Comparative Textual Analysis of Content Quality and Sentiment on 'Wikipedia.com' vs 'Encyclopedia.com' Articles

Introduction:

In this project, we aim to compare the content quality and sentiment of article on eleven selected topics from both Wikipedia and Encyclopedia. We employ web scraping techniques using Python to extract text data from these sources, followed by text cleaning and feature extraction. The extracted features include word counts, sentence counts, complex word counts, syllable counts, and readability metrics such as the FOG index. Additionally, sentiment analysis is performed to calculate positive score, negative score, polarity score, and subjectivity score for each entry.

Objectives:

- Extract text data from Wikipedia.com and Encyclopedia.com on eleven chosen topics.
- Clean and preprocess the text data to prepare it for analysis.
- Extract various features including word counts, sentence counts, complexity metrics, fog index and sentiment scores.
- Create a comparative analysis of content quality and sentiment between Wikipedia and Encyclopedia articles.
- Prepare a dataframe using panda library which holds all the vectorized data about each and every article
- Visualize the results using various visualization techniques such as bar charts, line charts, box-plot and scatter plots.
- Topics used for this project are: 'Space', 'Globalization', 'Global Warming', 'Italian Cuisine', 'The Mayans', 'Health and wellness', 'Real Estate', 'Natural Science', 'Spirituality,' 'Artificial Intelligence' & 'Education Reform' (I have used same topics to extract the text data from both Wikipedia and Encyclopedia for our analysis)

Methods:

- Web scraping: Utilize Python libraries such as Newspaper3k - Article and requests to scrape text data from Wikipedia and Encyclopedia articles.
- Text cleaning: Remove HTML tags, punctuation, stopwords, and perform lemmatization to preprocess the text data by using functions from NLTK library.
- Feature extraction: Calculate word counts, sentence counts, complexity metrics (e.g., FOG index), and sentiment scores using natural language processing techniques and various individual calculations for each vector.
- Data visualization: Construct pandas dataframes using panda library with the extracted features and utilize matplotlib library to create visualizations for comparative analysis with different chat and plots.

Findings:

- Comparison of content quality metrics (e.g., word counts, sentence counts) between Wikipedia and Encyclopedia articles.
- Analysis of readability metrics (e.g., FOG index) to assess the complexity of content on both platforms.
- Evaluation of sentiment scores to understand the overall tone and subjectivity of each article on the selected topics.

Implications:

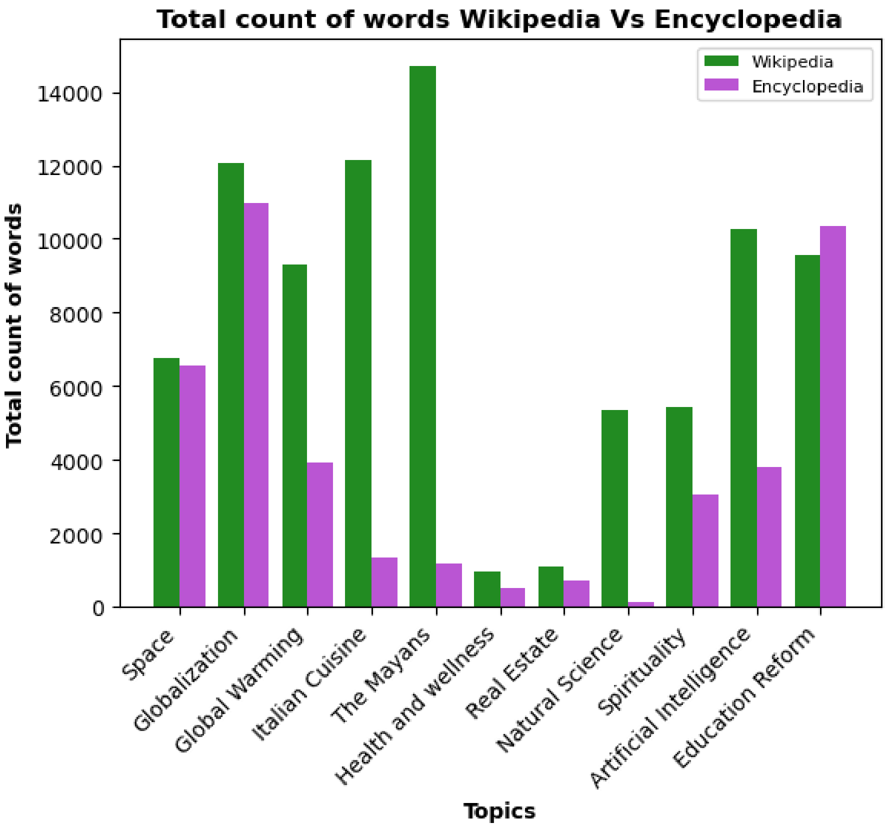
- Insights into the differences in content quality and sentiment between Wikipedia and Encyclopedia articles.
- Understanding how different sources present information on similar topics can aid readers in discerning reliable sources.
- Potential applications in information retrieval systems and content recommendation algorithms.

