📅Day 4: Word Frequency Analysis (Keyword Extraction)

📋User Story

Title: Word Frequency Analysis for Keyword Extraction

As a Developer,\ **I want to:**

- Read and parse HTML files.
- Tokenize and extract words from the content.
- Strip HTML tags manually (without using external libraries).
- Count the frequency of all valid (non-grammatical) words.
- Identify and save the most frequently occurring valid word as the page's keyword.

🔗Acceptance Criteria

- HTML tags must be stripped using a custom-built state machine or manual logic.
- A custom, manually provided stop-word list should be used to filter out common grammatical words (e.g., "the," "is," "and").
- Only non-stopwords should be considered for frequency analysis.
- The word with the highest frequency (excluding stopwords) must be identified and recorded as the page’s keyword.
- Each HTML page should be mapped to its corresponding extracted keyword for reference or indexing.
