Image from: https://countryhouselibrary.co.uk/products/kenilworth-by-sir-walter-scott-readers-library

Detecting and Reporting the Polarity of Sentiments Associated with Characters of Kenilworth by Sir Walter Scott

Module tutor: Andreas Vlachidis

Natural Language Processing and Text Analysis (INST0073)

SRN: 23223014

Introduction

This report presents the results of a text mining project focusing on *Kenilworth*, a novel by Sir Walter Scott which is free to access on Project Gutenberg. The primary goal is to identify up to two key characters from the text and analyse the sentiment of the discussions surrounding these characters, namely Tressillian and Queen Elizabeth, which are then visualized via two respective plots for each character. Through a systematic approach using various natural language processing (NLP) tools, this report will demonstrate how characters are portrayed and how their narrative contexts influence the sentiment of textual mentions.

Part 1: Design of the Pipeline

Framework and Tools

For this project, the chosen computational environment supported an integrated approach to natural language processing, leveraging widely recognized libraries and tools for a robust text analysis. The primary tools and libraries used include:

- **SpaCy**: A powerful NLP library used for tasks such as tokenization, part-of-speech tagging, and named entity recognition.
- **NLTK**: Complemented SpaCy, especially with its Sentiment Intensity Analyzer for sentiment analysis, and provided additional resources like tokenizers and lexicons.
- **TextBlob**: Another NLP library used for easy and quick sentiment analysis, offering a straightforward API to obtain polarity scores.
- **Matplotlib**: Employed for creating visualizations to present the analysis results effectively.

Stages of the Pipeline

1. **Data Acquisition and Preprocessing**

The text of *Kenilworth* was loaded from a local storage path to ensure reliable access during processing. The file was read into Python using standard file handling methods, with error handling to manage potential issues such as file not being found or other IO errors. This stage is crucial as it ensures the text data is correctly and efficiently loaded into the system for further processing.

- **Input**: File path.
- **Output**: Text data loaded into memory.

2. **Text Preprocessing and Environment Setup**

Before processing, necessary libraries were setup and configured. This included loading SpaCy's en_core_web_sm model, suitable for recognizing entities in English text, and setting the max_length property of the SpaCy object to accommodate the large size of the text, ensuring the entire document could be processed without truncation.

- **Input**: Raw text data.
- **Output**: Prepared and configured NLP environment.

3. Named Entity Recognition (NER)

SpaCy was used to process the loaded text to detect named entities, specifically focusing on personal names, which represent potential characters in the novel. This step is critical for identifying which characters will be analyzed for sentiment.

- **Input**: Clean text data.
- **Output**: Set of character names identified as named entities.

4. Character Frequency Analysis

Using the characters identified by SpaCy, a frequency count was conducted to determine how often each character was mentioned. This helped in pinpointing the key characters for further sentiment analysis based on their prominence in the text.

- **Input**: List of named entities recognized as characters.
- **Output**: Frequency distribution of character mentions.

5. Data Visualization

The frequency data was then visualized using Matplotlib to create a horizontal bar chart showing the top 10 characters by mention frequency. This visualisation aids in quickly identifying the main characters and understanding their importance in the narrative based on how frequently they are mentioned.

- **Input**: Character frequency data.
- **Output**: Bar chart visualizing the top characters.

Contribution to the Pipeline

Each stage contributed to the pipeline by preparing the data sequentially for the next stage, ensuring a smooth transition from raw text to insightful visualizations. The use of pre-configured NLP tools allowed for efficient processing, while the visualizations provided clear, immediate insights into the data extracted which set the next stage of the assignment which is the Sentiment Analysis. I then went on to analyse the sentiments by the most frequently mentioned characters in the text, Tressillian and Queen Elizabeth. This detailed walkthrough of the pipeline showcases the sequential processing and contributions of each stage but also highlights the practical application of NLP tools in literary analysis.

Part 2: Findings and Reflections

NER Output and Character Selection

The Named Entity Recognition (NER) process successfully identified a wide range of character mentions within *Kenilworth*. Two key characters, Tressilian and Queen Elizabeth, were chosen for detailed sentiment analysis due to their frequent mentions and pivotal roles in the narrative. To handle variations in how these characters were referred to in the text, a set of name variants were created:

- **Tressilian Variants:** Included 'Edmund', 'Edmund Tressilian', 'Master Tressilian', and several others that appeared with prefixes or in different contexts.
- **Queen Elizabeth Variants:** Covered names from 'Queen', 'Lady Elizabeth', to 'Queen Elizabeth', accommodating both formal titles and informal mentions.

