>
  </xsl:template>
</xsl:stylesheet>
```

```

2021-02-24 22:17:39 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://isport.ua/football/europe/3011339-ankhelino-gvardiola-unichtozhil-moyu-uverennost-v-sebe> (referer: https://isport.ua/)
2021-02-24 22:17:39 [scrapy.core.scraper] DEBUG: Scraped from <200 https://isport.ua/>
{'text': <filter object at 0x000002CB5FF7B580>, 'images': <map object at 0x000002CB60044E0>, 'url': 'https://isport.ua/'>
2021-02-24 22:17:39 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://isport.ua/football/ukraine/3011315-petresku-mozhet-vozglavit-dnepr-1> (referer: https://isport.ua/)
2021-02-24 22:17:39 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://isport.ua/football/europe/3011363-buffona-priznali-luchshin-golkiperon-sovremennosti-po-versii-iffhs> (referer: https://isport.ua/)
2021-02-24 22:17:39 [scrapy.core.scraper] DEBUG: Scraped from <200 https://isport.ua/football/europe/3011339-ankhelino-gvardiola-unichtozhil-moyu-uverennost-v-sebe>
{'text': <filter object at 0x000002CB5FF7B580>, 'images': <map object at 0x000002CB60044E0>, 'url': 'https://isport.ua/football/europe/3011339-ankhelino-gvardiola-unichtozhil-moyu-uverennost-v-sebe'>
2021-02-24 22:17:39 [scrapy.core.scraper] DEBUG: Scraped from <200 https://isport.ua/football/ukraine/3011315-petresku-mozhet-vozglavit-dnepr-1>
{'text': <filter object at 0x000002CB5FF7B580>, 'images': <map object at 0x000002CB60044E0>, 'url': 'https://isport.ua/football/ukraine/3011315-petresku-mozhet-vozglavit-dnepr-1'>
2021-02-24 22:17:39 [scrapy.core.scraper] DEBUG: Scraped from <200 https://isport.ua/football/europe/3011363-buffona-priznali-luchshin-golkiperon-sovremennosti-po-versii-iffhs>
{'text': <filter object at 0x000002CB5FF7B580>, 'images': <map object at 0x000002CB60044E0>, 'url': 'https://isport.ua/football/europe/3011363-buffona-priznali-luchshin-golkiperon-sovremennosti-po-versii-iffhs'>
2021-02-24 22:17:39 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://isport.ua/football/uefa/3011337-matvienko-net-nikakikh-garantij-chno-shakhter-projdet-makkabi> (referer: https://isport.ua/)
2021-02-24 22:17:39 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://isport.ua/power/3011307-mejvezer-nachal-vstrechatysya-so-striptizershej> (referer: https://isport.ua/)
2021-02-24 22:17:39 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://isport.ua/football/uefa/3011347-ally-zabil-shikarnyj-gol-udaron-cherез-sbva-v-vorota-volfsbergera> (referer: https://isport.ua/)
2021-02-24 22:17:39 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://isport.ua/boxing/3011378-stalo-izvestno-uslovie-pri-kotorom-sostoitsya-tretij-boj-alvares-golovkin> (referer: https://isport.ua/)
2021-02-24 22:17:39 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://isport.ua/football/europe/3011372-eks-igroka-nyukasla-obvinili-v-seksualnom-nasilii-nad-odnoklubnikami> (referer: https://isport.ua/)

```

Рис. 1 Парсинг сайту isport.ua

```

<data>
  <page url="https://isport.ua/">
    <fragment type="text">iSport.ua - новости спорта: футбол, бокс, баскетбол, хоккей, теннис</fragment>
    <fragment type="text">Футбол</fragment>
    <fragment type="text">Бокс</fragment>
    <fragment type="text">Биатлон</fragment>
    <fragment type="text">Теннис</fragment>
    <fragment type="text">ММА</fragment>
    <fragment type="text">ВИДЕО</fragment>
    <fragment type="text">Другие</fragment>
    <fragment type="text">Автоспорт</fragment>
    <fragment type="text">Баскетбол</fragment>
    <fragment type="text">Хоккей</fragment>
    <fragment type="text">Киберспорт</fragment>
    <fragment type="text">Новости</fragment>
    <fragment type="text">Лига Чемпионов</fragment>
    <fragment type="text">22:00</fragment>
    <fragment type="text">Аталанта - Реал 0:0 онлайн-трансляция матча Лиги чемпионов</fragment>
    <fragment type="text">Европа</fragment>
    <fragment type="text">21:58</fragment>
    <fragment type="text">Дубль Месси помог Барселоне разгромить Эльче</fragment>
    <fragment type="text">Лига Европы</fragment>
    <fragment type="text">21:51</fragment>
    <fragment type="text">Тренер Маккаби: Сделаем все возможное в матче против Шахтера</fragment>
    <fragment type="text">Лига Чемпионов</fragment>
    <fragment type="text">21:30</fragment>
    <fragment type="text">Зинченко не попал в стартовый состав Ман Сити на матч против Боруссии М</fragment>

```

Рис. 2 Результуючий XML файл

