1. Loading & Initial Inspection

- Dataset includes review metadata like review_id, review_title, review_text, rating, etc.
- Combined review_title and review_text into one field: combined_review.
- Dropped entries with empty reviews.

2. Normalization & Cleaning

- Lowercased all text for uniformity.
- Removed:
    - HTML tags using BeautifulSoup
    - Digits and punctuation using re
    - Extra whitespace and line breaks
- Resulted in a clean, lowercase, plain-text version of each review.

3. Stopword Removal

- Used NLTK's English stopwords list.
- Removed common non-informative words like "the," "is," "and," "was," etc.

4. Lemmatization vs. Stemming (Comparative Analysis)

- **Stemming**: Chops words (e.g., "reading" → "read", "characters" → "charact")
- **Lemmatization**: Converts words to dictionary form while preserving meaning (e.g., "reading" → "read", "was" → "be")

Observations:

- **Lemmatization preserves meaning** better, crucial for sentiment/context analysis.
- **Stemming is faster**, but more aggressive and can distort important review keywords.

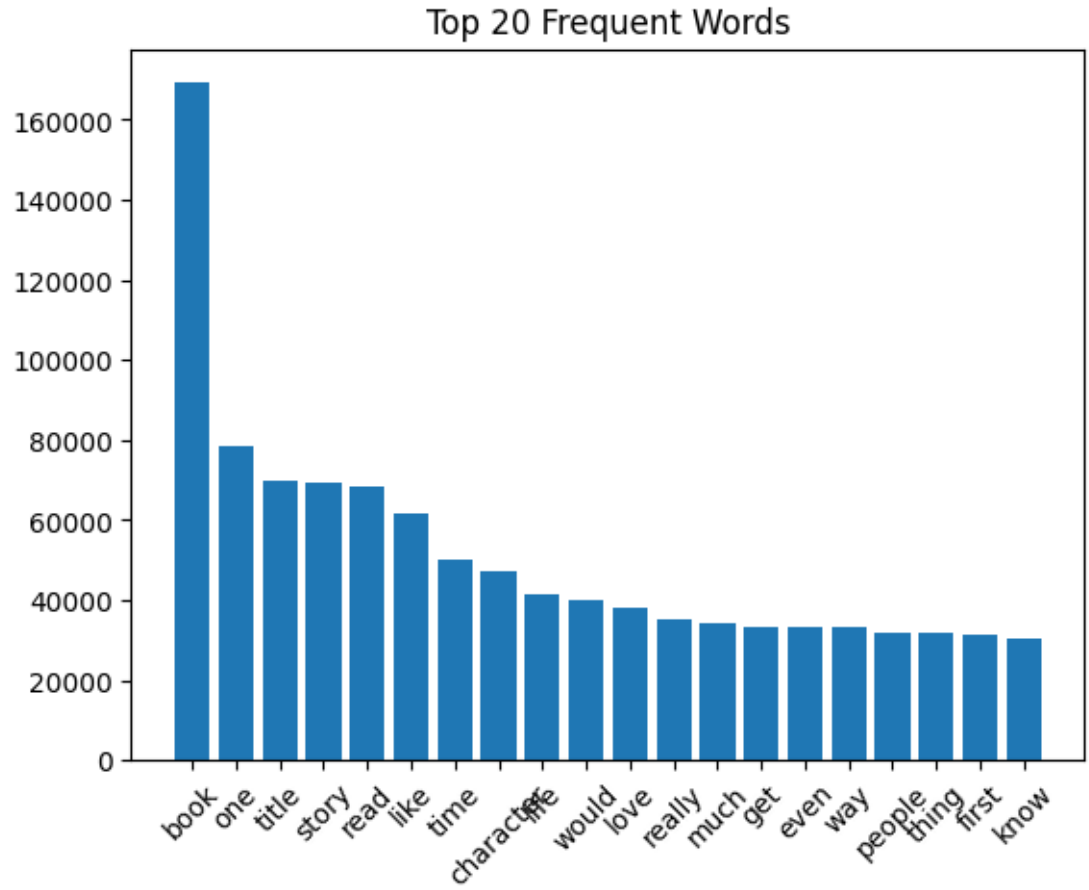
Decision: **Lemmatization** was chosen for better semantic retention.

5. Tokenization & Frequency Analysis

- Tokenized each cleaned review using NLTK.
- Created a global frequency count of all words.

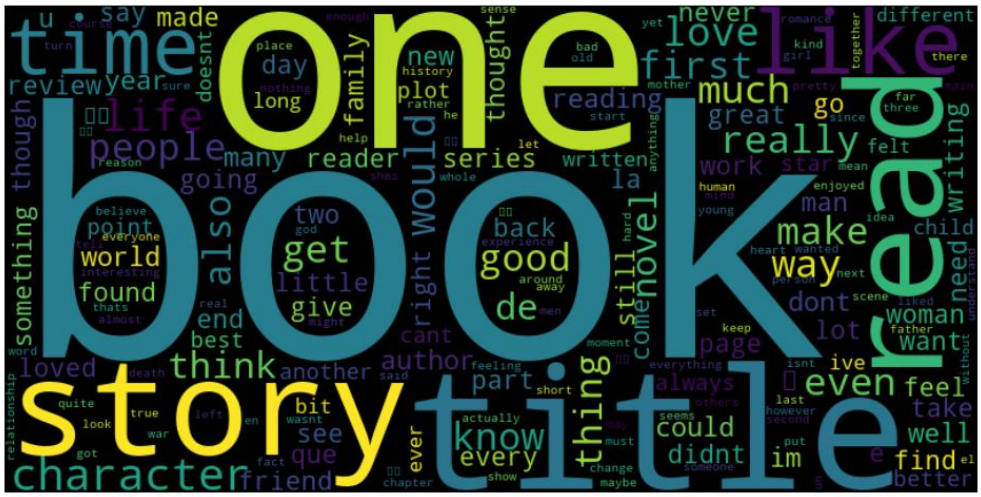
Token Statistics & Visualization

Top 20 Frequent Words (Bar Chart)



- **Most Frequent Tokens**: book, one, story, like, time, character, read, title, get, really
- Indicates that the reviews heavily focus on plot, characters, and the overall experience of reading.

Word Cloud



- Visually represents the most dominant words by size.
- Reaffirms insights from the bar chart: reader sentiment revolves around "book", "story", "character", "read", and "time".