# NLP Assignment 2

i201822 Rayed Muhammad Saeed

September 2023

## 1 Introduction

In this assignment we were tasked with scraping the website: "https://www.urduzone.net/" using the Python library Scrapy. Using this we were to extract the stories from the pages and store them in a csv file.

## 2 Approach

Firstly I set up the scrapy project and made my spider to crawl on the various links of stories present on the website. Next, I set the start url of the crawler to https://www.urduzone.net/. After doing this I went to the search page of the website by an empty search call. This was done by sending a POST request to the search option present on the website and then sending an empty search, this took us to the page where there were all the story links present. After this I set up the crawler to the maximum number of pages which was 226 and set the selector to the div element tag that had all the stories links present, this allowed me to get all the links on that particular page and then move to the next page and getting the links in the same way. These links were stored in a list. After this I proceeded to visit each of the link and then from there selected the div that had the story in it and using CSS selectors, extracted the story and saved it in a variable, later writing it into a List. Upon successful extraction of the stories I then stored the content which was the story and the title of that story into a formatted csv file. This whole process allowed us to efficiently collect and store the desired data in a well-formatted CSV document.

## 3 Challenges Faced

There were many challenges faced during the implementation of this assignment.

**1.** We had not studied web scraping using Scrapy and this was a challenge as we had to learn it first by watching tutorials and through trial and error.

**2.** The second challenge I faced was that accessing all the stories was too tricky as the main page of the website did not have the links to all the stories

that there were. So, accessing the search page to get all the links of the stories was a challenge.

**3.** Moreover, I faced the issue of cleaning urdu language. This was tricky as using the regex I first wrote, the code also cleaned any commas or dashes (fullstops) from the urdu writing.

These were some of the troubles that I had while doing this assignment and how I found the solutions to those problems encountered.