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»

„**Скрапінґ та кроулінґ веб сторінок**”

Виконав(ла) *ІП-11 Прищепа В.С.* _____
(шифр, прізвище, ім'я, по батькові)

Перевірив *Баришич Л. М.* _____
(прізвище, ім'я, по батькові)

Київ 2023

Завдання

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

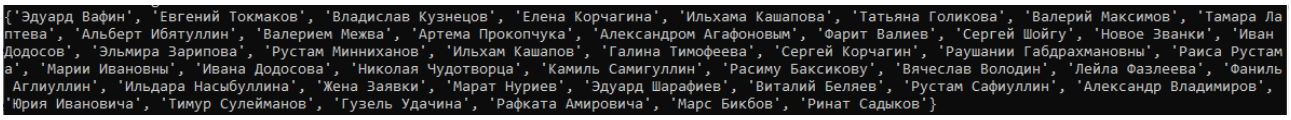
Хід роботи:

Код програми:

```
from selenium import webdriver
from selenium.webdriver.common.by import By
import re

browser = webdriver.Chrome()
browser.get("https://pestrecy-rt.ru/news/tag/list/specoperaciia/")
headlines = []

while True:
    heads = browser.find_elements(By.XPATH, '//h2[@class="oneNews__link news__bold-text"]')
    parags = browser.find_elements(By.XPATH, '//p[@class="oneNews__link"]')
    for hdline, parag in zip(heads, parags):
        headlines.append(f'{hdline.text.strip()}. {parag.text.strip()}')
    try:
        next_button = browser.find_element(By.XPATH, '//a[@class="all-news__button_forward"]//span[contains(text(), "Далее")]')
        next_button.click()
    except Exception:
        break
browser.quit()

namreg = re.compile(r'([А-ЯЁІЇЄҐ][а-яёіїєґ]+) ([А-ЯЁІЇЄҐ][а-яёіїєґ]+)')
corpses = set()
oddwords = {'СВО', 'РТ', '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 headline in headlines:
    headline = ' '.join(word for word in headline.split() if not any(pattern in word for
pattern in oddwords))
    matches = namreg.findall(headline)
    for match in matches:
        name = match[0]
        surname = match[1]
        corpses.add(f'{name} {surname}')
print(corpses)
```

Результат виконання:



Висновок: під час виконання лабораторної роботи я виконав скрапінґ заголовків за допомогою Xpath та кроулінґу сторінок та вичленив імена загиблих росіян з заголовків. Код і результат виконання наведені в лабораторній роботі.