Advanced Data Analysis And Machine Learning Sequential And Data Modeling

November 11, 2024

Bright Wiredu Nuakoh

Title: Pre-process The The black ship (with other allegories and parables by Elizabeth Rundle Charles) Text To Predict The Next Character In A Sequence

In this task, our aim is to prepare the Black ship text to train a character-level language model to predict the next character in a sequence of the text. The Black Ship: With Other Allegories and Parables by Elizabeth Rundle Charles is a collection of thought-provoking allegories and parables. Each story offers moral and spiritual lessons, exploring themes of faith, human struggles, and redemption. Through symbolic narratives, the book aims to inspire introspection and convey timeless truths. Using Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), or Transformers to learn the sequential dependency between characters will enable us to perform tasks such summarizing the stories in the Black ship text and taking key notes in its content. For us to carry out such task, we have to pre-processing the book to make it feasible to feed to our chosen model.

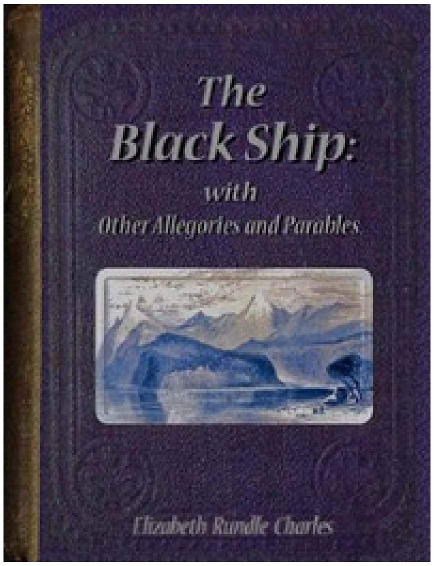

Figure 1: The Black Ship book, Sources;https://www.gutenberg.org/ebooks/74714

Data Description

The text comprises 3,201 lines, encapsulating a series of interconnected stories. It includes special characters such as *, ., and –, which require careful handling during preprocessing. Upon tokenizing The Black Ship, a total of 33,481 tokens were generated, representing 4,626 unique words, with ”The” being the most frequent, appearing 138 times. After filtering out excessively short and long words, the vocabulary size decreased to 3,079, with the most frequent word now appearing 112 times.

The token distribution is highly skewed, with a small subset of words accounting for a significant proportion of the text. This highlights the presence of common stop words, which can be further analyzed or removed to improve downstream tasks. Moreover, the diversity of the vocabulary indicates a rich narrative structure, suitable for our text generation applications.

Data Pre-processing

Our preprocessing pipeline applies several steps to clean and prepare the text data for analysis. First, part-of-speech details are added to assist in the removal of stopwords, ensuring that only contextually irrelevant words are eliminated. Next, common stopwords such as ”and” and ”the” are removed to focus on meaningful content. The text is then normalized through lemmatization, converting words to their base forms (e.g., ”running” to ”run”) to reduce inflectional variations and enhance consistency. Punctuation marks are erased to simplify the text structure, and length-based filtering is applied to remove words shorter than 2 characters or longer than 15 characters, minimizing noise and outliers. Finally, a bag-of-words model is created, representing the frequency of unique terms in the cleaned text, which is essential for further analysis and model training. This preprocessing ensures a clean and consistent dataset, ready for natural language processing tasks.

Statistics on the pre-processing

We initially use word cloud visualization compares the frequency of words in the raw data versus the cleaned data. From Figure 2, the common stopwords such as ”the,” ”and,” ”of,” ”to,” ”in,” and ”was” dominate the word cloud. These words provide little semantic content but are highly frequent in unprocessed text.

Punctuation marks (e.g., ;, :) and other non-alphanumeric symbols are visible, indicating the presence of noise in the raw dataset as discussed previously. The cleaning process successfully removed common stopwords, punctuation, and other irrelevant tokens, improving the dataset's focus on semantically important words. From Figure 2, the cleaned word cloud highlights core themes and keywords related to the narrative or topics discussed in the text, such as ”come” ”work,” ”project,” ”gutenberg,” and ”make”, suggest a reduction in redundancy and improved focus on meaningful words after cleaning.

