



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”

Виконав  
студент III курсу  
групи КП-82

Мельничук Олексій Геннадійович  
*(прізвище, ім'я, по батькові)*

варіант № 12

Зарахована  
“ \_\_\_\_ ” “ \_\_\_\_ ” 20\_\_ р.  
викладачем

Петрашенко Андрій Васильович  
*(прізвище, ім'я, по батькові)*

## Мета роботи

Здобуття практичних навичок створення програм, орієнтованих на обробку XML-документів засобами мови Python.

## Завдання

1. На основі базової адреси [www.uartlib.org](http://www.uartlib.org) 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="www.server.com/index.html">
    <fragment type="text">
... знайдений текст
    </fragment>
    <fragment type="image">
... url зображення
    </fragment>
  </page>
  <page url="www.server.com/index1.html">
    <fragment type="text">
... знайдений текст
    </fragment>
    <fragment type="image">
... url зображення
    </fragment>
  </page>
...
</data>
```

2. Вивести середню кількість текстових фрагментів на сторінку на ресурсі.
3. Проаналізувати вміст [www.hozmart.com.ua](http://www.hozmart.com.ua). 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.

## Код програми

### Main.py

```
import os
from scrapy import cmdline
from lxml import etree

mainpath = os.path.dirname(__file__) + '/tutorial'
root = None

# cmdline.execute("scrapy crawl uartlib".split())

# result = open(mainpath + '/results/o_uartlib.xml', 'rb')
# root = etree.parse(result)
# result.close()

# pageCount = root.xpath('count(//page)')
# textCount = root.xpath('count(//fragment[@type="text"])')
# result = 'Average count of text fragments per page: %f' % (textCount / pageCount)

# f = open(mainpath + '/results/l_artlib.txt', 'w')
# f.write(result)
# f.close()

def task_1_parse():
    cmdline.execute("scrapy crawl uartlib".split())

def task_1_processing():
    result = open(mainpath + '/results/o_uartlib.xml', 'rb')
    root = etree.parse(result)
    result.close()

    pageCount = root.xpath('count(//page)')
    textCount = root.xpath('count(//fragment[@type="text"])')
    result = 'Average count of text fragments per page: %f' % (textCount / pageCount)

    f = open(mainpath + '/results/r_uartlib.txt', 'w')
    f.write(result)
    f.close()

def task_2_parse():
    cmdline.execute("scrapy crawl hozmart".split())

def task_2_processing():
    dom = etree.parse(mainpath + '/results/o_hozmart.xml')
    xslt = etree.parse(mainpath + '/hozmart.xslt')
    transform = etree.XSLT(xslt)
    newdom = transform(dom)
    with open(mainpath + '/results/r_hozmart.html', 'wb') as f:
        f.write(etree.tostring(newdom, pretty_print=True))

task_2_processing()
```

## Pipelines.py

```
from lxml import etree
import os

class TutorialPipeline(object):
    def open_spider(self, spider):
        self.root = etree.Element("data")

    def close_spider(self, spider):
        f = open(os.path.dirname(__file__) + '/results/o_' + spider.name + '.xml',
'wb')

        f.write(etree.tostring(
            self.root, encoding="UTF-8",
            pretty_print=True
        ))
        f.close()

    def process_item(self, item, spider):
        if spider.name == "uartlib":
            page = etree.SubElement(self.root, "page", url=item["url"])
            for text in item['texts']:
                etree.SubElement(page, 'fragment', type='text').text = text
            for url in item['images']:
                etree.SubElement(page, 'fragment', type='image').text = url
            self.root.append(page)
        else:
            product: etree.Element = etree.Element("product")

            name = etree.Element("name")
            name.text = item["names"]

            price = etree.Element("price")
            price.text = item["prices"]

            image = etree.Element("image")
            image.text = item["images"]

            avail = etree.Element("availability")
            avail.text = item["avails"]

            product.append(name)
            product.append(price)
            product.append(image)
            product.append(avail)

            self.root.append(product)
        return item
```

## Uartlib\_spyder.py

```
import scrapy

class uartlibSpider(scrapy.Spider):
    name = "uartlib"
    start_urls = [
        'http://uartlib.org'
    ]
    selectors = {
        'text': "//*[not(self::script)][not(self::style)]//text()[normalize-space()][not(contains(.,'{')][not(contains(.,';')))]",
        'img': '//img/@src',
        'url': "//a/@href[starts-with(., '" + start_urls[0] + "')or starts-with(., '/'')]"
    }
    def parse(self, response):
        texts = response.xpath(self.selectors['text']).extract()
        images = response.xpath(self.selectors['img']).extract()
        urls = response.xpath(self.selectors['url']).extract()
        yield {
            'url': response.url,
            'texts': texts,
            'images': images
        }
        if response.url == self.start_urls[0]:
            links = [
                link for link in urls if link != "/"
            ]
            for link in links[:19]:
                if link.startswith("/"):
                    link = self.start_urls[0] + link
                yield response.follow(link, callback=self.parse)
```

## hozmart\_spyder.py

