NLP Assignment 2 Report

Mahnoor Akhtar

September 19, 2023

Contents

0.1 Introduction

The goal of this project was to scrape Urdu stories from the website "https://www.urduzone.net/" using the Scrapy framework in Python. The website contains a collection of stories, and we wanted to extract the text content of these stories for analysis or other purposes.

0.2 Approach

0.2.1 Project Setup: A Scrapy project was created using the 'scrapy startproject' command. This generated the basic project structure, including a 'spiders' directory.

0.2.2 Web Inspection: The target website was inspected using the browser's developer tools to identify the HTML tags that contained the text of the stories. It was discovered that each story's title was encapsulated within an `<h3>` tag, which also contained a link leading to the individual story pages. The story text was enclosed within `<p>` tags on each story page.

0.2.3 Spider Implementation: A spider named "i200635_urdu_stories_spider" was created using the 'scrapy genspider' command. Two callback methods were defined:

`parse`: This method was responsible for extracting the URLs of the individual story pages. It accomplished this by selecting the `<h3>` tags and their child `<a>` elements.

`parse_story`: This method focused on scraping the text content of the stories. It used CSS selectors to target and extract text from the `<p>` tags on each story page.

0.2.4 Text Cleaning: The scraped text was cleaned to remove newline characters (`\n`) and ensure data consistency. This was achieved using Python's `replace` method and stripping leading and trailing whitespaces.

0.2.5 Data Storage: The cleaned story text was stored in a CSV file named "urdu_stories.csv". This file was created in the same directory as the Scrapy project. New data rows were appended to the file for each story.

0.3 Challenges Faced

The following challenges were faced during the implementation of the approach:

- Identifying the Relevant HTML Tags: The primary challenge was to identify the HTML tags that contained the text of the stories. The inspection of the website's structure was instrumental in discovering the appropriate tags (`<h3>` for titles and `<p>` for story text).
- Text Cleaning: Cleaning the scraped text to remove newline characters and ensure data consistency was a minor challenge. However, it was crucial for maintaining well-structured data.
- CSV File Handling: Proper management of the CSV file, including appending new data rows without overwriting existing data or creating duplicates, required careful handling.
- Logging and Debugging: Debugging the spider and ensuring that it correctly scraped and stored data were vital. Scrapy's built-in logging capabilities were used to monitor the process.

0.4 Conclusion

The approach outlined in this report was successfully used to scrape Urdu stories from a website using Scrapy. The scraped data was cleaned and stored in a CSV file for further analysis or use. The project provided practical experience in web scraping, data cleaning, and data storage with Scrapy, while also addressing various challenges along the way.