**Hemingway and Carroll Sentiment Analysis: Exploratory Analysis**

1. **Description of the cleaning and analysis processes**
2. Purpose:

In Our Time by Ernest Hemingway and Alice’s Adventures in the Wonderland by Lewis Carroll, the novels from homework one are further evaluated in homework 2. The purpose of homework 2 is to conduct exploratory analysis by extracting sentences that contain adjectives or adverb phrases to gain insights into how to approach the sentiment analysis of these novels. Both novels are labeled as Hemingway and Carroll for ease of discussion in this assignment.

1. Cleaning process:
2. Packages:

In Our Time and Alice’s Adventures in the Wonderland were transformed into text files in homework 1. Natural Language Toolkit (NLTK) and Regular Expression (re) were imported for processing the text files in Jupyter. From NLTK imported sent\_tokenize, nltk.RegexpParser, and FreqDist for the cleaning and analysis processes.

1. Sentence Tokenization:

Two-step tokenization was used on each text. The first step is to tokenize the sentences using  based on punctuation like “.”. The second step is using the nltk.word\_tokenize the text within the tokenized sentence,  .

Then, Stanford Part-of-Speech (POS) tagger POS was applied to assign part of speech to the tokens of each sentence, . The POS tagging was necessary to retrieve the adjective (JJ) and adverb (RB) tags for the next step. Both Hemingway and Carroll were labeled as “taggedtext” for the analysis processing.

1. Chunking

The chunking technique is applied to parse the tokenized sentence to extract the adjective and adverb phrases. This technique uses regular expression (re) to segments and labels multi-token sequences to detect adjective and adverb phrases. The regular expressions define how adjective phrases (ADJPH), and adverb phrases (ADVPH) are identified in the chunk as the following:



The ADJPH expression,  reads to find a phrase of (<>) adverb (RB.?) and adjective (JJ.?) with “.” as a wildcard for RBR, RBS, JJR, or JJS. Additionally, the “?” can be considered RB or JJ alone for both adverb and adjective.



The ADVPH expression reads to find a phrase of two consecutive adverbs (RB).

1. NLTK Parser

Then, nltk.RegexParser is imported to process each sentence for adjective or adverb phrase. ADJPH or ADVPH expression is inputted for chunk\_parser\_adj or chuck\_parser\_adv, respectively. Then, python code processed the taggedtext through the regex parser tree to extracted for adjph\_tags (Figure 1) or advph\_tags (Figure 2) for the analysis process.

Text

Description automatically generated

Figure 1: Adjective Phrase Parser

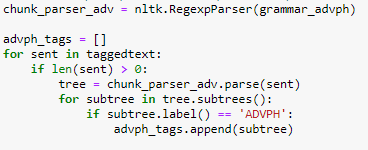


Figure 2: Adverb Phrase Parser

1. Analysis Process
2. Adjective and Adverb Phrases

The adjph\_tags and advph\_tags were input and coded for extracting the lists of adjective\_phrases (figure 3) and adverb\_phrases (figure 4). Then, statistical analysis can be performed to determine the length (len(advph\_tag) or len(advph\_tag)), and the frequency of the top 50 adjectives (figure 5 & 7) and adverb (figure 6 & 8) phrases of Carroll and Hemingway.

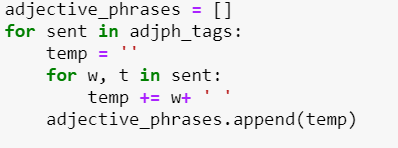


Figure 3: Adjective Phrase

Text

Description automatically generated with medium confidence

Figure 4: Adverb Phrase

A picture containing text

Description automatically generated

Figure 5: Carroll’s Top 50 Adjective Phrases

Map

Description automatically generated

Figure 6: Carroll’s Top 50 Adverb Phrases

A picture containing graphical user interface

Description automatically generated

Figure 7: Hemingway’s Top 50 Adjective Phrases

Text

Description automatically generated with low confidence

Figure 8: Hemingway’s Top 50 Adverb Phrases

1. Adjective and Adverb Tokens

Carroll and Hemingway’s tagged texts were input and coded for extracting the lists of adjective\_tokens (figure 9) and adverb tokens (figure 10). Then, statistical analysis can be performed to determine the length (len(advph\_tokens) or len(advph\_tokens)), and the frequency of the top 50 adjectives (figure 11 & 13) and adverb (figure 12 & 14) tokens of Carroll and Hemingway.

