Міністерство освіти і науки України
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Фізико-технічний інститут

Лабораторна робота №4
з дисципліни
«Web - аналітика»

Виконав:
студент групи ФБ-31мп
Щур Павло
Перевірив:
Ткач В. М.

Київ-2024

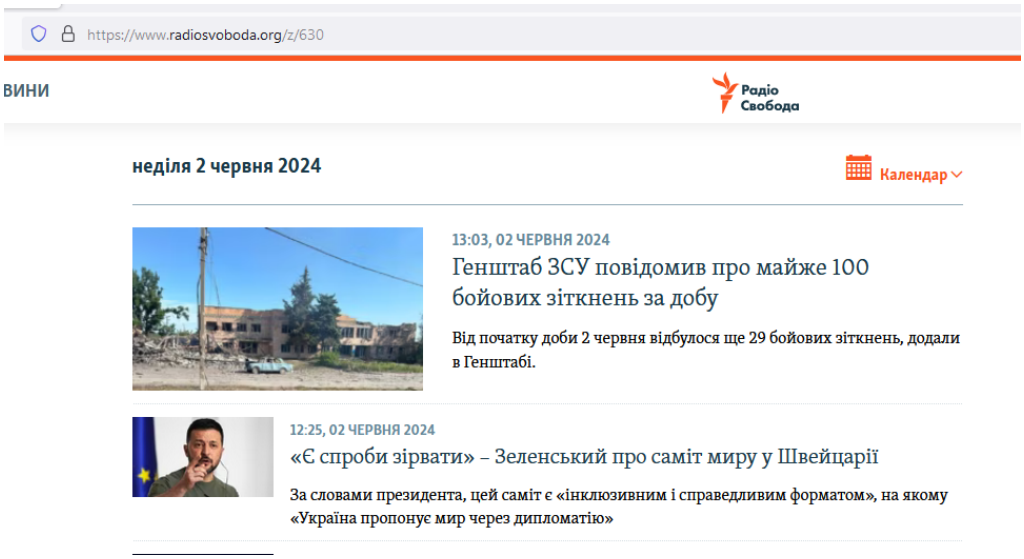
Посилання на GitHub: https://github.com/ShchurPavlo/web-analytics-2024/tree/main/lab4

Завдання

1. Збір великих даних із новинних ресурсів. Формування датасетів. Можна один ресурс не менше 1000 записів.

Виконання:

1) В якості піддослідного новинного ресурсу будемо використовувати сайт Радіо Свобода - https://www.radiosvoboda.org/z/630



Для збору даних використаємо бібліотеку для Python BeautifulSoup, а для відкриття веб-сторінки (обходу перевірки на бота) використаємо Selenium та його Chrome Driver.

2) Відкривши сторінку в режимі розробника знайдемо клас *col-xs-12 col-sm-12 col-md-12 col-lg-12 fui-grid__inner* в якому міститься інформація про новини:

```
</ul>
<ul id="ordinaryItems">
<li class="col-xs-12 col-sm-12 col-md-12 col-lg-12 fui-grid__inner">
<div class="media-block ">
<a href="/a/news-zelenskyj-samit-myru-sproby-zirvaty/32975840.html" class="img-wrap img-wrap--t-spac img-wrap--size-3 img-wrap--float img-wrap--xs"
<div class="thumb thumb16_9">
<noscript class="nojs-img">
<img src="https://gdb.rferl.org/01000000-0aff-0242-6331-08dc7f622faf_w100_r1.jpg" alt="Президент України Володимир Зеленський" />
</noscript>
<img data-src="https://gdb.rferl.org/01000000-0aff-0242-6331-08dc7f622faf_w66_r1.jpg" src="" alt="Президент України Володимир Зеленський" class=""/>
</div>
</a>
<div class="media-block__content media-block__content--h media-block__content--h-xs">
<span class="date date--mb date--size-3" >12:25, 02 червня 2024</span>
<a href="/a/news-zelenskyj-samit-myru-sproby-zirvaty/32975840.html">
<h4 class="media-block__title media-block__title--size-3" title="&#171;Є спроби зірвати&#187; – Зеленський про саміт миру у Швейцарії">
&#171;Є спроби зірвати&#187; – Зеленський про саміт миру у Швейцарії
</h4>
<p class="perex perex--mb perex--size-3">За словами президента, цей саміт є &#171;інклюзивним і справедливим форматом&#187;, на якому &#171;Україна
</a>
```