The aggregate counts of these variants provided a comprehensive frequency measure, ensuring that all relevant mentions were accounted for in subsequent analyses.

Sentiment Analysis Results

Sentiment analysis was conducted using TextBlob, which provided polarity scores for sentences containing mentions of the selected characters. This analysis helped to assess the overall sentiment (positive, negative, or neutral) towards these characters across different parts of the text.

- **Tressilian:** Exhibited an average sentiment polarity score that was predominantly neutral to positive. This reflects his portrayal as an active character within the narrative.

- **Queen Elizabeth:** Showed a more complex sentiment distribution, with scores ranging from negative to neutral.

Data Visualizations

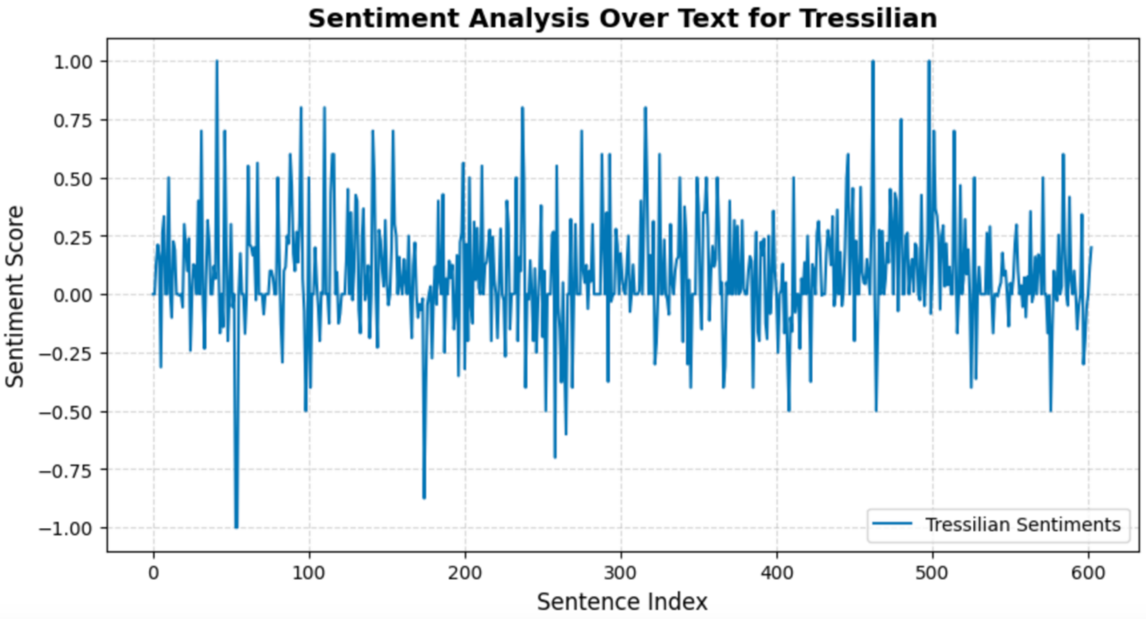
To visually represent the sentiment trends, line plots were created for each character. These plots displayed the sentiment polarity scores across the narrative, reflecting the changes in the narrative portrayal of the characters based on sentiment analysis.

- **Tressilian's Sentiment Plot:**

  Variability: The sentiment scores for Tressilian show significant variability, with values ranging from -1 to 1. This wide range indicates that the text associated with Tressilian includes both highly positive and highly negative sentiments.

