**Міністерство освіти і науки України**

**Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського"**

**Факультет інформатики та обчислювальної техніки**

**Кафедра інформатики та програмної інженерії**

**Звіт**

з лабораторної роботи №5 з дисципліни

«Програмування інтелектуальних інформаційних систем»

**Виконав(ла)**

(шифр, прізвище, ім'я, по батькові)

*ІП-11 Головня Олександр Ростиславович*

**Перевірив**

(прізвище, ім'я, по батькові)

*Баришич Л. М.*

Київ 2023

Лаба 5:

**1 Зіскрапити заголовки новин з сайту (https://pestrecy**

**rt.ru/news/tag/list/specoperaciia/) https://archive.is/pestrecy-rt.ru за допомогою  xpath (можливо знадобиться впн)**

**2 Реалізувати кроулінг через натискання на кнопку “Далі” 3 По отриманих заголовках створити список померлих росіян**

**Скрапити і кроулити можна як створивши застосунок, так і  використовуючи плагіни чи сторонні ресурсиэ**

Виконання:  
import requests

from bs4 import BeautifulSoup

import re

from urllib.parse import urlparse, urlunparse

url = "https://pestrecy-rt.ru/news/tag/list/specoperaciia/"

response = requests.get(url)

soup = BeautifulSoup(response.text, 'html.parser')

headlines = []

name\_regex = re.compile(r'([А-ЯЁІЇЄҐ][а-яёіїєґ]+) ([А-ЯЁІЇЄҐ][а-яёіїєґ]+)')

patterns\_to\_remove = {'СВО', 'PT', 'РФ', 'ВСЕ', 'Президент', 'Татарстан', 'Глава', 'Сил', 'России',

                      'Геро', 'Родин',  'Казан', 'Житель', 'Совета', 'Указ', 'Госдум', 'Заместитель',

                      'Защитники', 'Отделением', 'Республик', 'Госсовет',  'Пестре', 'Пенсионерка',

                      'Минобороны', 'Жизнь', 'Священник', 'Шигалеево', 'Мама', 'Фонда', 'Главнокомандующему',

                      'Также', 'Казни', 'Кремле', 'Верховный', 'Донбасса', 'Семье', 'Казанском', 'Около',

                      'Волге', 'Орденом', 'Мужества', 'Мост'}

for page\_number in range(2, 15):

    page\_headlines = soup.find\_all('h2', class\_='oneNews\_\_link news\_\_bold-text')

    page\_paragraphs = soup.find\_all('p', class\_='oneNews\_\_link')

    headlines.extend(f'{headline.text.strip()}. {paragraph.text.strip()}' for headline, paragraph in zip(page\_headlines, page\_paragraphs))

    base\_url = "https://pestrecy-rt.ru/news/tag/list/specoperaciia/page/"

    next\_url = f"{base\_url}{page\_number}"

    print(next\_url)

    response = requests.get(next\_url)

    soup = BeautifulSoup(response.text, 'html.parser')

dead\_russians = set()

for headline in headlines:

    headline = ' '.join(word for word in headline.split() if not any(pattern in word for pattern in patterns\_to\_remove))

    matches = name\_regex.findall(headline)

    for match in matches:

        first\_name = match[0]

        last\_name = match[1]

        dead\_russians.add(f'{first\_name} {last\_name}')

count=0;

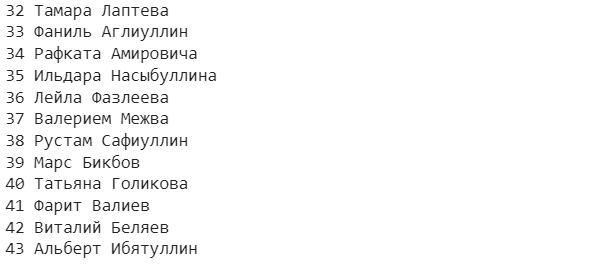
for item in dead\_russians:

    count +=1

    print(str(count)+" " + item)

Результати





Висновок:

Під час виконання лабораторної роботи було зіскраплено сайт для пошуку заголовків та  параграфів. Також виконано кроулінг для пересування по сторінкам сайту. Зібравши всі

заголовки та параграфи ми створили список імен мертвих росіян.