🕸️ Website Crawling Project Specification

Project Title:

Intelligent Web Crawler & Analyzer

🎯 Objective:

Design and implement a smart web crawler capable of analyzing a specific website's crawlability, extracting key metadata, and proposing the best approach to access its data (HTML, Sitemap, RSS, or API). Each group will work on a different website and provide a report + working prototype.

🧩 Tasks Breakdown (for 5 students per group)

👥 Member 1 – Crawlability Specialist

- Analyze the robots.txt file (allowed/disallowed paths, crawl-delay, sitemap links).
- Implement logic to check if crawling is permitted.
- Deliver a summary of crawlability rules.

🕷️ Member 2 – Content Extractor

- Use BeautifulSoup or Scrapy to extract titles, descriptions, links, or specific content.
- Focus on meaningful data related to the website's category (e.g., products, jobs, news).
- Implement retry mechanisms and handle pagination if possible.

🧠 Member 3 – JS & API Handler

- Determine if the site is JavaScript-heavy.
- Propose or implement solutions using Playwright/Selenium if required.
- Check for open APIs or RSS feeds.

📊 Member 4 – Visual & Report Designer

- Build a simple Streamlit dashboard to visualize: Crawlability score, Top extracted data, Recommendations for crawling tools.
- Create visual sitemap if applicable.

📝 Member 5 – Documentation & Deployment

- Document the entire process.
- Create a README with instructions and findings.
- Deploy the project on Streamlit Cloud or a local web server.

Deliverables:

- ✅A deployed Streamlit dashboard
- Full project code on GitHub (public or private repo)
- Final PDF report including: Crawlability analysis, Crawled data preview, Challenges faced, Group reflections
- Optional (for higher scores): CSV or database storage, scheduled crawling

Evaluation Criteria (Total 25 points):

Criteria	Points
Functionality (Crawl, Parse, Display)	5
Creativity in Extracted Data	5
API/RSS/JS Handling	3
Streamlit UI + Visualization	3
Code Quality & GitHub Use	3
Documentation & Report	3
Presentation (Demo)	3

Website Assignment

Each group will be assigned a different website. You may request categorized websites (e.g., news, jobs, recipes, ecommerce) or a randomized list based on popularity and diversity.