CS4063 - Natural Language Processing

Due Date: Thursday, September 21st by 11:55pm.

Assignments are to be done individually. No late assignments will be accepted.
Submissions that do not comply with the specifications given in this document will not be marked and a zero grade will be assigned.
Write your name and e-mail id in a comment line or Markdown cell in on top of each source file. You are required to submit a zip containing all the project files, a CSV file containing the cleaned data and a PDF report. Your zip file must be named i20-XXXX.zip where i20-XXXX represents your student id.

Web Scraping with Scrapy - Collecting Urdu Stories

1 Objective

The objective of this assignment is to develop a Scrapy spider to scrape the website (`https://www.urduzone.net`) and extract Urdu stories while removing tags and non-Urdu words.

2 Instructions

2.1 1. Setup

- Create a new Scrapy project for this assignment using the command: 'scrapy startproject urdu_stories'.

2.2 2. Spider Development

- Inside your Scrapy project, create a new spider named 'I20XXXXurdu_stories_spider' to scrape (https://www.urduzone.net). Where I201234 is your roll no.
- Set the start URL to (https://www.urduzone.net).
- Configure the spider to crawl through the website and extract the necessary data.
- Use Scrapy selectors to extract text content from the website.

2.3 3. Data Extraction

- Extract the text of the stories from the website while removing any HTML tags.
- Implement a mechanism to filter out non-Urdu words and numbers and keep only the Urdu text.

2.4 4. Data Storage

- Store the collected Urdu stories in a structured format such as a CSV file within the Scrapy project directory.

2.5 5. Documentation

- Provide clear and concise documentation explaining your approach, faced challenges written in LaTeX.
- Include comments in your spider's code to explain the logic and any important steps.

3 Submission

- Submit your Scrapy project folder, including the spider and any necessary files.
- Include a report that explains your approach, challenges faced, and any improvements or optimizations made to the spider.

4 Evaluation Criteria

Your assignment will be evaluated based on the following criteria:

- Successful extraction of Urdu stories from (https://www.urduzone.net) while removing tags and non-Urdu words.
- Proper documentation and comments in the code.
- The quality of the report, including an explanation of your approach and any challenges faced.
- Clean and well-structured code.

Honor Policy

This assignment is a learning opportunity that will be evaluated based on your ability to think in a group setting, work through a problem in a logical manner and write a research report on your own. You may however discuss verbally or via email the assignment with your classmates or the course instructor, but you are to write the actual report for this assignment without copying or plagiarizing the work of others. You may use the Internet to do your research, but the written work should be your own. **Plagiarized reports or code will get a zero**. If in doubt, ask the course instructor.