Problem:

The task at hand involves building an application to collect news articles from various RSS feeds, store them in a database, and categorize them into predefined categories. This entails challenges such as efficiently parsing and extracting relevant information from diverse RSS feeds, managing the storage of data in a structured manner, and implementing an effective categorization system for the articles.

Approach:

In addressing this problem, I took a systematic approach to ensure a comprehensive and efficient solution. The first step involved creating a script, RSS FEED.py, which systematically parses RSS feeds, extracts crucial details from news articles, and handles duplicate entries to maintain data integrity. To organize the extracted data, I opted for a relational database structure using SQLAlchemy.

For seamless processing and scalability, I integrated Celery, enabling asynchronous management of news articles. The Celery worker played a pivotal role in executing tasks such as category classification using spaCy. A clear-cut keyword-based categorization approach was chosen to enhance understanding and interpretation.

Robust logging mechanisms were implemented throughout the script to track events and handle errors gracefully, ensuring transparency and reliability. My choice of libraries, including Feedparser, SQLAlchemy, Celery, and spaCy, was deliberate, considering their widespread acceptance and suitability for the assignment's demands.

The resulting comprehensive documentation, coupled with the exported CSV file named classified_articles.csv, aims to provide users with a lucid understanding of the script's logic and design choices. This approach was crafted to address the core challenges posed by the assignment, providing an effective, scalable, and maintainable solution.