RISE RESEARCH INNOVATION SCHOLARSHIP ENTREPRENEURSHIP

Mentor Name. : John Rachlin
Subject/Category :
Class : DS3500
Group # : 44

Northeastern University

Mark Twain Analysis

Yihan Luo, Yiming Yuan

Background, Motivation, and Goals

Mark Twain is celebrated as one of America's greatest humorists and writers. His works span a wide array of genres and themes, reflecting societal norms, criticisms, and humor of the late 19th and early 20th centuries. By analyzing Twain's works from different periods, we aim to shed light on how his writing style, thematic concerns, and language usage developed in response to his maturing worldview, personal experiences, and the changing times in which he lived.

Process and Methods

1.Data Collection: We selected five texts:
"The Innocents Abroad" (Early Career, 1869)
"Adventures of Huckleberry Finn" (Mid Career, 1884)
"A Connecticut Yankee in King Arthur's Court" (1889)
"£1,000,000 Bank-Note" (Late Career, 1893)
"The Tragedy" (Late Career, 1894)

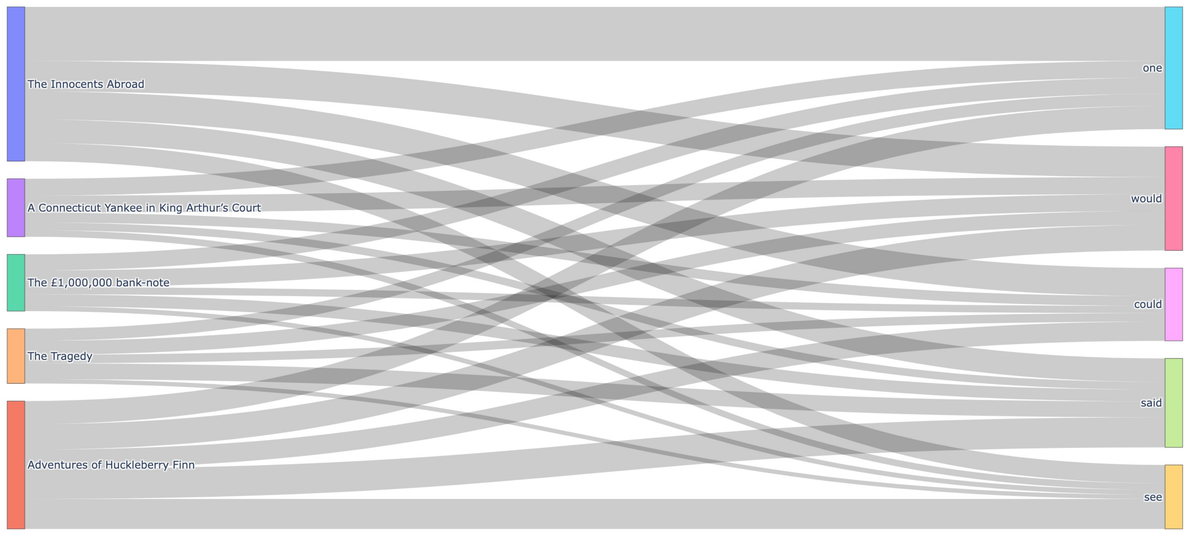
2.Data Preprocessing: Using a text analysis library, we preprocessed the texts by converting to lowercase and removing punctuation to standardize the data for comparison.

3.Stop Words Removal: Employ NLTK's comprehensive list of English stop words to filter out common words that offer little analytical value.

4.Analysis: We performed three main analyses:
Word Frequency: Determined the most common words across the different texts by Sankey diagram.
Type-Token Ratio and Flesch Reading Ease: Analyzed the lexical diversity and reading difficulty of each text.
Cosine Similarity: Computed the cosine similarity between the texts to understand the variance in word usage over time.

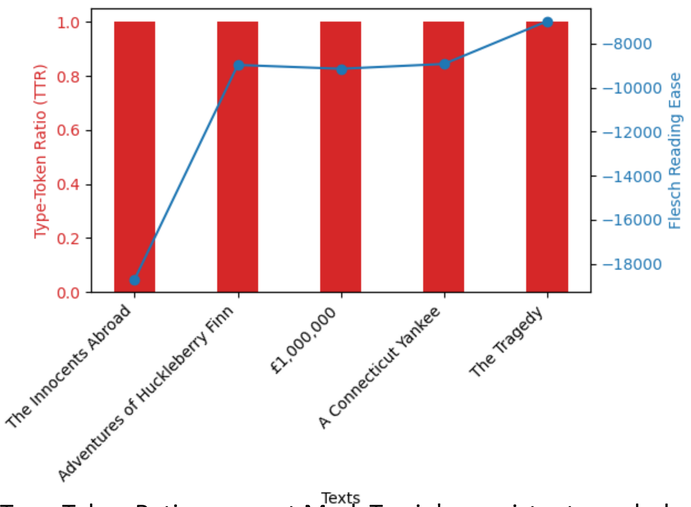
Findings and Products

Figure 1: Word Count Sankey Diagram: Tracing the Literary Currents of Mark Twain



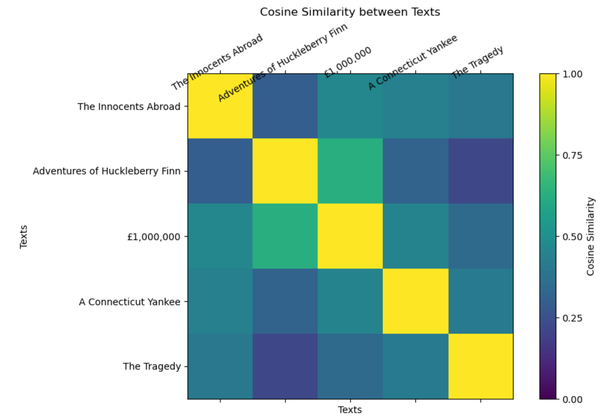
Analysis:
The Word Count Sankey Diagram: Tracing the Literary Currents of Mark Twain reveals the most prevalent words in Twain's narrative arsenal, underscoring a consistent use of certain terms despite the thematic evolution over time. It highlights his stylistic constancy and the shifting focus of his lexicon from his early to late works.

Figure 2: Type-Token Ratio and Flesch Reading Ease Chart



The Type-Token Ratios suggest Mark Twain's consistent vocabulary diversity, with "The Innocents Abroad" standing out for its richness. Contrastingly, "The Tragedy" presents as his most complex text, while "Adventures of Huckleberry Finn" is marked by its readability.

Figure 3: Cosine Similarity Heatmap



The Cosine Similarity Heatmap showcases the variance and parallels in word usage across Twain's selected texts, with early works sharing higher lexical similarities that diverge in his later writings. This indicates a possible maturation of style and a diversification of themes as his career progressed.

Conclusion and Next Steps

The text analysis of Mark Twain's selected works suggests a discernible evolution in his writing style and choice of vocabulary. Early works showcase a more adventurous and descriptive tone, while his mid to late works exhibit a richer vocabulary and more complex sentence structures. The similarity analysis further corroborates that Twain's literary style became more distinctive and nuanced as he matured as a writer.