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 `nltk.corpus.stopwords.words('english')` 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.

```
nick      0.024606361955759545
said      0.022056479887805187
back      0.008478357875948238
man       0.007840887358959648
went      0.0071396697902722
water     0.0071396697902722
big       0.006310958118187034
old       0.006119716963090457
one       0.00592847580799388
would     0.005864728756295021
bill      0.005864728756295021
could     0.005800981704596163
get       0.005737234652897304
like      0.005737234652897304
right     0.005482246446101868
came      0.00535475234270415
looked    0.0050360170842098555
got       0.0049722700325109965
going     0.004781028877414419
go        0.004526040670618984
way       0.004079811308726971
around    0.004016064257028112
trout     0.004016064257028112
george    0.003952317205329254
good      0.0038248231019315355
river     0.0038248231019315355
know      0.0034423407917383822
fire      0.00337859374003952 3
see       0.003314846688340664
time      0.003251099636641805
along     0.0030598584815452285
two       0.0029961114298463695
put       0.0029961114298463695
come      0.0029961114298463695
made      0.0029323643781475105
sat       0.002868617326448652
away      0.002804870274749793
little    0.002804870274749793
took      0.00274112322305093 4
head      0.002677376171352075
felt      0.002677376171352075
line      0.002677376171352075
father    0.0025498820679543573
asked     0.0024861350162554982
looking          0.0024861350162554982
walked    0.0024223879645566392
hand      0.0024223879645566392
want      0.0024223879645566392
stream    0.00235864091285778
current          0.00235864091285778
```

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

```
said      0.036983669548511046
alice     0.03170028818443804
little    0.010246557796990073
one       0.00792507204610951
would     0.007204610951008645
know      0.007044508485430675
could     0.0068844060198527054
like      0.006804354787063721
went      0.006644252321485751
queen     0.006003842459173871
thought           0.005923791226384886
time      0.005443483829650976
see       0.005363432596861992
king      0.004963176432917067
began     0.004629715017611275
turtle    0.004629715017611275
hatter    0.004482869036183158
mock      0.004482869036183158
quite     0.004402817803394172
gryphon           0.004322766570605188
think     0.004242715337816202
way       0.004242715337816202
much      0.004082612872238232
say       0.004082612872238232
first     0.004002561639449248
head      0.004002561639449248
thing     0.003922510406660265
go        0.003842459173871275
voice     0.003842459173871275
rabbit    0.003762407941082292
looked    0.003602305475504323
never     0.003602305475504323
got       0.003602305475504323
get       0.003522254242715338
must      0.003522254242715338
mouse     0.003362151777137368
duchess           0.003362151777137368
round     0.003282100544348383
came      0.003202049311559398
tone      0.003202049311559398
dormouse          0.003202049311559398
great     0.003121998078770413
well      0.002961895613192443
back      0.002961895613192443
two       0.002961895613192443
cat       0.002881844380403458
march     0.002721741914825488
large     0.002641690682036503
last      0.002641690682036503
long      0.002561639449247518
```

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

```
Bigrams from file with top 50 frequencies
(('nick', 'said'), 0.0022732135216089002)
(('old', 'man'), 0.0020592640136927686)
(('bill', 'said'), 0.0014709028669234062)
(('said', 'nick'), 0.0005883611467693624)
(('big', 'two-hearted'), 0.00045464270432178005)
(('two-hearted', 'river'), 0.00045464270432178005)
(('war', 'cloud'), 0.0004278990158322636)
(('mister', 'adams'), 0.0003476679503637142)
(('nick', 'asked'), 0.0003476679503637142)
(('uncle', 'george'), 0.0003476679503637142)
(('could', 'see'), 0.00032092426187419767)
(('george', 'said'), 0.00032092426187419767)
(('three-day', 'blow'), 0.00032092426187419767)
(('little', 'man'), 0.0002941805733846812)
(('long', 'time'), 0.0002941805733846812)
(('man', 'said'), 0.0002941805733846812)
(('doctor', 'said'), 0.00026743688489516476)
(('big', 'trout'), 0.00024069319640564828)
(('krebs', 'said'), 0.00024069319640564828)
(('marjorie', 'said'), 0.0002139495079161318)
(('nick', 'knew'), 0.0002139495079161318)
(('nick', 'sat'), 0.0002139495079161318)
(('went', 'back'), 0.0002139495079161318)
(('billy', 'tabeshaw'), 0.00018720581942661531)
(('dick', 'boulton'), 0.00018720581942661531)
(('indian', 'camp'), 0.00018720581942661531)
(('nick', 'looked'), 0.00018720581942661531)
(('nick', 'took'), 0.00018720581942661531)
(('pine', 'trees'), 0.00018720581942661531)
(('said', 'nothing'), 0.00018720581942661531)
(('walked', 'along'), 0.00018720581942661531)
(('young', 'indian'), 0.00018720581942661531)
(('ad', 'said'), 0.00016046213093709883)
(('cross-country', 'snow'), 0.00016046213093709883)
(('every', 'day'), 0.00016046213093709883)
(('left', 'hand'), 0.00016046213093709883)
(('nick', 'felt'), 0.00016046213093709883)
(('nick', 'put'), 0.00016046213093709883)
(('nick', 'walked'), 0.00016046213093709883)
(('said', 'george'), 0.00016046213093709883)
(('sweet', 'fern'), 0.00016046213093709883)
(('would', 'come'), 0.00016046213093709883)
(('ever', 'seen'), 0.00013371844244758238)
(('george', 'gardner'), 0.00013371844244758238)
(('get', 'married'), 0.00013371844244758238)
(('man', 'looked'), 0.00013371844244758238)
(('mother', 'said'), 0.00013371844244758238)
(('negro', 'said'), 0.00013371844244758238)
(('never', 'saw'), 0.00013371844244758238)
(('nick', 'held'), 0.00013371844244758238)
```