```
import scrapy

class hozmartSpider(scrapy.Spider):
    name = "hozmart"
    start_urls = [
        'https://hozmart.com.ua/uk/15-benzopili',
        'https://hozmart.com.ua/uk/15-benzopili?p=2'
    ]
    selectors = {
        'all_items': "//ul[contains(@id, 'product_list')]/li",
        'names': ".//a[contains(@class, 'b1c-name-uk')]/text()",
        'images': ".//img[contains(@class, 'b1c-img')]/@src",
        'prices': ".//span[contains(@class, 'price')]/text()",
        'available': ".//p[contains(@class, 'availability')]/span//text()[normalize-space()]"
    }
    def parse(self, response):
        for result in response.xpath(self.selectors['all_items'][:20]):
            yield {
                'names': result.xpath(self.selectors['names']).extract_first(),
                'images': result.xpath(self.selectors['images']).extract_first(),
                'prices': result.xpath(self.selectors['prices']).extract_first(),
                'avails':
                    result.xpath(self.selectors['available']).extract_first().strip()
            }
```

## Результати роботи програм

```
pipelines.py  uartlib_spider.py  settings.py  bruh.py  hozmart_spider.py

tutorial > results > o_uartlib.xml

1  <data>
2    <page url="http://uartlib.org">
3      <fragment type="text">Бібліотека українського мистецтва - Головна</fragment>
4      <fragment type="text">EN </fragment>
5      <fragment type="text">UA </fragment>
6      <fragment type="text">ПІДТРИМАТИ БІБЛІОТЕКУ</fragment>
7      <fragment type="text">0 Елементи</fragment>
8      <fragment type="text">УКРАЇНСЬКІ ХУДОЖНИКИ</fragment>
9      <fragment type="text">КНИГИ</fragment>
10     <fragment type="text">МИСТЕЦТВОЗНАВЧИ</fragment>
11     <fragment type="text">ІСТОРІЯ УКРАЇНСЬКОГО МИСТЕЦТВА</fragment>
12     <fragment type="text">УКРАЇНСЬКИЙ АВАНГАРД</fragment>
13     <fragment type="text">УЖИТКОВЕ МИСТЕЦТВО</fragment>
14     <fragment type="text">ГУЦУЛЬСЬКА КЕРАМІКА</fragment>
15     <fragment type="text">ПЕТРИКІВСЬКИЙ РОЗПИС</fragment>
16     <fragment type="text">ДИТЯЧІ КНИГИ</fragment>
17     <fragment type="text">ПОЗА МИСТЕЦТВОМ</fragment>
18     <fragment type="text">КНИГИ ПРО КИЇВ</fragment>
19     <fragment type="text">РІК ВИДАННЯ</fragment>
20     <fragment type="text">ЖУРНАЛИ</fragment>
21     <fragment type="text">НОВА ГЕНЕРАЦІЯ</fragment>
22     <fragment type="text">НОТАТКИ З МИСТЕЦТВА</fragment>
23     <fragment type="text">ХУДОЖНИКИ УКРАЇНИ</fragment>
24     <fragment type="text">ОБРАЗОТВОРЧЕ МИСТЕЦТВО</fragment>
25     <fragment type="text">МИСТЕЦТВОЗНАВЦІ</fragment>
26     <fragment type="text">ФОТОАРХІВ</fragment>
27     <fragment type="text">СТАТТІ</fragment>
28     <fragment type="text">ЕКСКЛЮЗИВ</fragment>
29     <fragment type="text">ПРО БІБЛІОТЕКУ</fragment>
```

