**Project Overview: Data Scraping using Scrapy Objective:**

Extract insightful data from various websites using Scrapy, a web crawling and scraping framework.

Project Details: Scraping Tool:

Utilized Scrapy for efficient web scraping. Duration: Completed data extraction from 98 out of 100 websites within a span of 2 days.

File Structure: Main Scraper File: url.py

Location: internshala/internshala/urldataextract/urldataextract/spiders/url.py

Description: Primary script responsible for scraping data from different websites.

Input Data Source: Excel File

Name: input.xlsx

Usage: Opened and read using the openpyxl library to obtain website data for scraping.

Output Data Directory: files Location:

internshala/internshala/urldataextract/urldataextract/spiders/files Contains extracted data from

various websites. Files may not be in sequential order due to the random nature of scraping.

Output Formatted File: Location: internshala/internshala/urldataextract/urldataextract/spiders

Contains the organized and formatted output data obtained from the scraping process.

help.py : is the main file where I have written my whole code for different functions that are to I have applied on extracted text to get the insightful data.

Location: internshala/internshala/urldataextract/urldataextract/spiders/help.py

Challenges Faced: Encoding Issue: Encountered encoding problems leading to the skipping of two

files out of 100 during scraping process.

Additionally I have added comments in url.py file and help.py file to get better understanding of the code I have written.