Figure 4: Hemmingway's top 50 bigrams by frequency

```
Bigrams from file with top 50 frequencies
(('said', 'alice'), 0.003433553279789807)
(('mock', 'turtle'), 0.0016421341772907773)
(('march', 'hare'), 0.0009255665362911653)
(('thought', 'alice'), 0.0007762816110829129)
(('white', 'rabbit'), 0.0006568536709163109)
(('alice', 'thought'), 0.00035828382049980594)
(('alice', 'could'), 0.00032842683545815545)
(('alice', 'said'), 0.00032842683545815545)
(('poor', 'alice'), 0.00032842683545815545)
(('alice', 'replied'), 0.00026871286537485446)
(('alice', 'looked'), 0.00023885588033320396)
(('king', 'said'), 0.00023885588033320396)
(('little', 'thing'), 0.00023885588033320396)
(('poor', 'little'), 0.00023885588033320396)
(('alice', 'began'), 0.00020899889529155347)
(('cried', 'alice'), 0.00020899889529155347)
(('good', 'deal'), 0.00020899889529155347)
(('oh', 'dear'), 0.00020899889529155347)
(('beautiful', 'soup'), 0.00017914191024990297)
(('could', 'see'), 0.00017914191024990297)
(('golden', 'key'), 0.00017914191024990297)
(('great', 'hurry'), 0.00017914191024990297)
(('little', 'door'), 0.00017914191024990297)
(('said', 'nothing'), 0.00017914191024990297)
(('three', 'gardeners'), 0.00017914191024990297)
(('alice', 'felt'), 0.00014928492520825248)
(('alice', 'went'), 0.00014928492520825248)
(('another', 'moment'), 0.00014928492520825248)
(('came', 'upon'), 0.00014928492520825248)
(('cheshire', 'cat'), 0.00014928492520825248)
(('feet', 'high'), 0.00014928492520825248)
(('kid', 'gloves'), 0.00014928492520825248)
(('little', 'golden'), 0.00014928492520825248)
(('next', 'witness'), 0.00014928492520825248)
(('offended', 'tone'), 0.00014928492520825248)
(('play', 'croquet'), 0.00014928492520825248)
(('right', 'size'), 0.00014928492520825248)
(('trembling', 'voice'), 0.00014928492520825248)
(('white', 'kid'), 0.00014928492520825248)
(('would', 'go'), 0.00014928492520825248)
(('alice', 'hastily'), 0.00011942794016660198)
(('alice', 'ventured'), 0.00011942794016660198)
(('come', 'back'), 0.00011942794016660198)
(('father', 'william'), 0.00011942794016660198)
(('hare', 'said'), 0.00011942794016660198)
(('inches', 'high'), 0.00011942794016660198)
(('lobster', 'quadrille'), 0.00011942794016660198)
(('low', 'voice'), 0.00011942794016660198)
(('never', 'heard'), 0.00011942794016660198)
(('old', 'fellow'), 0.00011942794016660198)
```

