Data Science

WEB SCRAPING

COMMAND PROMPT

conda activate myenv

conda create -n myenv python=3.6

conda activate myenv

conda install jupyter

jupyter notebook

JUPYTER NOTEBOOK

1. Инсталирање библиотеки

(документација: https://www.crummy.com/software/BeautifulSoup/bs4/doc/)

```
!pip install requests

!pip install pandas

!pip install beautifulsoup4
```

2. Импорт на библиотеки

```
import pandas as pd

from bs4 import BeautifulSoup

import requests
```

3. Земање сајт

```
url= https://publicitet.mk/category/kultura/
response = requests.get(url)
raw_html=response.text
soup = BeautifulSoup(raw_html, 'html.parser')
print(soup.prettify())
```

4. Селектирање на картички

```
articles = soup.select("body > div.jeg_viewport > div.jeg_main > div > div.jeg_content >
div.jeg_section > div > div.jeg_cat_content.row > div.jeg_main_content.jeg_column.col-sm-8 > div >
div.jnews_category_content_wrapper > div > div.jeg_block_container >
div.jeg_posts.jeg_load_more_flag > article")

// земање една картичка
article = articles[0]
//земање на делови од неа користејќи селектори
```

title = article.select_one('body > div.jeg_viewport > div.jeg_main > div > div.jeg_content > div.jeg_section > div > div.jeg_cat_content.row > div.jeg_main_content.jeg_column.col-sm-8 > div > div.jnews_category_content_wrapper > div > div.jeg_block_container > div.jeg_posts.jeg_load_more_flag > article > div.jeg_postblock_content > h3').**text.strip()**
date = article.select_one('body > div.jeg_viewport > div.jeg_main > div > div.jeg_content > div.jeg_section > div > div.jeg_cat_content.row > div.jeg_main_content.jeg_column.col-sm-8 > div > div.jnews_category_content_wrapper > div > div.jeg_block_container > div.jeg_posts.jeg_load_more_flag > article > div.jeg_postblock_content > div.jeg_post_meta > div.jeg_meta_date > a').text.strip()
img = article.select_one('body > div.jeg_viewport > div.jeg_main > div > div.jeg_content > div.jeg_section > div > div.jeg_cat_content.row > div.jeg_main_content.jeg_column.col-sm-8 > div > div.jnews_category_content_wrapper > div > div.jeg_block_container > div.jeg_posts.jeg_load_more_flag > article > div.jeg_thumb > a > div > img')**.get('data-src')**
descp = article.select_one("body > div.jeg_viewport > div.jeg_main > div > div.jeg_content > div.jeg_section > div > div.jeg_cat_content.row > div.jeg_main_content.jeg_column.col-sm-8 > div > div.jnews_category_content_wrapper > div > div.jeg_block_container > div.jeg_posts.jeg_load_more_flag > article > div.jeg_postblock_content > div.jeg_post_excerpt > p")**.text.strip()**

//земање нова страница за да се земе атрибут од друга страница за ист објект (повторување на горните чекори)
read_news = soup.select_one("body > div.jeg_viewport > div.jeg_main > div > div.jeg_content > div.jeg_section > div > div.jeg_cat_content.row > div.jeg_main_content.jeg_column.col-sm-8 > div > div.jnews_category_content_wrapper > div > div.jeg_block_container > div.jeg_posts.jeg_load_more_flag > article > div.jeg_postblock_content > div.jeg_post_excerpt > a").get('href')
views_response=requests.get(read_news)
views_html=views_response.text
views_soup = BeautifulSoup(views_html, 'html.parser')
views_count= views_soup.select_one("body > div.jeg_viewport > div.post-wrapper > div.post-wrap > div.jeg_main > div > div > div > div.row > div.jeg_main_content.col-md-8 > div > div.entry-content.no-share > div.content-inner > div > span.post-views-count").text.strip()

5. Додавање на сите во DICTIONARY

```
import requests
from bs4 import BeautifulSoup

# Assuming you have already fetched the page and have the 'articles' list ready
# Example: articles = soup.select("your-article-selector-here")

# Initialize the list to store parsed articles
parsed_articles = []

# Loop through each article element
for article in articles:
(селектирање на сите делови од горе, код се повторува)
    title = article.select_one('body > div.jeg_viewport > div.jeg_main > div > div.jeg_content > div.jeg_section > div > div.jeg_cat_content.row > div.jeg_main_content.jeg_column.col-sm-8 > div > div.jnews_category_content_wrapper > div > div.jeg_block_container > div.jeg_posts.jeg_load_more_flag > article > div.jeg_postblock_content > h3').text.strip()
```

```
    date = article.select_one('body > div.jeg_viewport > div.jeg_main > div > div.jeg_content > div.jeg_section > div > div.jeg_cat_content.row > div.jeg_main_content.jeg_column.col-sm-8 > div > div.jnews_category_content_wrapper > div > div.jeg_block_container > div.jeg_posts.jeg_load_more_flag > article > div.jeg_postblock_content > div.jeg_post_meta > div.jeg_meta_date > a').text.strip()
    img = article.select_one('body > div.jeg_viewport > div.jeg_main > div > div.jeg_content > div.jeg_section > div > div.jeg_cat_content.row > div.jeg_main_content.jeg_column.col-sm-8 > div > div.jnews_category_content_wrapper > div > div.jeg_block_container > div.jeg_posts.jeg_load_more_flag > article > div.jeg_thumb > a > div > img').get('data-src')
    descp = article.select_one('body > div.jeg_viewport > div.jeg_main > div > div.jeg_content > div.jeg_section > div > div.jeg_cat_content.row > div.jeg_main_content.jeg_column.col-sm-8 > div > div.jnews_category_content_wrapper > div > div.jeg_block_container > div.jeg_posts.jeg_load_more_flag > article > div.jeg_postblock_content > div.jeg_post_excerpt > p').text.strip()

    read_news = article.select_one("body > div.jeg_viewport > div.jeg_main > div > div.jeg_content > div.jeg_section > div > div.jeg_cat_content.row > div.jeg_main_content.jeg_column.col-sm-8 > div > div.jnews_category_content_wrapper > div > div.jeg_block_container > div.jeg_posts.jeg_load_more_flag > article > div.jeg_postblock_content > div.jeg_post_excerpt > a").get('href')

    # Get views count from the article page
    views_response = requests.get(read_news)
    views_html = views_response.text
    views_soup = BeautifulSoup(views_html, 'html.parser')
    views_count = views_soup.select_one("body > div.jeg_viewport > div.post-wrapper > div.post-wrap > div.jeg_main > div > div > div > div.row > div.jeg_main_content.col-md-8 > div > div.entry-content.no-share > div.content-inner > div > span.post-views-count").text.strip()

    # Create a dictionary to store the article information
    product_articles = {
        "Title": title,
        "Date": date,
        "Image": img,
        "Description": descp,
        "Views": views_count
    }

    # Append the dictionary to the parsed_articles list
    parsed_articles.append(product_articles)

# Now parsed_articles contains all the scraped articles
```

6. Користење ПАНДА за изглед на табела

```
df = pd.DataFrame(parsed_articles)
```

7. Импортирање на фајлот

```
df.to_csv("publicitet.csv", index=False, encoding='utf-8-sig')
```