Graphical user interface, text, application

Description automatically generated

Figure 9: Adjective Tokens

Graphical user interface, text

Description automatically generated

Figure 10: Adverb Tokens

A picture containing table

Description automatically generated

Figure 11: Carroll’s Top 50 Adjective Tokens

A picture containing text

Description automatically generated

Figure 12: Carroll’s Top 50 Adjverb Tokens

Table

Description automatically generated with medium confidence

Figure 13: Hemingway’s Top 50 Adjective Tokens

A picture containing diagram

Description automatically generated

Figure 14: Hemingway’s Top 50 Adverb Tokens

1. Statistic

Basic statistic codes are applied to provide the total number of a corpus, sentence tokens, tokens, average length of sentence, and phrase (table 1). Hemingway’s statistic shows a higher corpus but a lower number of adjective and adverb phrases or tokens to Carroll’s statistic. This observation could indicate Hemingway might have a higher usage of nouns or verbs, which need to be further investigated and confirmed.

Table 1: Basic Statistic for Hemingway and Carroll Corpus.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Code | Hemingway | Carroll |
| # of Total Corpus | total\_corpus = sum(len(sent) for sent in textsplit) | 163807 | 141812 |
| # of Sentence | len(tokentext) | 2750 | 1625 |
| Average Length of Sentence | total\_corpus/ len(tokentext) | 59.2 | 87.3 |
| # of Adjective Phrase | len(adjph\_tags) | 160 | 222 |
| # of Adverb Phrase | len(advph\_tags) | 140 | 235 |
| # of Adjective Tokens | len(adjective\_tokens) | 1745 | 1488 |
| # of Adverb Tokens | len(adverb\_tokens) | 1895 | 2107 |
| # of Adjective Whole Sentence | len(adjph\_whole\_sentences) | 343 | 464 |
| # of Adverb Whole Sentence | len(advph\_whole\_sentences) | 291 | 484 |
| Average length of an adjective phrase sentence | total\_adjph\_sentences / len(adjph\_whole\_sentences) | 9.15 | 9.20 |

1. **Results of the analysis and interpretation**

Hemingway’s writing style is known to be concise, factual, and unadorned style. His sentences are usually short, which has 2750 sentences in the corpus. Carroll’s writing style is a nonsensical style with a whimsical way of using words and long sentences, which has 1625 sentences in the corpus and is lesser than Hemingway. Hemingway’s total corpus and sentence tokens are higher than Carroll’s corpus by 13% and Carroll’s sentence by 40%.

Compared to both sentence examples (Figure 15), Hemingway has 13 tokens, whereas Carroll has 60. With a comparable number of tokens in both corpora, Carroll’s example has showcased the tendency to write longer sentences than Hemingway. Also, this is further supported by the average length of sentence by statistic performed in Carroll is ~32% more than Hemingway.

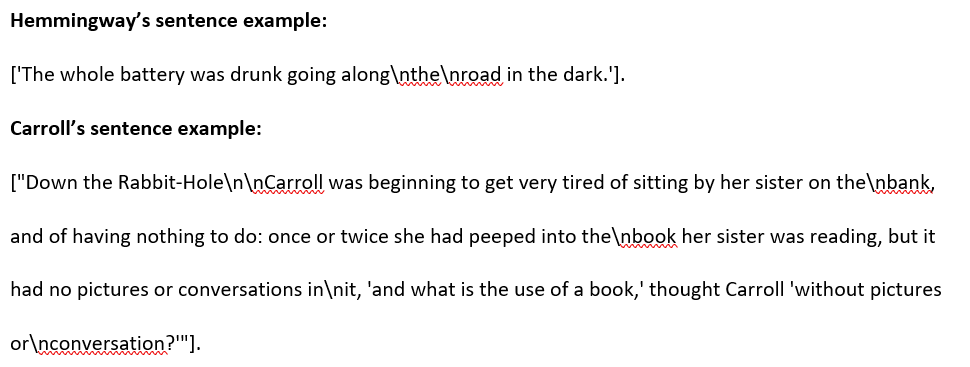


Figure 15: Sentence Example of Hemingway and Carroll

Adjective and adverb phrases are extracted from the Hemingway and Carroll corpora. Figure 16 shows Carroll has 27% more in adjective phrases and 40% more in adverb phrases than Hemingway. Adjective and adverb tokens are also extracted from the Hemingway and Carroll corpora. Figure 17 shows Carroll has 15% less in adjective tokens and 11% more in adverb tokens than Hemingway. Also, this indicated that Hemingway wrote fewer adjective phrases than Carroll, but he preferred a single word adjective in his sentence instead of a phrase or preferred verbs or nouns.