Figure 5: Carroll's top 50 bigrams by frequency

```
Bigrams from file with top 50 mutual information scores
(('san', 'siro'), 12.868513923895463)
(('billy', 'tabeshaw'), 12.190442018782825)
(('sweet', 'fern'), 12.020517017340513)
(('three-day', 'blow'), 11.374525083221798)
(('cross-country', 'snow'), 10.798124596004065)
(('mister', 'adams'), 10.71595605442324)
(('dick', 'boulton'), 10.538365322203134)
(('war', 'cloud'), 10.01053292876789)
(('young', 'indian'), 9.60547951806167)
(('every', 'day'), 9.283551423174307)
(('two-hearted', 'river'), 9.283551423174305)
(('uncle', 'george'), 9.236245708395952)
(('ever', 'seen'), 8.973211302562156)
(('george', 'gardner'), 8.973211302562156)
(('indian', 'camp'), 8.889272484062262)
(('pine', 'trees'), 8.680384327075561)
(('big', 'two-hearted'), 8.561085398703215)
(('pulled', 'hard'), 8.476196501116702)
(('old', 'man'), 7.929751553417329)
(('long', 'time'), 7.848165278503663)
(('never', 'saw'), 7.424907272419848)
(('left', 'hand'), 7.357552004618086)
(('walked', 'along'), 7.164906926675689)
(('get', 'married'), 6.933054176090172)
(('could', 'see'), 6.567170161164196)
(('little', 'man'), 6.247927513443587)
(('one', 'day'), 6.066320706953636)
(('bill', 'said'), 6.013611548613749)
(('big', 'trout'), 5.753730476645611)
(('would', 'come'), 5.697253711769333)
(('went', 'away'), 5.245583572975283)
(('ad', 'said'), 5.170851290424945)
(('nick', 'asked'), 5.013022480793591)
(('marjorie', 'said'), 4.948458869088499)
(('doctor', 'said'), 4.948458869088494)
(('krebs', 'said'), 4.925738792588412)
(('said', 'nothing'), 4.919312523428982)
(('nick', 'knew'), 4.643788671127874)
(('negro', 'said'), 4.618310267396165)
(('nick', 'held'), 4.5979849815147436)
(('nick', 'said'), 4.572747690015724)
(('george', 'said'), 4.3865799814803825)
(('went', 'back'), 4.327804661224034)
(('man', 'looked'), 4.266074860153845)
(('nick', 'sat'), 4.106131885185073)
(('mother', 'said'), 4.033347766675007)
(('nick', 'took'), 3.979075148870253)
(('nick', 'walked'), 3.935019968792316)
(('nick', 'stood'), 3.875518957043653)
(('nick', 'felt'), 3.7906300594571416)
```

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

```
Bigrams from file with top 50 mutual information scores
(('play', 'croquet'), 11.768537579120732)
(('golden', 'key'), 11.639254562175768)
(('kid', 'gloves'), 11.57214036631723)
(('white', 'kid'), 10.124681389346009)
(('beautiful', 'soup'), 9.987177865596077)
(('three', 'gardeners'), 9.972678295900963)
(('march', 'hare'), 9.944109143704189)
(('good', 'deal'), 9.669001905569822)
(('cheshire', 'cat'), 9.59861257767842)
(('trembling', 'voice'), 9.183575078399578)
(('mock', 'turtle'), 9.147595781294015)
(('next', 'witness'), 9.124681389346009)
(('feet', 'high'), 9.105572566398305)
(('white', 'rabbit'), 9.029524156305671)
(('great', 'hurry'), 8.871700648176141)
(('right', 'size'), 8.74616976609228)
(('offended', 'tone'), 8.709643890067165)
(('oh', 'dear'), 8.572140366317232)
(('another', 'moment'), 7.939872150817717)
(('little', 'golden'), 7.546145157784286)
(('came', 'upon'), 7.3311322668134355)
(('poor', 'little'), 6.331132266813439)
(('little', 'door'), 5.709643890067168)
(('little', 'thing'), 5.416862140839324)
(('would', 'go'), 5.276684482791058)
(('poor', 'alice'), 5.161207265371125)
(('could', 'see'), 5.124180540515814)
(('thought', 'alice'), 4.893201717387061)
(('cried', 'alice'), 4.887642192045162)
(('alice', 'replied'), 4.7141593711896626)
(('said', 'alice'), 4.395956374403237)
(('alice', 'felt'), 4.200581503705267)
(('alice', 'looked'), 3.9103622685452457)
(('said', 'nothing'), 3.8578948486511084)
(('alice', 'thought'), 3.7777244999671247)
(('alice', 'could'), 3.43538222881012)
(('alice', 'began'), 3.3515892918049524)
(('king', 'said'), 3.225626633151597)
(('alice', 'went'), 2.349104028415356)
(('alice', 'said'), 1.0098979420961598)
```

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

2b.

