NLP

ASSIGNMENT 2

September 21, 2023

1 Introduction

Web scrapping allows to extract data from websites. Scrapy is a web crawling and scrapping framework for python used to collect structured data from websites. We extracted urdu stories from website, ”https://www.urduzone.net/.” The spider will scrap through all stories and save them in CSV file.

2 Approach used

2.1 Scrapy setup

- Install Scrapy using pip.
- Create a new Scrapy project for this assignment using the command: ‘scrapy startproject urdu stories‘.
- Create a new spider named ‘I20XXXXurdu stories spider‘ to scrape (https://www.urduzone.net).
- Set the start URL to (https://www.urduzone.net).

2.2 Navigating to website

- Start from the ”start url” and navigate to the main page by searching space bar in search
- Follow links to search for urdu stories
- Increment the page number in current page link to navigate through 226 pages

2.3 Extract Urdu stories

- At first, extract all lines from text page
- Write a Urdu pattern expression to compare all text you extracted
- Find all Urdu words in a story and concatenate the Urdu words into a single string
- Append the concatenated Urdu word string to the CSV file
- Use yield to write output in terminal

3 Challenges faced

The main challenge faced in this assignment was navigating through pages from start URL to extract all stories. At first, there were only 14 stories on main page with URL: (https://www.urduzone.net). Then i tried to navigate to other pages by using HTML tags of next page icon which didn't work. Then I tried to search space bar in search bar to navigate to the main page with URL: ”https://www.urduzone.net/?s=+” and then I moved to other pages by incrementing page number in the URL to extract all stories from 226 pages.