Word Frequency in Classic Novels: Moby Dick

Project Overview

This project analyzes the most frequent words in the classic novel **Moby Dick** by Herman Melville. It demonstrates web scraping, text cleaning, and natural language processing (NLP) using Python. The results are visualized and saved for further exploration.

Step-by-Step Documentation

1. Downloading the Novel

We use the `requests` library to fetch the plain text of Moby Dick from Project Gutenberg.

```
import requests
url = 'https://www.gutenberg.org/files/2701/2701-0.txt'
response = requests.get(url)
response.encoding = 'utf-8'
text = response.text
```

Explanation: - Fetch the text using `requests.get()`. - Set encoding to UTF-8.

---

2. Extracting the Main Content

Project Gutenberg texts include headers and footers. We extract only the main content.

```
def extract_main_text(text):
    start_marker = 'CHAPTER 1. Loomings.'
    end_marker = 'End of the Project Gutenberg EBook of Moby Dick; or The Whale, by Herman Melville'
    start = text.find(start_marker)
    end = text.find(end_marker)
    if start == -1 or end == -1:
        return text
    return text[start:end]
main_text = extract_main_text(text)
```

Explanation: - Locate the start and end of the actual novel using unique markers.

---

3. Cleaning and Tokenizing the Text

Remove punctuation, convert to lowercase, and split into words.

```
import re
def clean_and_tokenize(text):
    text = text.lower()
    text = re.sub(r'[^a-z\s]', '', text)
    words = text.split()
    return words
words = clean_and_tokenize(main_text)
```

Explanation: - Lowercasing ensures uniformity. - Regex removes all non-letter characters.

---

4. Removing Stopwords

Stopwords are common words (like “the”, “and”) that are not meaningful for frequency analysis.

```
import nltk
nltk.download('stopwords', quiet=True)
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))
filtered_words = [w for w in words if w not in stop_words]
```

Explanation: - Use NLTK’s list of English stopwords.

---

5. Counting Word Frequencies

Count the frequency of each word using `collections.Counter`.

```
from collections import Counter
word_counts = Counter(filtered_words)
top_100 = word_counts.most_common(100)
```

Explanation: - `Counter` creates a dictionary of word counts.

---

6. Saving and Visualizing Results

Save the results as a CSV and plot the top 30 words.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
top_100_df = pd.DataFrame(top_100, columns=['word', 'count'])
top_100_df.to_csv('moby_dick_top100_words.csv', index=False)
plt.figure(figsize=(10, 12))
sns.set_style('whitegrid')
sns.barplot(data=top_100_df.head(30), y='word', x='count', palette='Blues_d')
plt.title('Top 30 Most Frequent Words in Moby Dick')
plt.xlabel('Count')
plt.ylabel('Word')
plt.tight_layout()
plt.savefig('moby_dick_top30_words.png', dpi=300)
plt.show()
```

Explanation: - Use pandas to save the results as a CSV. - Seaborn and matplotlib for a clean, readable bar chart.

Example Output

- **moby_dick_top100_words.csv**: Top 100 most frequent words and their counts
- **moby_dick_top30_words.png**: Bar chart of the top 30 words

---

README for GitHub

```
# Word Frequency in Classic Novels

This project analyzes the most frequent words in the classic novel **Moby Dick** by Herman Melville. It demonstrates web scraping, text cleaning, and natural language processing (NLP) using Python.

## Features
- Downloads the full text of Moby Dick from Project Gutenberg
- Cleans and tokenizes the text
- Removes common English stopwords
- Counts and visualizes the most frequent words
- Saves results as CSV and PNG

## Requirements
- Python 3.x
- requests
- nltk
- pandas
- matplotlib
- seaborn

Install requirements with:
```bash
pip install requests nltk pandas matplotlib seaborn
```

Usage

1. Run the main script:

```
python moby_dick_word_frequency.py
```

1. The script will:
2. Download and process the text
3. Save `moby_dick_top100_words.csv`
4. Save and display `moby_dick_top30_words.png`

Output

- **moby_dick_top100_words.csv**: Top 100 most frequent words and their counts
- **moby_dick_top30_words.png**: Bar chart of the top 30 words

License

This project is licensed under the MIT License. ```
```