Limitations:

- Dependency on the accuracy of web scraping techniques for data extraction.
- Variability in content quality and style across different Wikipedia and Encyclopedia articles.
- The scope of sentiment analysis may be limited to the availability of labeled data or lexicons.

Visualization:

1. Bar Chart: Total Word Count Comparison¶



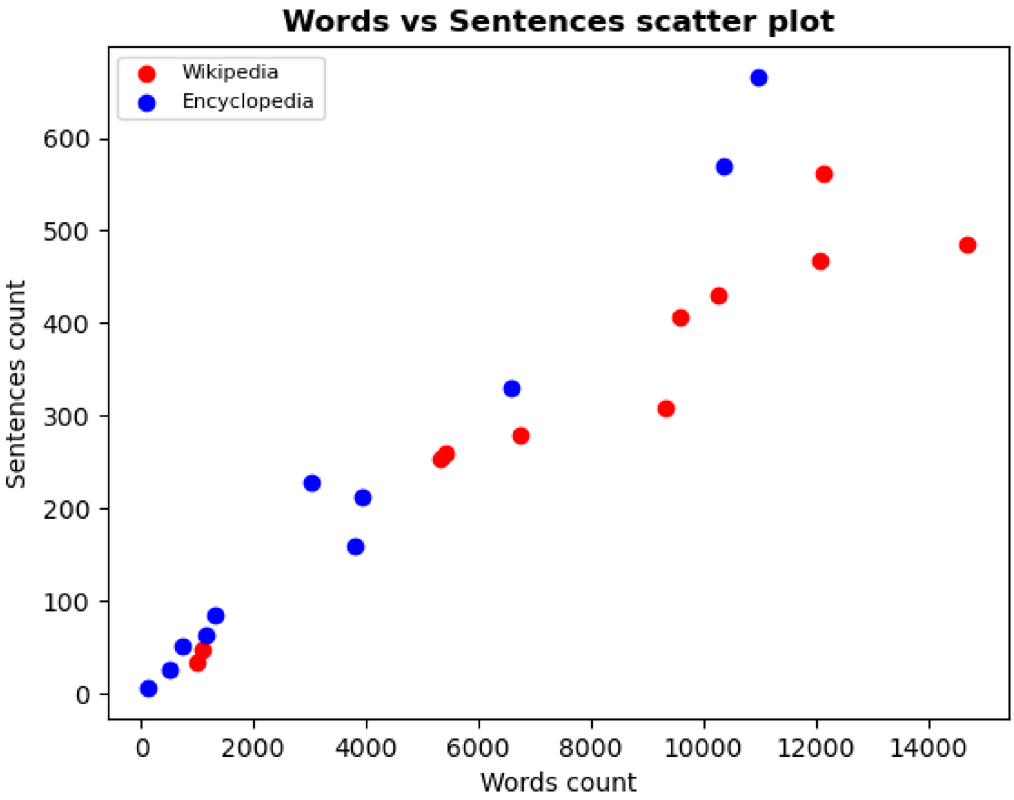
This bar chart displays the total word count for each topic extracted from both Wikipedia and Encyclopedia articles. Each bar represents a different topic, with the height of the bar indicating the total word count.

