Splore – Data Engineer Assignment

Congratulations! This is a fantastic opportunity to highlight your coding skills and problem-solving abilities to your potential employer.

Before you start, carefully read, and understand the assignment's instructions and requirements. Take some time to plan out your approach and consider any potential edge cases or issues that may arise.

Remember to write clean, well-structured code and to comment on your code so that others can easily understand what you have done.

Do not be afraid to ask questions or seek clarification if you are unsure about anything. Give it your best effort and have fun!

Objective:

The objective of this programming assignment is to develop a web scraper that can extract the content of Wiki articles and save them in a collection of documents.

Specifically, your task is to extract all pages from https://animalcrossing.fandom.com/wiki/Animal_Crossing_Wiki

Description:

Your task is to create a web scraper that can collect data from these articles and store them in a collection of documents. The documents should contain the title of the article, its summary, and its content.

Instructions:

1. You can use any programming language of your choice (though Python is preferred)
2. You can use any web scraping library of your choice.
3. The web scraper should take a list of URLs as input.
4. The scraper should crawl the URL and extract the title, and content of the Wiki article.
5. The scraper should store the extracted data in a collection of documents, where each document should contain the title, and content of a single Wiki article.
6. The scraper should handle any errors or exceptions that may arise during the scraping process.

7. You should provide a README file that explains how to run the program, any libraries that need to be installed, and any other relevant information.
8. The scraper should be designed so that it can be run weekly. In each weekly run, it should include logic to determine if there are any changes in any of the pages and report the changes.

Deliverables:

9. The source code of the web scraper.
10. A collection of documents containing the title, summary, and content of the Wiki articles.
11. A README file explaining how to run the program and any relevant information.

Evaluation Criteria:

12. Correctness of the web scraper.
13. Efficiency and speed of the web scraper.
14. Quality and completeness of the collected data.
15. Quality and readability of the source code.
16. The completeness and clarity of the README file.

Note:

17. It is important to adhere to ethical scraping practices and respect the terms and conditions of the website you are scraping.

All the best!