```
/football/369140-ukraine
/football/369141-europe
/football/369142-other-countries
/football/369143-champions-league
/football/369144-uefa
/659685-worldcup
#Submenu544954d6b2e2090cc9321161397af217
/other/369166-gaming
/other/1097246-bitva-redaktsij
/other/369167-athletics
/other/369168-gymnastics
/other/369173-chess
/other/369174-aquatics
/other/369175-olympics
/other/369176-cycling
#Submenu29038d28032e6aa2b6f1855368805a1a
/auto/369157-f1
/auto/369159-moto
/auto/369158-rally
/auto/369160-else
#Submenufb70ec174e663471840096e0d304a24b
/basketball/369147-ukraine
/basketball/369148-nba
/basketball/369149-eurocup
/basketball/369150-world
#Submenu2e8761e0d49138d90e79e266efbd1dae
/hockey/369153-ukraine
/hockey/369154-nhl
/hockey/369155-europe
/1172270-pravila-polzovaniya-interaktivnymi-resursami-sajta-isportua
```

Рис. 3 за допомогою XPath виводимо список посилань



```
<?xml version="1.0" encoding="utf-8"?>
<items>
<item><link><value>https://portativ.ua/product_27368.html</value></link><price>490</price><img><value>https://portativ.ua/media/cache
Наушники QCY T1c Black
</name></item>
<item><link><value>https://portativ.ua/product_21782.html</value></link><price>699</price><img><value>https://portativ.ua/media/cache
Наушники JAM Live Large Black (HX-EP303BK)
</name></item>
<item><link><value>https://portativ.ua/product_23252.html</value></link><price>599</price><img><value>https://portativ.ua/media/cache
Наушники Panasonic RP-NJ310BGE-A
</name></item>
<item><link><value>https://portativ.ua/product_12779.html</value></link><price>599</price><img><value>https://portativ.ua/media/cache
Наушники JAM Transit Mini BT Gray
</name></item>
<item><link><value>https://portativ.ua/product_23958.html</value></link><price>599</price><img><value>https://portativ.ua/media/cache
Наушники Xiaomi Redmi AirDots Black (ZBW4467CN)
</name></item>
<item><link><value>https://portativ.ua/product_28119.html</value></link><price>699</price><img><value>https://portativ.ua/media/cache
Наушники Xiaomi Redmi AirDots 2 Black (TWSEJ061LS/BHR42726L)
</name></item>
<item><link><value>https://portativ.ua/product_28591.html</value></link><price>599</price><img><value>https://portativ.ua/media/cache
Наушники Realme Buds Q Black (RMA215)
</name></item>
<item><link><value>https://portativ.ua/product_21783.html</value></link><price>799</price><img><value>https://portativ.ua/media/cache
Наушники JAM Live Fast Black (HX-EP404BK)
</name></item>
<item><link><value>https://portativ.ua/product_25905.html</value></link><price>649</price><img><value>https://portativ.ua/media/cache
Наушники JBL T115BT Coral (JBLT115BTICOR)
</name></item>
<item><link><value>https://portativ.ua/product_17593.html</value></link><price>999</price><img><value>https://portativ.ua/media/cache
Наушники Soul Impact Wireless Bluetooth Earphones White
</name></item>
<item><link><value>https://portativ.ua/product_27369.html</value></link><price>699</price><img><value>https://portativ.ua/media/cache
Наушники QCY T9s Black
</name></item>
<item><link><value>https://portativ.ua/product_22337.html</value></link><price>899</price><img><value>https://portativ.ua/media/cache
Наушники Soul Pure Wireless Plus White
</name></item>
```

Рис. 4 XML файл з даними навушники сайту portativ.ua





Name	Price	Img
Наушники QCY T1c Black	490	
Наушники JAM Live Large Black (HX-EP303BK)	699	
Наушники Panasonic RP-NJ310BGE-A	599	
Наушники JAM Transit Mini BT Gray	599	

Рис. 5 Трансформований XHTML файл

**Висновки:** при виконанні даної лабораторної роботи я дізнався як легко діставати потрібну інформацію з сайтів за допомогою XPath, XSLT та XHTML, здобув навички створення програм для обробки XML документів.