Observations:

- The article about 'The Mayans' in Wikipedia has the highest word count among all topics analyzed, indicating a comprehensive coverage of the subject matter.
- In contrast, the article 'Globalization' in Encyclopedia exhibits the highest word count, suggesting detailed coverage of this topic in that source.
- 'Natural Science' in Encyclopedia has the lowest word count among all topics, implying a relatively concise treatment of this subject.
- Notably, most articles from Wikipedia have word counts exceeding 6,000, indicating extensive content and in-depth exploration of the topics.

Key Insights:

This visualization highlights the variations in word counts between Wikipedia and Encyclopedia article for different topics. It underscores the detailed coverage and extensive content typically found in Wikipedia articles, particularly evident in topics like 'The Mayans', while also showcasing specific topics where Encyclopedia article may provide more comprehensive information.



2. Scatter Plot: Word Count vs. Sentence Count

This scatter plot illustrates the relationship between word count (x-axis) and sentence count (y-axis) for articles from both ‘Wikipedia.com’ and ‘Encyclopedia.com’. Each data point represents an article, with blue dots representing Encyclopedia articles and red dots representing Wikipedia articles.

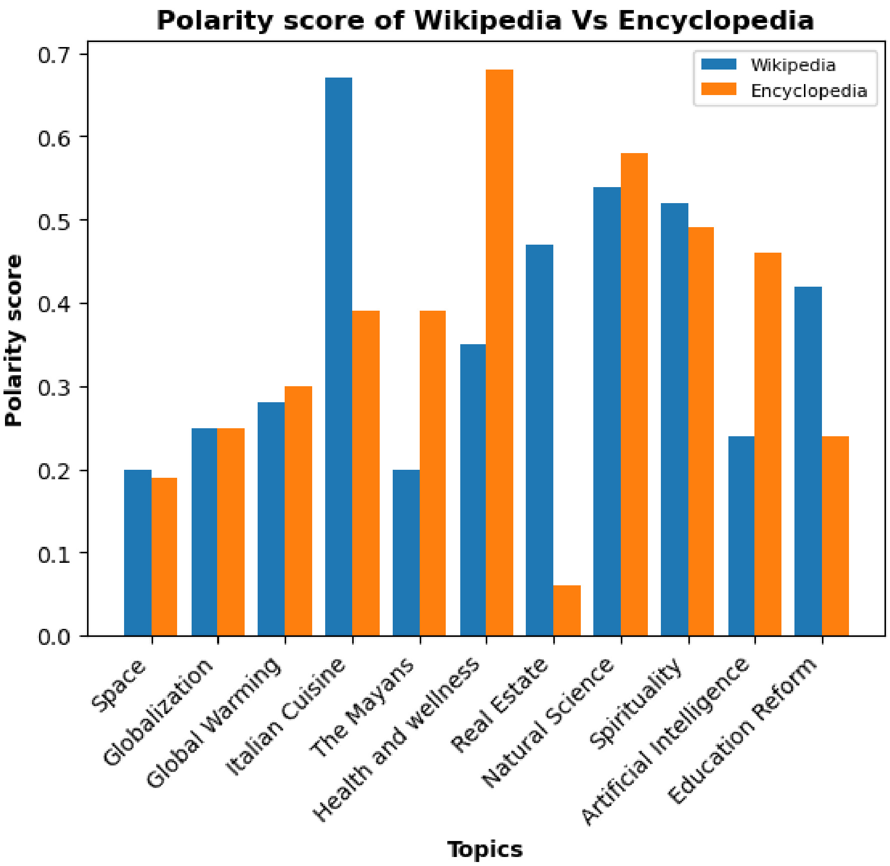
Observations:

- The scatter plot reveals a positive linear pattern, indicating a moderate positive correlation between word count and sentence count.
- Articles with higher word counts tend to have more sentences, suggesting a relationship between the length of the text and the complexity or depth of the content.
- While the pattern is not perfectly straight, it still demonstrates a noticeable trend where articles with higher word counts generally exhibit more sentences.
- The distribution of data points across the plot shows variability in both word count and sentence count for articles from both sources.