The combination of phrases and tokens for adjective and adverb in whole sentences (figure 18) consistently shows that Carroll has 35% higher in adjectives and 66% higher in adverbs than Hemingway. Also, this is supported by the average length of an adjective sentence in Carroll is 5% more than Hemingway, which is surprisingly not significantly more.

The conclusion from these results confirmed that Hemingway tended to use simple words and fewer adjectives in his writing style and adverbs. Also, these results have shown that Carroll crafted elaborate long sentences with lots of adjective and adverb phrases. Statistical analysis on both corpora has been demonstrated that Carroll has a higher usage of adjectives and adverbs than Hemingway. Lastly, Carroll used more adverbs than adjectives, which was an unexpected finding from this assignment.

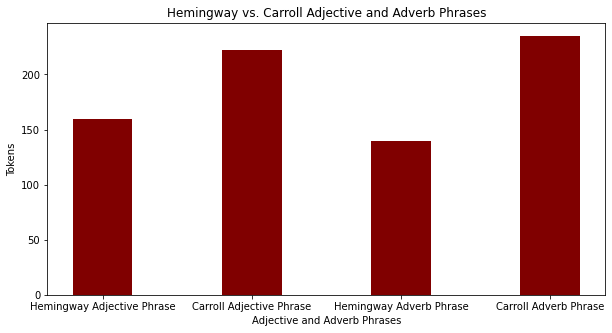


Figure 15: Hemingway vs. Carroll Adjective and Adverb Phrases

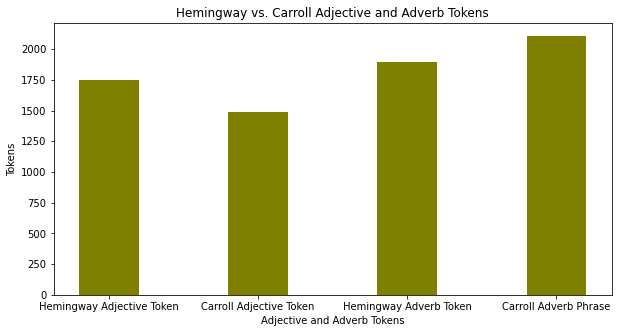


Figure 16: Hemingway vs. Carroll Adjective and Adverb Tokens

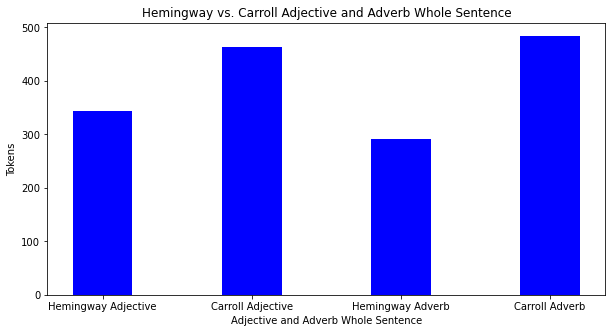


Figure 17: Hemingway vs. Carroll Adjective and Adverb Whole Sentence

1. **Thoughts on sentiment analysis for Hemingway and Carroll Corpora**

Hemingway and Carroll’s tokens should be further cleaned by removing certain words like “n’t” from the adverb\_phrase and adverb\_token lists that are not suitable for sentimental analysis. The generated list of adjective phrases, tokens, and combined phrases plus tokens and adverbs can be applied for sentiment analysis. The sentiment analysis can score if Hemingway or Carroll corpus is positive, negative, or neutral. Hemingway tends to be a serious and factual writer, which would be interesting to learn from the sentiment analysis if his corpus has more a negative or neutral score. Unlike Hemingway, Carroll tends to be whimsical and satirical, which could be more positive than negative. Overall, the sentimental analysis will provide further insights into the author’s writing style, emotions, and content of both corpora.