3) Реалізуємо окрему функцію для отримання html розмітки сторінки з новинами:

```python
def Get_soup(url,driver):
    driver.get(url.replace(' ', ''))
    time.sleep(5)
    source_data = driver.page_source
    soup = BeautifulSoup(source_data, features="html.parser")
    return soup
```

4) Також з уже отриманої розмітки зможемо отримати інформацію, для цього теж реалізуємо окрему функцію:

```python
def Parse_data(soup):
    news_blocks = soup.find_all('li', class_='col-xs-12 col-sm-12 col-md-12 col-lg-12 fui-grid__inner')
    result = []
    for block in news_blocks:
        title_tag = block.find('h4', class_='media-block__title')
        perex_tag = block.find('p', class_='perex perex--mb perex--size-3')
        date_tag = block.find('span', class_='date date--mb date--size-3')
        link_tag = block.find('a', href=True)

        title = title_tag.get_text(strip=True) if title_tag else None
        perex = perex_tag.get_text(strip=True) if perex_tag else None
        date = date_tag.get_text(strip=True) if date_tag else None
        link = link_tag['href'] if link_tag else None
        result.append({
            'title': title,
            'perex': perex,
            'date': date,
            'link': 'https://www.radiosvoboda.org'+link
            })

    return pd.DataFrame(result)
```

Будемо збирати заступні дані:

- Заголовок
- Короткий опис
- Дату публікації
- Посилання на сторінку з новиною

5) Функція яка приймає в якості аргументів місць та рік та повертає pandas датафрейм з новинами за заданий період часу:

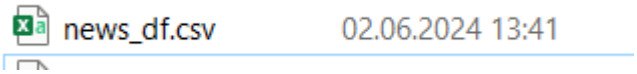
```python
def Get_data(year,month):
    result_df = pd.DataFrame()
    for month in month:
        for day in range (1,31):
            try:
                print('Parse date: '+year+'.'+month+'.'+str(day))
                soup = Get_soup( url: f"https://www.radiosvoboda.org/z/630/{year}/{month}/{day}", driver)
                result_df = pd.concat( objs: [result_df, Parse_data(soup)], ignore_index=True)
            except Exception as e:
                print(f"Error occurred for {month}/{day}: {e}")
    return result_df
```

Результат роботи:

```
Parse date: 2024.4.1
Parse date: 2024.4.2
Parse date: 2024.4.3
Parse date: 2024.4.4
Parse date: 2024.4.5
Parse date: 2024.4.6
Parse date: 2024.4.7
Parse date: 2024.4.8
Parse date: 2024.4.9
Parse date: 2024.4.10
```

Отриманий датафрейм збережемо в csv файл:

```
51  result=Get_data( year: '2024', month: ['4','5'])
52  result.to_csv( path_or_buf: 'news_df.csv', index=False,encoding='utf-8')
```



В результаті за період квітень-травень 2024р було отримано 1741 запис:

	A	B	C	D	E	F
1717	Уряд Словенії визнав Палестинську державу, рішення потребує схвален	Рішення уряду Словенії	22:09, 30 травня 202	https://www.radiosvoboda.org/a/news-uriad-slovenii-palestynska-derzhava/32972893.html		
1718	«Маємо підтвердження щодо збільшення наших спроможностей ППО» -	«Буде найближчим часом	21:55, 30 травня 202	https://www.radiosvoboda.org/a/news-zelenskyy-protypovitryana-oborona/32972888.html		
1719	У Керченському порту, ймовірно, пришвартований пором, про пошкодже	Генштаб ЗСУ заявив, що	21:44, 30 травня 202	https://www.radiosvoboda.org/a/news-krym-porom/32972880.html		
1720	Міністр оборони Німеччини приїхав в Україну і оголосив про новий пакет	Деталі візиту німецького	21:29, 30 травня 202	https://www.radiosvoboda.org/a/news-nimechchyna-pistorius-dopomoha/32972872.html		
1721	Представників США не буде на церемонії в пам'ять про Раїсі в ООН	Плани провести церемо	21:17, 30 травня 202	https://www.radiosvoboda.org/a/news-ssha-oon-raisi/32972864.html		
1722	Українські військові заявили про ураження ракетами ATACMS поромної	У Генштабі кажуть, що ви	21:03, 30 травня 202	https://www.radiosvoboda.org/a/news-atacms-krym-porom/32972859.html		
1723	«Укренерго» вже четвертий день поспіль не планує вимкнень світла в Ук	31 травня відключення с	20:46, 30 травня 202	https://www.radiosvoboda.org/a/news-ukrenerho-svitlo-vidklyuchennya/32972847.html		
1724	Генштаб: на фронті за день було 84 зіткнення, половина – на двох напр	На Кураківському і Покр	20:34, 30 травня 202	https://www.radiosvoboda.org/a/news-henshtab-zvedennia/32972840.html		
1725	Зеленський: Україна уклала вже 12 безпекових угод на понад 23 млрд д	Напередодні були уклад	20:25, 30 травня 202	https://www.radiosvoboda.org/a/news-zelenskyy-bezpekovi-uhody/32972834.html		
1726	Єврокомісія виступає за присутність у марафоні ЗМІ, які «виражають по	В органі виконавчої влад	20:22, 30 травня 202	https://www.radiosvoboda.org/a/news-eurokomisia-telemarafon/32972833.html		
1727	Воєнкори Радіо Свобода стали лауреатами премії «Честь професії»	Їхній матеріал «ЗСУ пішл	19:54, 30 травня 202	https://www.radiosvoboda.org/a/news-chest-profesii-kushnir-nuzhnenko/32972814.html		
1728	Двоє людей поранені через обстріл Руської Лозової на Харківщині – ОВ	Спочатку в ОВА повідоми	19:16, 30 травня 202	https://www.radiosvoboda.org/a/news-kharkivshchyna-poraneni/32972686.html		
1729	У ГУР уточнили, що уразили чотири російські катери в Криму, з них два -	«Згідно з уточненими да	19:02, 30 травня 202	https://www.radiosvoboda.org/a/news-hur-tunets-krym/32972621.html		
1730	Шрі-Ланка посилює контроль, щоб не допустити використання своїх тер	Коломбо відправить деле	18:41, 30 травня 202	https://www.radiosvoboda.org/a/news-shri-lanka-rosia-viyna/32972603.html		
1731	Латвія не обговорює присутність своїх військ в Україні – голова МЗС	«Я не спекулюю. Наразі	18:22, 30 травня 202	https://www.radiosvoboda.org/a/news-latvia-ukraina-viyska/32972581.html		
1732	Литва приєдналася до коаліції з надання ППО Україні – Науседа	«Україні потрібна зброя,	17:57, 30 травня 202	https://www.radiosvoboda.org/a/news-lytva-koalitsia-ppo/32972551.html		
1733	У МЗС Росії заявили, що готують відповідь на обмеження для російськи	Заходи у відповідь, за сл	17:34, 30 травня 202	https://www.radiosvoboda.org/a/news-polshcha-rosia-dyplomaty/32972521.html		
1734	В Ірані відкрили реєстрацію кандидатів на президентських виборах	Період реєстрації триват	17:18, 30 травня 202	https://www.radiosvoboda.org/a/news-iran-reestratsia-vybory-prezydenta/32972505.html		
1735	Неурядові організації Грузії планують оскаржувати закон про «іноагентів	Кілька НУО оголосили, ш	17:06, 30 травня 202	https://www.radiosvoboda.org/a/news-hruziia-neuriadovi-orhanizatsii-inoahenty/32972369.html		
1736	Чеська ініціатива забезпечуватиме Україну десятками тисяч снарядів щ	«Чехія також усвідомлює	16:45, 30 травня 202	https://www.radiosvoboda.org/a/news-chekhia-ukraina-boeprypasy/32972349.html		
1737	Генштаб: інтенсивні бої на Покровському напрямку тривають, найгарячі	Одну з двох атак Сили об	16:34, 30 травня 202	https://www.radiosvoboda.org/a/news-henshtab-pokrovskyi-napriamok-boii/32972339.html		
1738	Україна все ще може перемогти, але з допомогою союзників – генсекрет	«Україна продовжує воюв	16:27, 30 травня 202	https://www.radiosvoboda.org/a/news-stoltenberg-ukraina-viyna/32972329.html		
1739	Данія не проти застосування F-16 для ударів по військових цілях у Росії	«Ми про це дуже чітко за	16:04, 30 травня 202	https://www.radiosvoboda.org/a/news-daniia-f-16-viiskovi-tsili-rosiia/32972304.html		
1740	Рятувальники гасять 15 лісових пожеж на Харківщині, більшість із них сп	За повідомленням ДСНС	15:30, 30 травня 202	https://www.radiosvoboda.org/a/news-lisovi-pozhezhi-kharkivshchyna-rosiiski-obstrily/32972265.html		
1741	Голова МЗС Чехії не бачить наразі перспективи надсилання чеських вій	«Ми маємо захист нашог	15:03, 30 травня 202	https://www.radiosvoboda.org/a/news-mzs-chehii-perspektyva-nadsylannia-cheskyh-viiskovyh/32972222.		