Key insight:

The observed positive linear pattern suggests that as the word count increases, the number of sentences tends to increase as well, indicating a potential correlation between content length and complexity.

3. Bar Plot: Polarity Score Comparison¶



In this bar plot, we compare the polarity scores of articles extracted from both Encyclopedia and Wikipedia. Polarity score measures the sentiment expressed in the text, indicating whether the overall sentiment of the content is positive, negative, or neutral. A polarity scores close to 1 indicates a highly positive sentiment, while a score close to -1 suggests a highly negative sentiment. A score around 0 indicates a neutral sentiment.

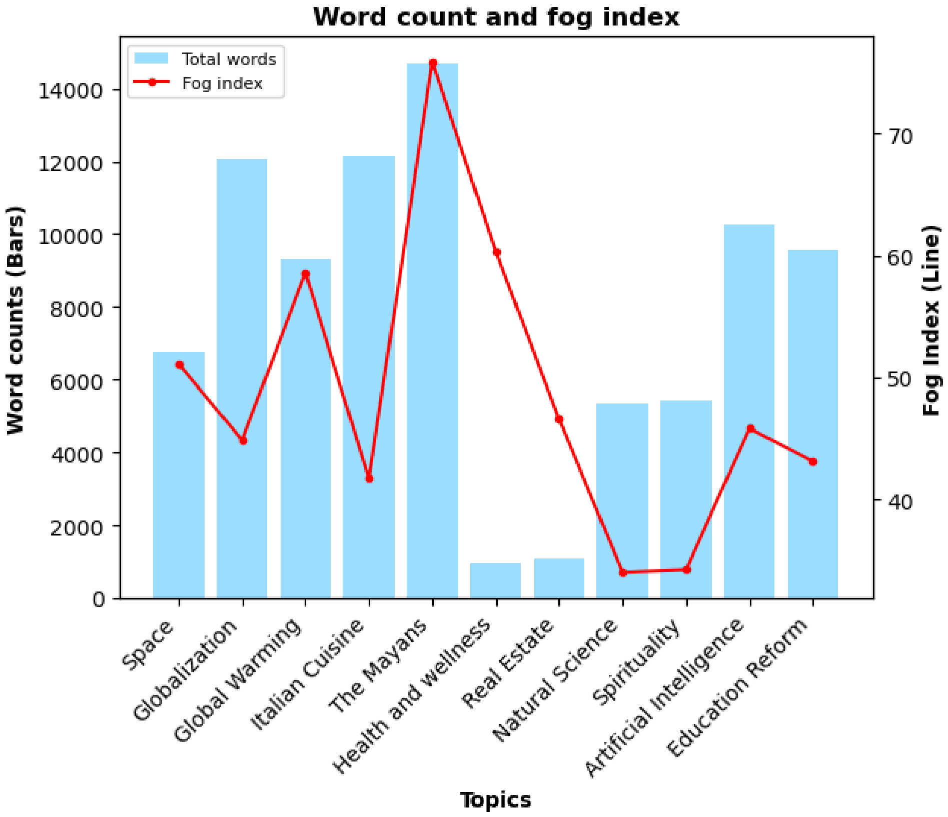
Observations:

- The polarity scores of articles from both Encyclopedia and Wikipedia are compared.
- Most articles have polarity scores exceeding 0.2, indicating a predominantly positive sentiment in the content.
- Specifically, the article on 'Health and Wellness' from Encyclopedia and 'Italian Cuisine' from Wikipedia have the highest polarity scores, suggesting overwhelmingly positive sentiment in these topics.
- Conversely, the article on 'Real Estate' from Encyclopedia exhibits the lowest polarity score of less than 0.01, indicating a relatively neutral or possibly mixed sentiment.

Key insight:

This predominance of positive polarity scores suggests that the majority of the content analyzed tends to convey positive sentiment. However, variations in polarity scores among different topics highlight the nuanced sentiment expressed in each article.

4. Pareto Chart: Word Count vs. FOG Index¶



In this Pareto chart, we compare the word count and FOG index of articles from Wikipedia. The bar chart represents the word count for each article, while the line chart depicts the corresponding FOG index.

