Kód v Pythone na scraping web stránky FHI

Vypracovala: Bc. Romana Halamová

Kód na extrakciu obsahu z webovej stránky Fakulty hospodárskej informatiky bol vytvorený v jazyku Python s využitím knižníc requests, BeautifulSoup a modulov zo štandardnej knižnice ako os, re a urllib.parse. Na sťahovanie súborov a vyhľadávanie v HTML štruktúre boli použité overené open-source riešenia, ktoré sú podrobne dokumentované a široko využívané v oblasti web scrapingu.

Knižnica 'requests' pre sťahovanie obsahu z webu:

REITZ, K. a kol. (2025). *Requests: HTTP for Humans*. [online]. Verzia 2.32.3. Dostupné na: https://docs.python-requests.org/

Knižnica 'BeautifulSoup' na parsovanie HTML dokumentov:

RICHARDSON, L. (2025). *Beautiful Soup Documentation*. [online]. Dostupné na: https://www.crummy.com/software/BeautifulSoup/

Moduly zo štandardnej knižnice – os, re, urlib.parse:

PYTHON SOFTWARE FOUNDATION. (2025). *The Python Standard Library*. [online]. Dostupné na: https://docs.python.org/3/library/

Kód:

```
# Nacitanie potrebnych kniznic

import os

import requests

from bs4 import BeautifulSoup

from urllib.parse import urljoin, urlparse

import re


# Funkcia na stiahnutie a extrahovanie textu z odsekov <p>, <li>, <span> na stranke

def get_page_paragraphs(url, visited, file):

    try:

        if url in visited:

            return  # Stranka uz bola navstivena

        visited.add(url)
```

```
response = requests.get(url)
response.raise_for_status()

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

# Hladanie vsetkych odsekov <p>, <li>, <span>
paragraphs = soup.find_all(['p', 'li', 'span'])

# Ak najde odseky, zapise ich do suboru
if paragraphs:
    print(f'\nURL: {url}')
    file.write(f'\nURL: {url}\n')
    for paragraph in paragraphs:
        paragraph_text = paragraph.get_text(separator=' ', strip=True)

        # Odstrani odkazy na socialne siete z textu
        if not contains_forbidden_content(paragraph_text):
            file.write(paragraph_text + '\n')

# Najdenie a prehladanie vsetkych odkazov na podstranky a PDF/Word subory
for link in soup.find_all('a', href=True):
    next_page = urljoin(url, link['href'])

    # Stiahnutie PDF alebo Word dokumentov
    if next_page.endswith(('.pdf', '.doc', '.docx')):
        download_file(next_page)
    # Preskoci URL, ktore su zakazane (socialne siete)
    if contains_forbidden_content(next_page):
        continue
    # Skontroluje, ci URL spada pod pozadovane adresy
    if is_allowed_url(next_page):
```

```
                get_page_paragraphs(next_page, visited, file)

    except requests.exceptions.RequestException as e:
        print(f"Chyba pri sťahovaní stránky: {e}")

# Funkcia na kontrolu, či URL patri do povolenych zakladov stranky
def is_allowed_url(url):
    allowed_bases = [
        'https://fhi.euba.sk/fakulta/profil-fakulty',
        'https://fhi.euba.sk/fakulta/historia-fakulty#strucna-historia-fakulty',
        'https://fhi.euba.sk/fakulta/dekanat-fakulty',
        'https://fhi.euba.sk/fakulta/organy-fakulty',
        'https://fhi.euba.sk/fakulta/predpisy',
        'https://fhi.euba.sk/fakulta/telefonny-zoznam',
        'https://fhi.euba.sk/fakulta/oznamy-fakulty',
        'https://fhi.euba.sk/katedry/katedra-aplikovanej-informatiky/oznamy',
        'https://fhi.euba.sk/katedry/katedra-aplikovanej-informatiky/profil-katedry',
        'https://fhi.euba.sk/katedry/katedra-aplikovanej-informatiky/clenovia-katedry',
        'https://fhi.euba.sk/katedry/katedra-aplikovanej-informatiky/externa-spolupraca',
