Hemingway and Carroll

1a.

Ernest Hemingway was an American contemporary 20th- century writer and a Pulitzer Prize winner for his novel, The Old Man and the Sea. Hemingway’s writing style is known to be concise, factual, and unadorned style. His sentences are usually short and do not use a lot of adjectives in his work. Lewis Carroll was an English writer of children’s friction in the 19th century. Carroll’s writing style is known to be a nonsensical style with a whimsical way of using words. Unlike Hemingway’s style, Carroll was very playful with his writing through fantasy, which used a lot of adjectives or wordplay in his works. In Our Time by Ernest Hemingway and Alice’s Adventures in the Wonderland by Lewis Carroll are the novels selected for the purpose to compare the corpora by using corpus statistics with specifically on their word usages.

1b.

In Our Time and Alice’s Adventures in the Wonderland were downloaded from the Gutenberg.org website. In Our Time was downloaded as a PDF file, converted to text files, saved as UTF-8, and imported to the Jupyter. Whereas Alice’s Adventures in Wonderland was already a text file in the NLTK Gutenberg corpus and readily uploaded into the Jupyter for further processing for both text files. The texts were labeled as Hemingway and Carroll for In Our Time, and Alice’s Adventures in the Wonderland, respectively for ease of discussion in this assignment.

2a.

Natural Language Toolkit (NLTK) and Regular Expression (re) were imported for processing the text files in Jupyter. The texts were tokenized through the process of splitting sentences into individual words known as tokens. The count of tokens for Hemingway was 37392 and the count for Carroll was 33493. Both corpora were converted to lower letter cases for easier removal of stop words and stemming. NLTK stop words  and additional stop words specific for each corpus were incorporated to remove repetitive words from the corpus. Also, non-alphabetic characters were removed by using a regular expression pattern. No stemming or lemmatization was applied to evaluate because Hemmingway was known to use short, one, or two-syllable words, which might overly stem the word if Porter or Lancaster stemmer were applied. Also, no stemming was applied to Carroll to observe the whimsical way of how Carroll plays with words in his writing.

With the preliminarily processed corpora, Frequency Distribution (FreqDist) and Collocation from the NLTK were imported to perform statistics to compare both corpora for unigram by frequency, bigram by frequency, and by pointwise mutual information (PMI). The count of tokens for Carroll was 12492 and the count for Hemingway was 15669. After processing the corpora, 42% of Hemingway’s token and 37% of Carroll’s token remained for the next steps. The unigram analysis generated the top 50 words and normalized frequency for both corpora (Figure 1 & 2). The top 50 bigrams by frequency (Figure 3 & 4) and top 50 bigrams by Point Mutual Information (PMI) score with minimum frequency for the bigram set at or above 5 (Figure 5 & 6). This setting will filter words that appeared together at least five or more times in the corpus, which indicate those words have the likely probability of being together.

Text

Description automatically generated with medium confidence

Figure 2: Hemingway’s top 50 words by frequency (normalized by the length of the text)

Text

Description automatically generated

Figure 3: Carroll’s top 50 words by frequency (normalized by the length of the text)

Text

Description automatically generated

Figure 4: Hemmingway’s top 50 bigrams by frequency

Text

Description automatically generated

Figure 5: Carroll’s top 50 bigrams by frequency

Text

Description automatically generated

Figure 6: Hemingway’s top 50 bigrams by using frequently occurring words in mutual information

A picture containing text, person, document

Description automatically generated

Figure 7: Carroll’s top 50 bigrams by using frequently occurring words in mutual information

2b.

Initially,  was used for processing both corpora. Both original token counts were significantly reduced, which resulted in a less than 50 top bigrams by Point Mutual Information (PMI) score. The stop words (Figure 7) found in were words that should keep as part of the text analysis. The NLTK, was used to further process the corpora and fewer tokens were removed as observed in the counts (Figure 8). A better list of 50 top bigrams by Point Mutual Information (PMI) score was generated for the Hemingway. But Carroll’s bigram was only able to have a less than 50 top bigrams.