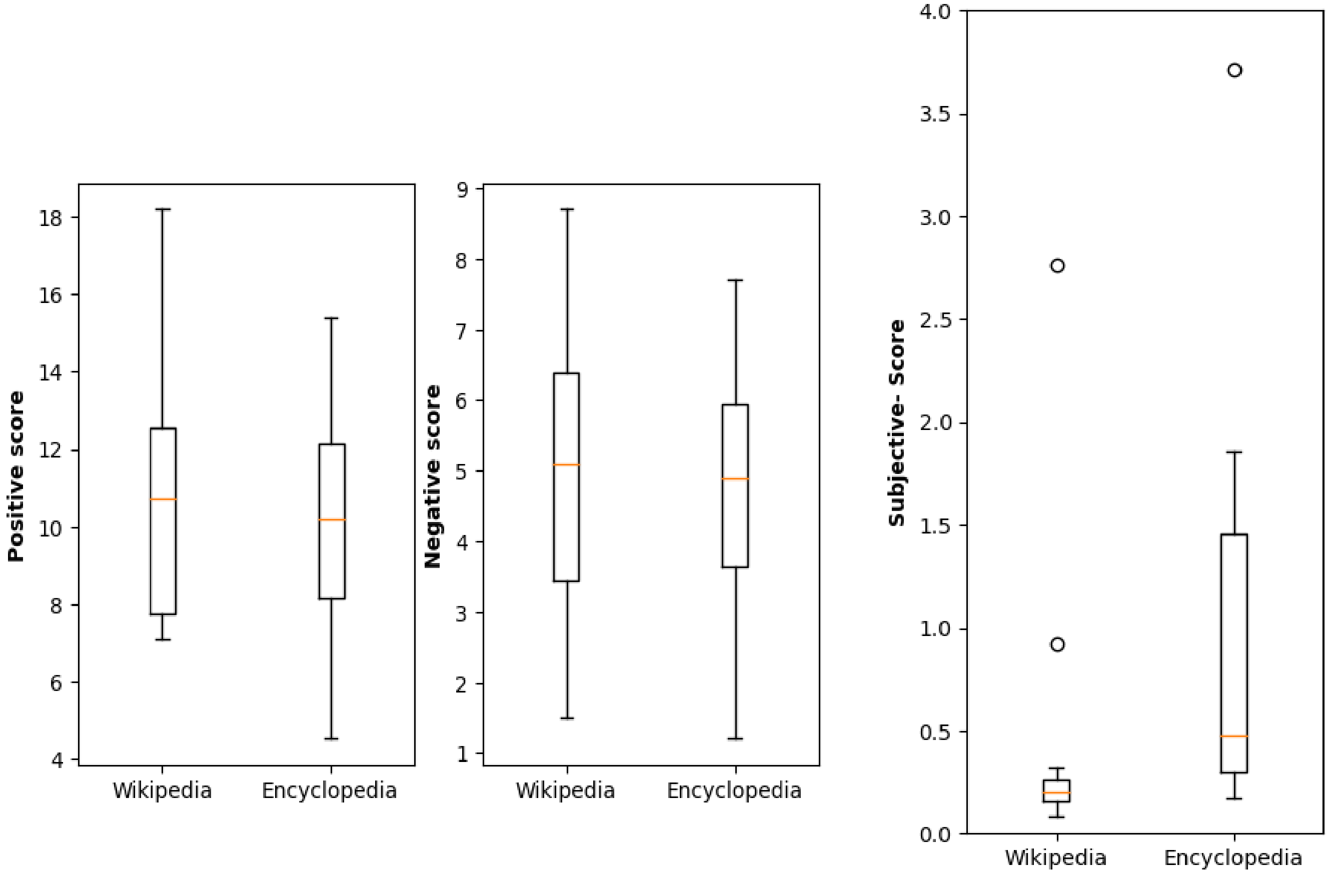
Observations:

- The Pareto chart provides a visual comparison of two key metrics: word count and FOG index, offering insights into the relationship between content length and readability complexity.
- While there is no clear linear relationship between word count and FOG index across all articles, some notable patterns emerge.
- 'The Mayans' article stands out with the highest word count and FOG index, indicating extensive content and highly complex language and sentence structures.
- Conversely, articles such as 'Health and Wellness' and 'Real Estate' have lower word counts but comparatively high FOG index scores, suggesting that despite shorter lengths, these articles exhibit complexity in language and sentence structures.
- The lack of a consistent trend between word count and FOG index underscores the nuanced nature of content complexity, which may be influenced by various factors beyond sheer word count.