        'https://fhi.euba.sk/katedry/katedra-matematiky-a-aktuarstva/oznamy',
        'https://fhi.euba.sk/katedry/katedra-matematiky-a-aktuarstva/profil-katedry',
        'https://fhi.euba.sk/katedry/katedra-matematiky-a-aktuarstva/clenovia-katedry',
        'https://fhi.euba.sk/katedry/katedra-matematiky-a-aktuarstva/externa-spolupraca',
        'https://fhi.euba.sk/katedry/katedra-operacneho-vyskumu-a-ekonometrie/oznamy',
        'https://fhi.euba.sk/katedry/katedra-operacneho-vyskumu-a-ekonometrie/profil-katedry',
        'https://fhi.euba.sk/katedry/katedra-operacneho-vyskumu-a-ekonometrie/clenovia-katedry',
        'https://fhi.euba.sk/katedry/katedra-operacneho-vyskumu-a-ekonometrie/externa-spolupraca',
        'https://fhi.euba.sk/katedry/katedra-statistiky/oznamy',
        'https://fhi.euba.sk/katedry/katedra-statistiky/profil-katedry',
        'https://fhi.euba.sk/katedry/katedra-statistiky/clenovia-katedry',
        'https://fhi.euba.sk/katedry/katedra-statistiky/externa-spolupraca',
```

```
        'https://fhi.euba.sk/katedry/katedra-uctovnictva-a-auditorstva/oznamy',
        'https://fhi.euba.sk/katedry/katedra-uctovnictva-a-auditorstva/profil-katedry',
        'https://fhi.euba.sk/katedry/katedra-uctovnictva-a-auditorstva/clenovia-katedry',
        'https://fhi.euba.sk/uchadzaci-o-studium/',
        'https://fhi.euba.sk/studium/',
        'https://fhi.euba.sk/veda-a-vyskum/svoc',
        'https://fhi.euba.sk/veda-a-vyskum/mobilitne-programy',
        'https://fhi.euba.sk/medzinarodne-vztahy',
        'https://programy.euba.sk/fakulta/fakulta-hospodarskej-informatiky/'

    ]
    return any(base in url for base in allowed_bases)




# Funkcia na kontrolu, ci URL obsahuje zakazane vzory(socialne siete)
def contains_forbidden_content(text):
    forbidden_patterns = [
        'facebook.com', 'instagram.com', 'linkedin.com',
        'whatsapp.com', 'twitter.com', 'tiktok.com',
        'youtube.com', 'pinterest.com', 'snapchat.com'
    ]
    return any(pattern in text for pattern in forbidden_patterns)

# Funkcia na kontrolu, ci nazov suboru obsahuje nepovolene znaky
def contains_invalid_characters(file_name):
    invalid_chars = r'[<>:"/\\|?*]'
    return bool(re.search(invalid_chars, file_name))

# Funkcia na stiahnutie PDF alebo Word suboru
```

```
def download_file(file_url):
    try:
        # Extrahovanie nazvu suboru
        file_name = file_url.split('/')[-1]

        # Kontrola na nepovolene znaky
        if contains_invalid_characters(file_name) or contains_forbidden_content(file_name):
            print(f"Súbor {file_name} obsahuje nepovolené znaky a bude preskočený.")
            return  # Preskočí sťahovanie súboru

        response = requests.get(file_url)
        response.raise_for_status()

        # Ulozenie suboru do priecinka 'fhi_download'
        if not os.path.exists('fhi_download'):
            os.makedirs('fhi_download')
        file_path = os.path.join('fhi_download', file_name)
        with open(file_path, 'wb') as file:
            file.write(response.content)

        print(f'Subor {file_name} bol uspesne stiahnuty.')
    except requests.exceptions.RequestException as e:
        print(f"Chyba pri stahovani suboru: {e}")

# Hlavna cast skriptu
if __name__ == '__main__':
    url = 'https://fhi.euba.sk/'
    visited = set()  # Mnozina navstivenych URL
    print(f"Prehľadávanie začalo na URL: {url}")

    # Otvorenie suboru na zapisovanie extrahovaneho textu
```

```
with open('text_subor_fhi.txt', 'w', encoding='utf-8') as file:
    get_page_paragraphs(url, visited, file)
```