Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Факультет інформатики та обчислювальної техніки

Кафедра інформаційних систем та технологій

Лабораторна робота №5

з дисципліни «Програмування інтелектуальних інформаційних систем»

Виконав:

студент групи ІП-11

Лисенко Андрій

Київ, 2023

Завдання

1. Зіскрапити заголовки новин з сайту https://pestrecy-rt.ru/news/tag/list/specoperaciia/ 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 ресурси

Код

```
import pymorphy3
import spacy
from selenium import webdriver
from selenium.common import NoSuchElementException
from selenium.webdriver.common.by import By

TARGET_URL = 'https://pestrecy-rt.ru/news/tag/list/specoperaciia'
NEWS_HEADER_XPATH = '/html/body/main/ul/li[{}]/a/div[1]/h2'
NEXT_BUTTON_FIRST_PAGE_XPATH = '/html/body/main/div[2]/div/a'
NEXT_BUTTON_XPATH = '/html/body/main/div[2]/div/a[2]'

driver = webdriver.Chrome()
driver.get(TARGET_URL)

news_headers = []
page = 1
while True:
    i = 1
    while True:
        try:
            news_header = driver.find_element(By.XPATH,
NEWS_HEADER_XPATH.format(i))
        except NoSuchElementException:
            break
        else:
            news_headers.append(news_header.text)
```

```
            i += 1
    try:
        next_button = driver.find_element(By.XPATH,
NEXT_BUTTON_FIRST_PAGE_XPATH if page == 1 else NEXT_BUTTON_XPATH)
    except NoSuchElementException:
        break
    else:
        next_button.click()
        page += 1

nlp = spacy.load('ru_core_news_md')
morph = pymorphy3.MorphAnalyzer()

names = []
for news_header in news_headers:
    doc = nlp(news_header)
    header_names = [entity.lemma_ for entity in doc.ents if entity.label_
== "PER"]
    names.extend(header_names)


for name in ['минмолодежи рт', 'соцфонд']:
    names.remove(name)

names = list(set(names))

for name in sorted(names):
    print(name)
```

Результати

```
александр агафонов
валерий межва
виталий беляев
елена корчагин
жуков
иван додосов
иван додосова
куюков
лейла фазлеева
минниханов
```

```
пестрецов
путин
расим баксикову
тамара лаптев
тинчурина
эдуард шарафиев
```