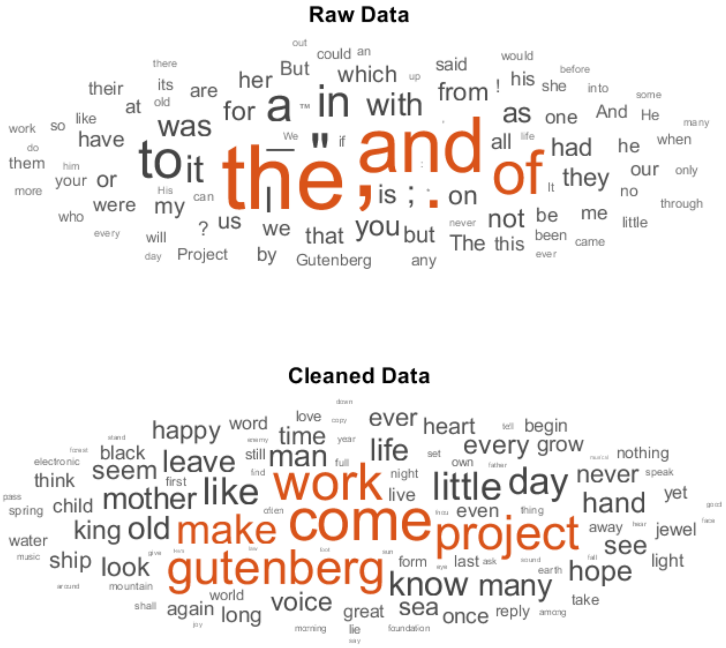

Figure 2: Word cloud visualization of raw data against cleaned data (cleared vocabulary)

After cleaning, the proportion of stopwords approaches zero, improving the dataset's focus on content words.

N-gram analysis

N-Gram analysis is a fundamental technique in natural language processing (NLP) used to model sequences of words or characters. This analysis captures the frequency and co-occurrence patterns of words, providing insights into the structure and context of a text in the Black Ship book. By analyzing n-grams, we can identify common phrases, predict the next word in a sequence, and improve tasks our goal.

From Figure 3, the bar chart (top) displays the frequency of the top 30 most common unigrams (1-grams) in the cleaned text. The line plot (bottom) shows a descending frequency trend for these unigrams, highlighting the dominance of a few high-frequency words like "come" and "work," while other words appear less frequently. This illustrates a typical long-tail distribution in natural language text.

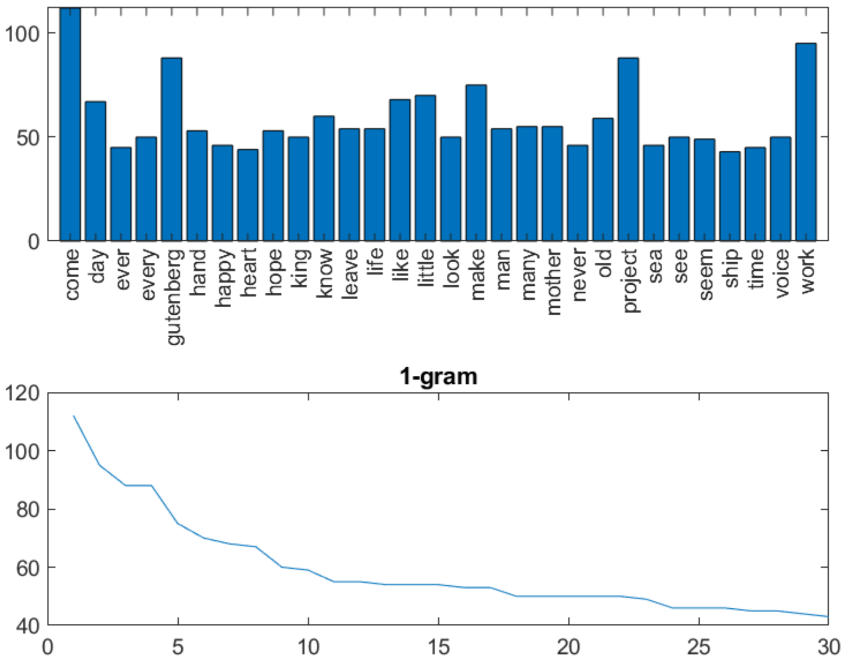

Figure 3: Frequency distribution of the top 30 unigrams in the cleaned text. The bar chart (top) highlights individual word frequencies, while the line plot (bottom) demonstrates the declining trend, typical of natural language text, where a few words dominate the dataset.