Initially, `'Smart.English.stop'` 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 `'Smart.English.stop'` were words that should keep as part of

the text analysis. The NLTK, `nltk.corpus.stopwords.words('english')` 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.

```
Display first 50 Stopwords:
['â€™s', 'a', "a's", 'able', 'about', 'above', 'according', 'accordingly', 'across', 'actually', 'after', 'afterwards', 'agai
n', 'against', "ain't", 'all', 'allow', 'allows', 'almost', 'alone', 'along', 'already', 'also', 'although', 'always', 'am', 'a
mong', 'amongst', 'an', 'and', 'another', 'any', 'anybody', 'anyhow', 'anyone', 'anything', 'anyway', 'anyways', 'anywhere', 'a
part', 'appear', 'appreciate', 'appropriate', 'are', "aren't", 'around', 'as', 'aside', 'ask', 'asking']
```

Figure 7: First 50 stop words from `'Smart.English.stop'`

```
Display first 50 Stopwords:
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'y
ourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those',
'am', 'is', 'are', 'was', 'were', 'be']
```

Figure 8: First 50 stop words from `nltk.corpus.stopwords.words('english')`

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

a. Hemingway addstopwords

```
addstopwords =["'s", "n't","'d", "part", "ii", "two-", "'you", "'re", "'t","'it","'no","'m", "'all", "'right/'", "'that",
              "'ve"]
```

b. Carroll addstopwords

```
addstopwords =["'and", "'ll", "'ve", "'but", "'what", "'oh","n't", "'s", "'m", "'it", "'you", "'re", "'that", "'d", "'well",
              "'of", "'why"]
```

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.