```

521 <fragment type="image">http://uartlib.org/wp-content/uploads/2016/11/united_kingdom_great_britain.png</f
522 <fragment type="image">http://uartlib.org/wp-content/uploads/2016/11/ukraine.png</fragment>
523 <fragment type="image">http://uartlib.org/wp-content/uploads/2016/05/uartlibtop-1.png</fragment>
524 <fragment type="image">http://uartlib.org/wp-content/uploads/2021/02/surprisemalevich-150x200.jpg</fragme
525 <fragment type="image">http://uartlib.org/wp-content/uploads/2021/02/slobotco-150x200.jpg</fragment>
526 <fragment type="image">http://uartlib.org/wp-content/uploads/2021/02/dovgunco-150x200.jpg</fragment>
527 <fragment type="image">http://uartlib.org/wp-content/uploads/2021/02/bvco906-150x200.jpg</fragment>
528 <fragment type="image">http://uartlib.org/wp-content/uploads/2021/02/coverbalaban-150x200.jpg</fragment>
529 <fragment type="image">http://uartlib.org/wp-content/uploads/2021/01/img222co-150x200.jpg</fragment>
530 <fragment type="image">http://uartlib.org/wp-content/uploads/2021/01/img220co-150x200.jpg</fragment>
531 <fragment type="image">http://uartlib.org/wp-content/uploads/2021/01/MDS18672co-150x200.jpg</fragment>
532 <fragment type="image">http://uartlib.org/wp-content/uploads/2021/01/img182co-150x200.jpg</fragment>
533 <fragment type="image">http://uartlib.org/wp-content/uploads/2020/12/karpotroco-150x200.jpg</fragment>
534 <fragment type="image">http://uartlib.org/wp-content/uploads/2020/12/MDS18233co-150x200.jpg</fragment>
535 <fragment type="image">http://uartlib.org/wp-content/uploads/2021/01/img211-150x200.jpg</fragment>
536 <fragment type="image">http://uartlib.org/wp-content/uploads/2020/12/img035co-150x200.jpg</fragment>
537 <fragment type="image">http://uartlib.org/wp-content/uploads/2020/12/bajayco-150x200.jpg</fragment>
538 <fragment type="image">http://uartlib.org/wp-content/uploads/2020/12/kaufmanvco-150x200.jpg</fragment>
539 <fragment type="image">http://uartlib.org/wp-content/uploads/2020/12/budnlvivco-150x200.jpg</fragment>
540 <fragment type="image">http://uartlib.org/wp-content/uploads/2020/12/ralkolvivco-150x200.jpg</fragment>
541 <fragment type="image">http://uartlib.org/wp-content/uploads/2020/12/img810co-150x200.jpg</fragment>
542 <fragment type="image">http://uartlib.org/wp-content/uploads/2020/12/coimg809cfo-150x200.jpg</fragment>
543 <fragment type="image">http://uartlib.org/wp-content/uploads/2020/12/img723co-150x200.jpg</fragment>
544 <fragment type="image">http://uartlib.org/wp-content/uploads/2021/01/img212co-150x200.jpg</fragment>
545 <fragment type="image">http://uartlib.org/wp-content/uploads/2020/11/img465co-150x200.jpg</fragment>
546 <fragment type="image">http://uartlib.org/wp-content/uploads/2021/01/img223co-150x200.jpg</fragment>
547 <fragment type="image">http://uartlib.org/wp-content/uploads/2020/11/img248cob-150x200.jpg</fragment>

```

Рис.1.1-2. Скріншоти результатів завдання 1

```

pipelines.py  uartlib_spider.py  settings.py  bruh.py  hozmart_spider.py  r_uartlib.txt X
tutorial > results > r_uartlib.txt
1 Average count of text fragments per page: 290.222222

```

Рис.2. Скріншот результатів завдання 2

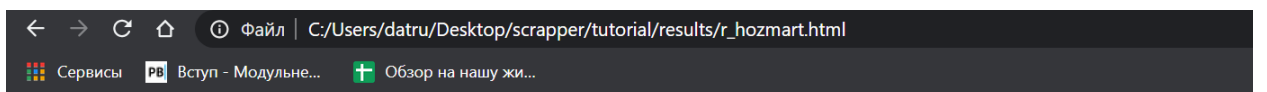


```

pipelines.py  uartlib_spider.py  settings.py  bruh.py  hozmart_spider.py  o_hozmart.xml  hozmart
tutorial > results > o_hozmart.xml
1  <data>
2  <product>
3  <name>Бензопила Gator GS-52, 52cc, Гатор (GS-52)</name>
4  <price>1 977 грн.</price>
5  <image>https://www.hozmart.com.ua/13719-home_default/benzopila-gator-gs-52-52cc.jpg</image>
6  <availability>Відсутній</availability>
7  </product>
8  <product>
9  <name>Бензопила Sadko GCS-254, Садко (8013172)</name>
10 <price>2 309 грн.</price>
11 <image>https://www.hozmart.com.ua/655-home_default/benzopyla-sadko-gcs-254.jpg</image>
12 <availability>Відсутній</availability>
13 </product>
14 <product>
15 <name>Бензопила Saber SC-52, Сабер (SC-52)</name>
16 <price>2 404 грн.</price>
17 <image>https://www.hozmart.com.ua/13717-home_default/benzopyla-saber-sc-52.jpg</image>
18 <availability>Відсутній</availability>
19 </product>
20 <product>
21 <name>Бензопила Sadko GCS-510E, Садко (8009303)</name>
22 <price>2 411 грн.</price>
23 <image>https://www.hozmart.com.ua/499-home_default/benzopyla-sadko-gcs-510e.jpg</image>
24 <availability>Відсутній</availability>
25 </product>
26 <product>
27 <name>Бензопила Sadko GCS-560E, Садко (8009894)</name>
28 <price>2 569 грн.</price>
29 <image>https://www.hozmart.com.ua/611-home_default/benzopyla-sadko-gcs-560e.jpg</image>
30 <availability>Відсутній</availability>

```

Рис.3. Скріншот результатів завдання 3



## Saws


Name	Price	Availability
Бензопила Gator GS-52, 52cc, Гатор (GS-52)		
	1 977 грн.	Відсутній
Бензопила Sadko GCS-254, Садко (8013172)		
	2 309 грн.	Відсутній
Бензопила Saber SC-52, Сабер (SC-52)		
	2 404 грн.	Відсутній

Рис.4.1. Скріншот результатів завдання 4

## **Висновки**

Виконавши дану лабораторну роботу я ознайомився з синтаксисом та принципом роботи XPath та XSLT, використав здобуті знання для проектування та розробки програми для діставання та обробки великих об'ємів інформації з веб-сторінок.