Team-MakeChange

Blog Generation Project part 1

This project scrapes random blog posts from WikiHow, extracts subheadings and paragraphs, and stores them in a CSV file. The project utilizes Python libraries like `requests` and `BeautifulSoup` for web scraping and `csv` for data storage.

Project Structure

```
blog_generation_project/
│
├── data/
│   └── wikihow.csv          │
├── notebooks/
│   ├── web_scraping.ipynb
│   └── model_fine_tuning.ipynb
model
│
├── models/
│   └── checkpoints/
│
├── src/
│   ├── scraping.py
│   └── fine_tuning.py
│
├── requirements.txt
```

└── README.md ```

Requirements

Ensure that you have the following Python libraries installed. You can install them using the provided ` requirements.txt` .

```
-----------pip install -r requirements.txt
```

Dependencies

- ` beautifulsoup4==4.12.2`
- ` requests==2.31.0`

Quick Start Guide

1. Check Internet Connection:

Before starting, ensure that your internet connection is working by running:

```
 import requests


try:
   response = requests.get('https://www.google.com')
   print("Internet connection is working.")
except requests.exceptions.ConnectionError as e:
   print("No internet connection:", e)
```

` ` `

2. Scrape a Random Blog Post from WikiHow:

Use the following code to scrape a random article:

```
import requests
import bs4

url = "https://www.wikihow.com/Special:Randomizer"
response = requests.get(url)
soup = bs4.BeautifulSoup(response.content, 'html.parser')
```

3. Extract Subheadings and Paragraphs:

Extract subheadings and paragraphs using appropriate HTML tags:

```
 import re

subheadings = []
paragraphs = []
steps = soup.find_all('div', {'class': 'step'})

for step in steps:
  subheading_element = step.find('b')
  if subheading_element is not None:
    subheading_text = subheading_element.text.strip().replace('\n', '')
    subheading_text = subheading_text.encode('ascii', errors='ignore').decode()
```

```
        subheading_text = re.sub(r'\r', '', subheading_text)
        subheadings.append(subheading_text)

        # Remove titles and extra links
        subheading_element.extract()
        for span_tag in step.find_all('span'):
            span_tag.extract()

    paragraph_text = step.text.strip().replace('\n', '').replace('  ', ' ')
    paragraph_text = paragraph_text.encode('ascii', errors='ignore').decode()
    paragraph_text = re.sub(r'\r', '', paragraph_text)
    paragraphs.append(paragraph_text)

print(subheadings)
print(paragraphs)
```

4. Save Data to CSV:

Save the extracted data to a CSV file:

```
import os
import csv

for count in range(4000):
    url = 'https://www.wikihow.com/Special:Randomizer'
    response = requests.get(url)
```

```
html_content = response.content
soup = bs4.BeautifulSoup(html_content, 'html.parser')
article_title = soup.find('title').text.strip()
print(article_title + " " + str(count))

subheadings = []
paragraphs = []
steps = soup.find_all('div', {'class': 'step'})

for step in steps:
    subheading_element = step.find('b')
    if subheading_element is not None:
        subheading_text = subheading_element.text.strip().replace('\n', '')
        subheading_text = subheading_text.encode('ascii', errors='ignore').decode()
        subheadings.append(subheading_text)
        subheading_element.extract()
    for span_tag in step.find_all('span'):
        span_tag.extract()
    paragraph_text = step.text.strip().replace('\n', '').replace('  ', ' ')
    paragraph_text = paragraph_text.encode('ascii', errors='ignore').decode()
    paragraphs.append(paragraph_text)

file_path = './data/wikihow.csv'
file_exists = os.path.exists(file_path)

if len(subheadings):
```

```
os.makedirs(os.path.dirname(file_path), exist_ok=True)
with open(file_path, mode='a', newline='', encoding='utf-8') as csv_file:
    writer = csv.writer(csv_file)

    if not file_exists:
        writer.writerow(['Article Title', 'Subheading', 'Paragraph'])

    for i in range(len(subheadings)):
        writer.writerow([article_title, subheadings[i], paragraphs[i]])
```