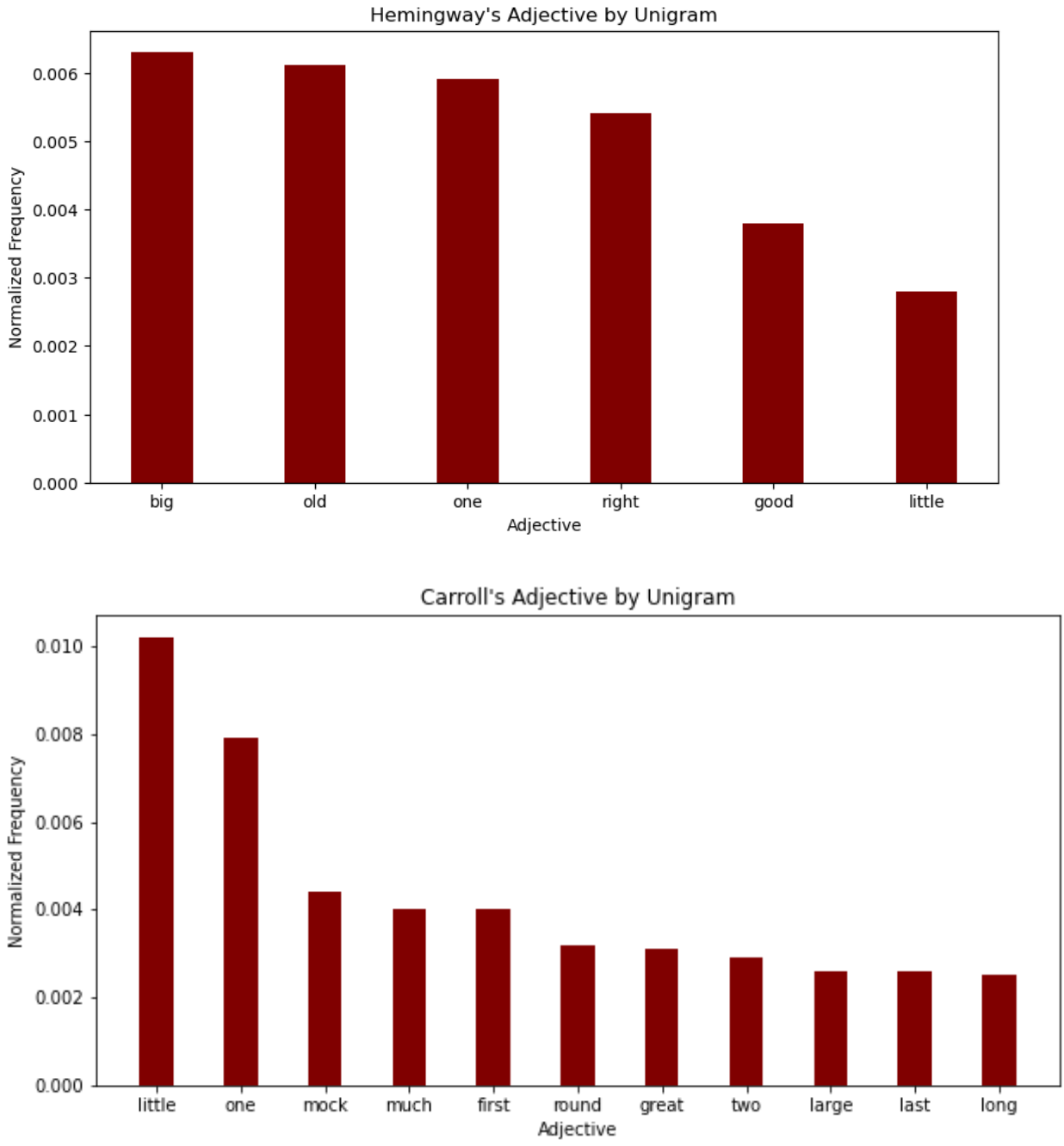

Figure 10: Hemingway and Carroll's Adjective in Top 50 Words by Frequency

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.

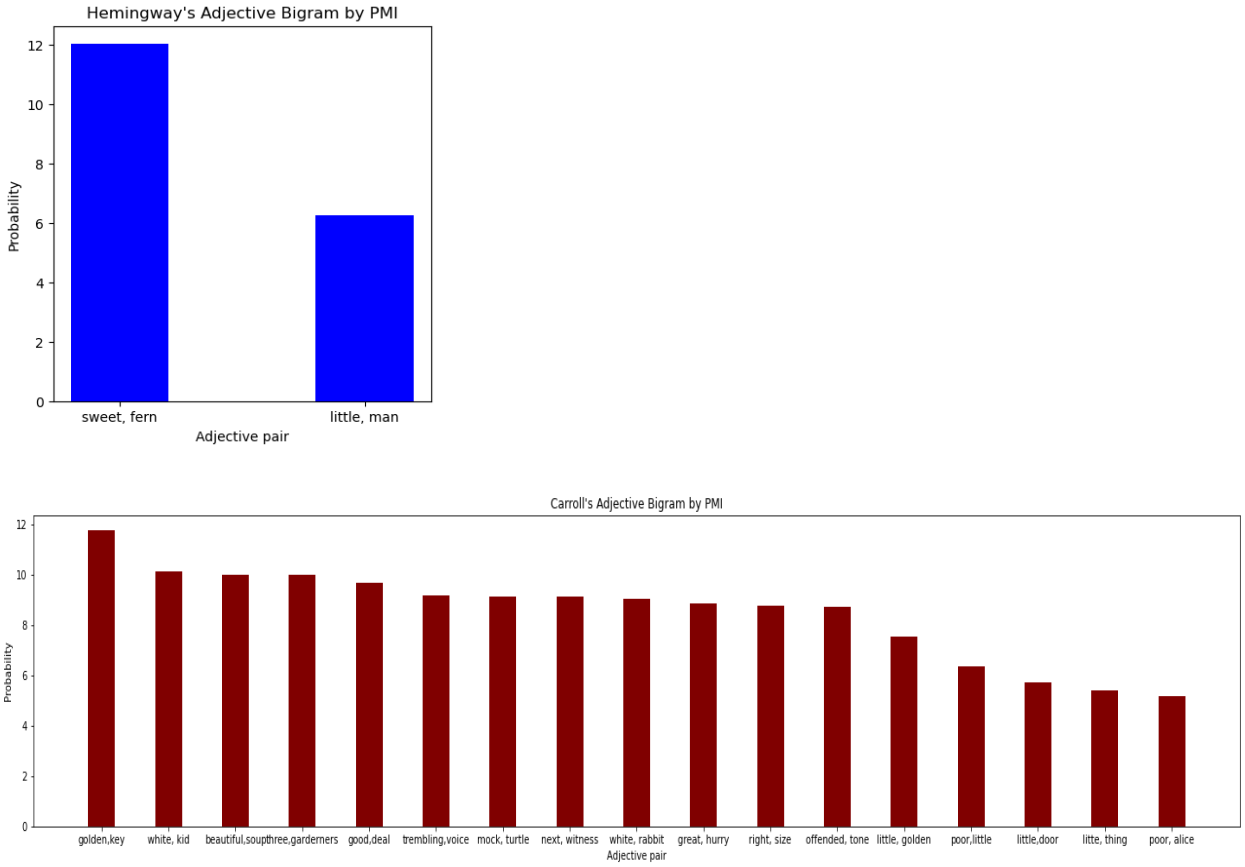

Figure 11: Hemingway and Carroll's Adjective in Top 50 Bigrams by PMI