2N Gram and 3N Gram

The word cloud visualizes the most frequent bigrams (word pairs) in the text, with larger font sizes indicating higher frequencies. We observe prominent phrases such as ”project gutenberg,” ”electronic work,” and ”black ship” stand out, suggesting their importance within the text. This representation helps identify recurring themes and associations, providing insights into the text’s structure and content. Bigrams like ”archive foundation” and ”united states” highlight specific topics or organizational mentions, while others such as ”training begin” and ”literary archive” suggest broader contextual themes.

The trigram word cloud below also displays the most frequent three-word phrases in the text. Larger phrases like ”project gutenberg electronic,” ”gutenberg electronic work,” and ”literary archive foundation” suggest recurring themes related to Project Gutenberg and literary content. The visualization provides insights into specific topics and relationships between words, such as legal aspects (”protect copyright law”) and thematic expressions (”make things musical”). These trigrams offer a deeper understanding of the text’s context and highlight significant thematic patterns.

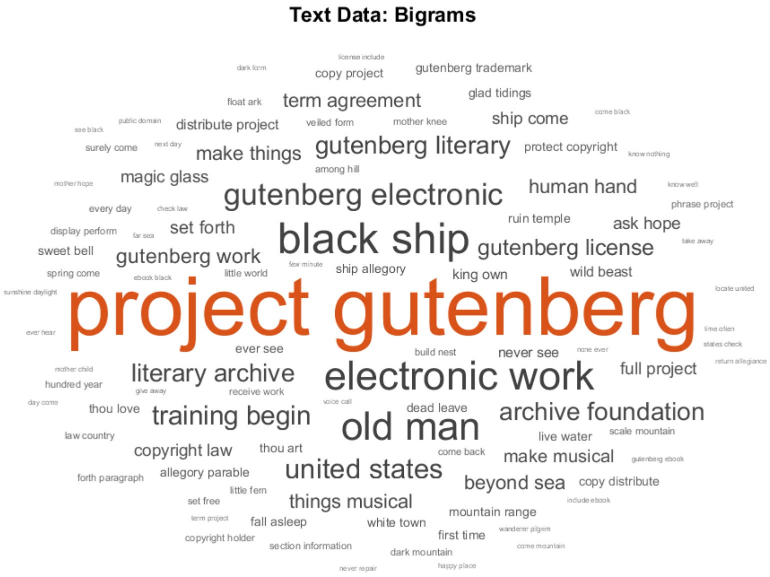

Figure 4: Bigram Word Cloud: Key Phrase Frequencies and Thematic Associations in Text Black Ship book.

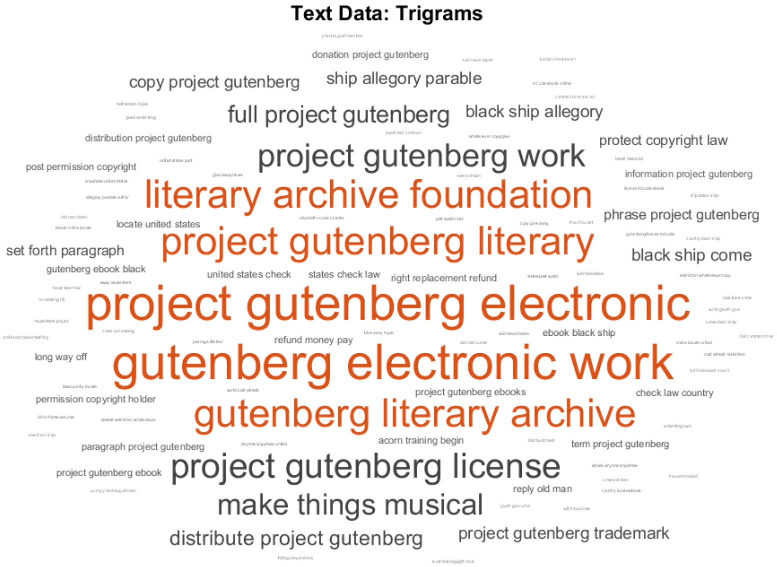

Figure 5: Trigram Word Cloud: Key Three-Word Phrases Highlighting Textual Patterns and Themes.