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

Лабораторна робота №5

Мертва русня

Виконав
студент групи ІП-11:
Панченко С. В.

Перевірила:
Баришич Л. М

Київ 2023

ЗМІСТ

1 МЕТА ЛАБОРАТОРНОЇ РОБОТИ

Знаходження мертвих чмобиків.

2 ЗАВДАННЯ

Завдання видане Будановим.

3 ВИКОНАННЯ

Для початку імпортуємо модулі.

In [21]:
```
import requests
import spacy
from lxml.html import fromstring
```

Рисунок 3.1.1 - Імпортовані модулі

Оголосимо назви основних URL.

In [22]:
```
base_url = 'https://pestrecy-rt.ru/news/tag/list/specoperaciia'
list_path = '/html/body/main/ul/li'
headline_xpath = 'a/div[1]/h2/text()'
next_page_xpath = '/html/body/main/div[2]/div/a'
next_page_xpath_2 = '/html/body/main/div[2]/div/a[2]'
```

Рисунок 3.1.2 - Назви url

Зіскрапимо всі заголовки.

In [23]:
```
headlines = []

def get_next_page_url(soup, is_first):
    try:
        if is_first:
            return soup.xpath(next_page_xpath)[0].get('href')
        else:
            return soup.xpath(next_page_xpath_2)[0].get('href')
    except IndexError:
        return None

def extract_headlines(response_text):
    responsed = fromstring(response_text)
    return [element.xpath(headline_xpath)[0] for element in responsed.xpath(list_path)]

is_first_page = True
while True:
    response = requests.get(base_url)
    headlines.extend(extract_headlines(response.text))
    next_page_url = get_next_page_url(fromstring(response.text), is_first_page)

    if not next_page_url:
        break
    else:
        is_first_page = False
        base_url = next_page_url
```

```
In [24]: headlines = [headline.strip() for headline in headlines]
         headlines
```

```
Out[24]: ['Представляем список необходимой гуманитарной помощи для бойцов СВО',
          'Военный комиссар РТ поздравил военных пенсионеров с профессиональным праздником',
          '«Надо Родину защищать!»: отец участников СВО гордится своими сыновьями',
          'В храме деревни Куюки откроют сбор гуманитарной помощи мобилизованным',
          'Ефрейтор из села Пестрецы рассказал о службе на СВО и планах на будущее',
          'На участвовавших в СВО татарстанцев и их детей ввели единовременную выплату',
          'Соцфонд Татарстана рассказал о выплатах участникам СВО и членам их семей',
          'Жизнью военнослужащего в зоне СВО из села Карповка поделились его родители',
          'Участники СВО, получившие инвалидность, могут получить страховую выплату до 2,3 млн',
          'Участвующие в СВО военные пенсионеры будут получать компенсацию пенсий',
          'Военнослужащие в Татарстане получат льготу при предоставлении земельных участков',
          'Пестречинцы могут принять участие в сборе гуманитарной помощи военным',
          'Зарплата участников СВО выросла до 210 тысяч рублей',
          'Названы топ-3 вещи, необходимые военнослужащему на СВО',
          'В пестречинском приюте «Шатлык» состоялась встреча с участником СВО',
          'В Пестрецах объявлен сбор гуманитарной помощи для участников СВО',
          'Жители села Пановка плетут маскировочные сети для бойцов в зоне СВО',
          'Глава Пестрецов поздравил маму Героя России Ивана Додосова с днем рождения',
          'Уроженец Пестречинского района попал в число героев на выставке в Казани',
          'Участникам СВО из Татарстана выплатят по 305 тыс. рублей',
          'В Татарстане организована постоянная поддержка военнослужащих СВО и членов их семей',
          'Участник СВО из села Карповка Пестречинского района рассказал о службе',
          'Проводили в последний путь участника СВО из Пестречинского района',
          'Пестречинские бойцы поблагодарили за нужные вещи в зоне СВО',
          'В Татарстане планируют принять законопроект о предоставлении земли участникам СВО'
```

Рисунок 3.1.3 - Назви заголовків

Відфільтруємо імена мертвих чмобиків.

```
In [25]: from spacy_download import load_spacy
         import pandas as pd

         core = spacy.load('ru_core_news_sm')
         chmobics = set()
         for l in headlines:
             chmobics.update([ent.text for ent in core(l).ents if ent.label_ == "PER"])

         pd.DataFrame(chmobics, columns=['Chmobics'])
```

Out[25]:

	Chmobics
0	Минниханов
1	Куюков
2	Тамара Лаптева
3	Елена Корчагина
4	Расиму Баксикову
5	Эдуард Шарафиев
6	Ивана Додосова
7	Тинчурина
8	Жукова
9	Лейла Фазлеева
10	Александром Агафоновым
11	Валерием Межва
12	Виталий Беляев
13	Путин
14	Иван Додосов
15	Соцфонд
16	Пестрецов

Рисунок 3.1.4 - Імена чмобиків

ДОДАТОК А ТЕКСТИ ПРОГРАМНОГО КОДУ

Тексти програмного коду

(Найменування програми (документа))

Жорсткий диск

(Вид носія даних)

(Обсяг програми (документа), арк.)

Студента групи ІП-113 курсу

Панченка С. В

```
      import requests
      import spacy
      from lxml.html import fromstring
      base_url = 'https://pestrecy-rt.ru/news/tag/list/specoperaciia'
      list_path = '/html/body/main/ul/li'
      headline_xpath = 'a/div[1]/h2/text()'
      next_page_xpath = '/html/body/main/div[2]/div/a'
      next_page_xpath_2 = '/html/body/main/div[2]/div/a[2]'
      headlines = []
      def get_next_page_url(soup, is_first):
      try:
      if is_first:
      return soup.xpath(next_page_xpath)[0].get('href')
      else:
      return soup.xpath(next_page_xpath_2)[0].get('href')
      except IndexError:
      return None
      def extract_headlines(response_text):
      responsed = fromstring(response_text)
      return [element.xpath(headline_xpath)[0] for element in
responsed.xpath(list_path)]
      is_first_page = True
      while True:
      response = requests.get(base_url)
      headlines.extend(extract_headlines(response.text))
      next_page_url = get_next_page_url(fromstring(response.text),
is_first_page)
      if not next_page_url:
      break
      else:
      is_first_page = False
      base_url = next_page_url
      headlines = [headline.strip() for headline in headlines]
      headlines
      from spacy_download import load_spacy
      import pandas as pd
      core = spacy.load('ru_core_news_sm')
      chmobics = set()
      for l in headlines:
      chmobics.update([ent.text for ent in core(l).ents if ent.label_ == "PER"])
      pd.DataFrame(chmobics, columns=['Chmobics'])
```