Key insight:

While some articles demonstrate a correlation between word count and FOG index, others defy this pattern, suggesting that factors beyond word count alone influence readability complexity. Understanding these nuances can aid in assessing the comprehensiveness and accessibility of textual content.

5. Box Plots: Positive Score, Negative Score, and Subjectivity Score Comparison¶



In these box plots, we compare the positive score, negative score, and subjectivity score of articles from both Encyclopedia and Wikipedia. Additionally, we'll delve into the concept of subjectivity score and its significance in textual analysis. Subjectivity score measures the degree of subjectivity or objectivity expressed in the text. A subjectivity scores close to 0 suggests a more objective presentation of information, while a score closer to 1 indicates a more subjective or opinionated tone. In the context of this analysis, subjectivity score helps discern the nature of content presentation. The box plots for positive score and negative score show the distribution of scores for articles from Encyclopedia and Wikipedia, highlighting any differences in sentiment between the two sources.

Observations:

- Wikipedia articles tend to have mostly positive sentiment, with low outliers indicating occasional negative sentiment. Conversely, Encyclopedia articles exhibit similar distributions for positive and negative scores, with comparable means.
- While positive and negative scores exhibit similar distributions between Encyclopedia and Wikipedia, subjectivity score reveals contrasting patterns.
- Wikipedia articles tend to have lower subjectivity scores, with the majority of scores falling within the range of 0 to 0.5, indicating a more objective presentation of information.
- In contrast, Encyclopedia articles display higher subjectivity scores, with a wider range spanning from 0.2 to 2.0, suggesting a broader spectrum of subjectivity and potentially more opinionated content.
- The presence of outliers in subjectivity score for both sources signify instances where articles deviate significantly from the typical subjectivity level observed within their respective datasets.

Key insight:

While both sources exhibit similar sentiment distributions, their subjectivity scores diverge, reflecting differences in editorial style and content presentation. Understanding these nuances aids in evaluating the reliability and objectivity of information provided by different sources.

Conclusion:

The comparative analysis of Wikipedia and Encyclopedia articles sheds light on the distinct nature and style of content presentation offered by these two sources. The examination of various metrics, including word count, readability, sentiment, and subjectivity, reveals nuanced differences in their approach to conveying information on common topics.

- **Content Length and Readability:** Wikipedia articles generally exhibit higher word counts and FOG index scores compared to Encyclopedia entries. This indicates that Wikipedia tends to provide more extensive and detailed coverage of topics, often with complex language and sentence structures.
- **Sentiment Analysis:** Wikipedia articles predominantly convey positive sentiment, with occasional instances of negative sentiment. In contrast, Encyclopedia entries demonstrate a balanced distribution of positive and negative sentiment. This suggests that Wikipedia articles may lean towards a more optimistic portrayal of topics, while Encyclopedia articles maintain a relatively neutral stance.
- **Subjectivity and Editorial Style:** Wikipedia articles tend to present information in a more objective manner, as evidenced by lower subjectivity scores. Conversely, Encyclopedia entries display higher subjectivity scores, indicating a broader spectrum of subjective viewpoints and potentially more opinionated content.
- Overall, the analysis underscores the distinct characteristics of Wikipedia and Encyclopedia articles in terms of content length, readability, sentiment, and subjectivity. While Wikipedia offers comprehensive coverage with a tendency towards positive sentiment and objective presentation, Encyclopedia provides concise information with a balanced sentiment distribution and a wider range of subjectivity. Understanding these differences is essential for readers seeking reliable and comprehensive information from diverse sources.