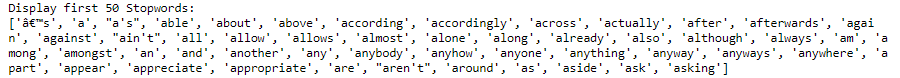
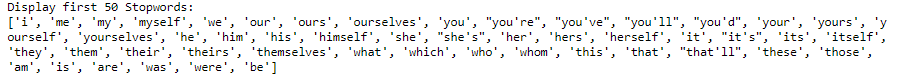


Figure 7: First 50 stop words from 

Figure 8: First 50 stop words from 

2c.

The top 50 bigrams by frequency are different from the top 50 bigrams scored by mutual information. Bigram by frequency shows the percentage occurrence of the bigram in the corpus. Whereas bigram in mutual information calculates the probability of two words occurring in sequence. Hemingway’s first bigram by frequency (’nick’, ‘said’) has 0.00227 but bigram by PMI has a low score of 4.572. Another example is the first bigram, (‘san’, ‘siro’) has a high PMI score of 12.868 but bigram by frequency is found outside of the top 50 bigrams by frequency is 0.0001337. Carroll’s first bigram by frequency (’said’,‘alice’ ) is 0.00343 but bigram by PMI has a low score of 4.395. Another example is the first bigram, (‘play’, ‘croquet’) has a high PMI score of 11.768 but bigram by frequency is 0.0001492. Therefore, the ranking of the bigram by frequency in both corpora does not share the same ranking as the bigram by mutual information. Additionally, bigrams by frequency (’nick’, ‘said’) and (’said’,‘alice’ ) show the high occurrence of those pair words in the corpus. Bigrams by mutual information (‘san’, ‘siro’) and (‘play’, ‘croquet’) show the high probability of those word pairs occurred as a joint instance that is when both words occur independently and most likely both words could be connected at the same time.

2d.

While performing initial unigram and bigram analysis, certain words such as ‘all, ‘right/’, and ‘that in Hemingway’s corpus (Figure 1a) and ‘well, ‘why, ‘of and ‘and in Carroll’s corpus (Figure 1b) were found not appropriate as part of the analysis. Additional stop words were prepared for each corpus to remove for the final analysis

1. Hemingway addstopwords



1. Carroll addstopwords



Figure 9: Hemingway(a) and Carroll(b) add stop words

3.

Did Hemingway use lesser adjective words than Carroll within word frequency and bigram?

Yes, Hemingway did use fewer adjective words in this corpus than Carroll. The word frequency (Figure 10) analysis observed 6 adjective words in Hemingway’s corpus and Carroll’s corpus has 10 adjective words. Additionally, the normalized frequency of most used adjective word, big, in Hemingway is 0.0063 and most used adjective word, little, in Carroll is 0.0102. The most used adjective word in Carroll’s corpus is higher than by Hemingway’s corpus. This confirmed that Hemingway had a tendency of using simple words and less adjectives in his writing style.

Chart, bar chart

Description automatically generated

Icon

Description automatically generated

Figure 10: Hemingway and Carroll’s Adjective in Top 50 Words by Frequency

In bigram analysis by PMI (Figure 11) further confirmed that Hemingway used of adjective words was lesser than Carroll. The analysis observed 2 adjective word pairs in Hemingway’s corpus and Carroll’s corpus has 17 adjective word pairs. Hemingway used simple adjective to descript an object or character as sweet fern or little man without using verbose adjective or wordiness to provide information. As for Carroll, did use adjective such as beautiful soup, poor little, or trembling voice to describe an object, a character, or a tone of voice to engage the readers.

Chart, bar chart

Description automatically generated

Icon

Description automatically generated

Figure 11: Hemingway and Carroll’s Adjective in Top 50 Bigrams by PMI