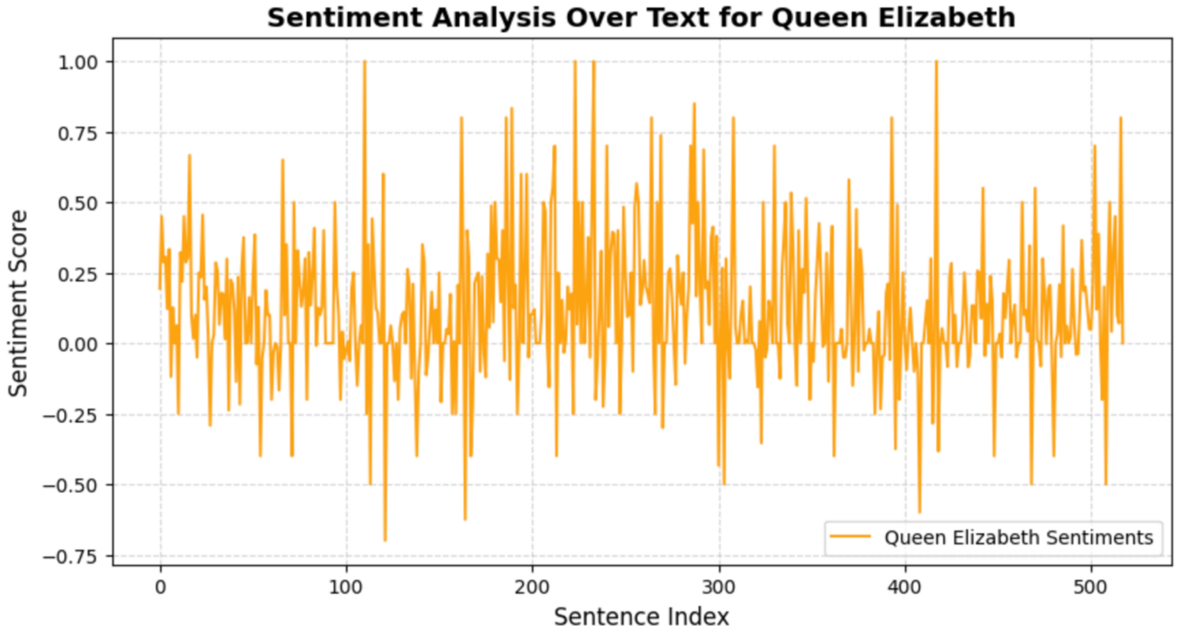
  Overall Trend: The plot does not show a clear long-term trend towards either positivity or negativity, which suggests that Tressilian's character may be involved in a variety of situations- some seen as positive, others as negative.

  Volatility: The sentiment seems quite volatile, with frequent sharp peaks and troughs. This could reflect a dynamic portrayal of Tressilian in the narrative, possibly indicating fluctuations in maybe his conflicts, or complex interactions with other characters.



Queen Elizabeth's Sentiment Plot:

> **Range and Consistency**: The sentiment scores for Queen Elizabeth are also varied, oscillating between -0.75 and 0.75, but they do not reach the extremes seen in Tressilian's plot. This might indicate that the narrative surrounding Queen Elizabeth, while still mixed, is less intense in its emotional extremes compared to Tressilian.
>
> **Fluctuations**: Like Tressilian, Queen Elizabeth's sentiment plot shows considerable fluctuation. However, the changes here are more frequent but less sharp, suggesting a constant shifting in how her actions or circumstances are viewed within the narrative.
>
> **General Observation**: The sentiment associated with Queen Elizabeth appears to be slightly more stable compared to Tressilian, albeit still showing no clear trend towards positivity or negativity over time.



Implications:

Character Dynamics: Both characters show complex sentiment dynamics, which implies that they are central to the narrative and likely involved in a range of emotionally charged situations.

Narrative Influence: The fluctuations and range of sentiments likely reflect key plot developments and character interactions. The absence of a clear positive or negative trend might suggest a balanced, nuanced portrayal of both characters, where both favorable and unfavorable traits or actions are depicted.

Challenges and Limitations

Several challenges were encountered during the project:

- **Large Text Handling:** Adjustments needed to be made to SpaCy's max_length setting to process the entire text, which presented initial hurdles in managing memory and performance.
- **Name Variations:** Capturing all relevant variants of the character names required meticulous configuration and could potentially miss less obvious references.
- **Sentiment Analysis Nuances:** The sentiment analysis tools, while effective, sometimes struggled with the subtleties of historical language and context-specific meanings, which could affect the accuracy of the sentiment scores.

These limitations highlight areas for potential improvement in future projects, such as the implementation of more sophisticated context-aware sentiment analysis techniques or enhanced entity recognition models that can better handle old literary texts and varied naming conventions.