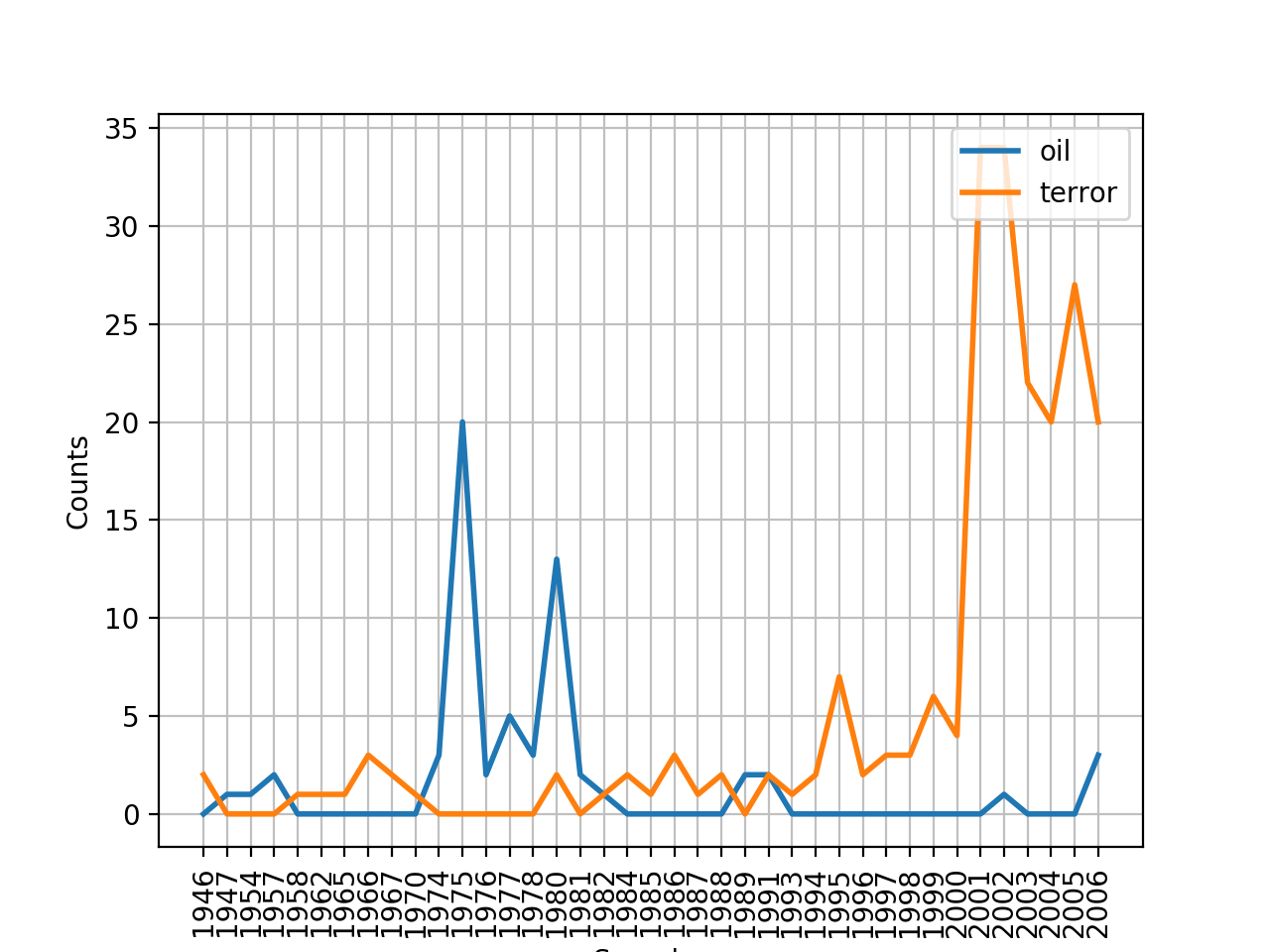
**Chapter 2 NLTK Exercises**

**State of the Union Address Graph**



**Exercise 7**

**Text (Wall Street Journal)**

s , such as telephones or watches . However , none of the big three weeklies re

ther intellectual-property rights . However , five other countries -- China , T

to be sold \*-50 in the U.S. \*T\*-1 . However , after two meetings with the Sovie

clining . The purchasing managers , however , also said that orders turned up i

of worries about potential abuse , however , he says 0 the state will begin \*-

ll game . '' Some fellow teachers , however , viewed Mrs. Yeargin as cocky and

ssary . '' He declined \*-2 to say , however , how much Sea Containers might rai

e going \*-3 to top us . '' Temple , however , harshly criticized Sea Containers

a bottle . In the last year or so , however , this exclusive club has taken in

. These prices seem rather modest , however , in light of other French wines fr

ons can steal a march on Burgundy , however . The six wines of the Domaine de l

it would study other alternatives . However , it has n't yet made any proposals

t of high-yield subordinated debt . However , the junk-bond market has collapse

T\*-227 usually stays in the tower , however . More often than not , ringers thi

y are n't accepted \*-1 everywhere , however . The oldest bell-ringing group in

on Wednesday . There are no signs , however , of China 's yielding on key issue

York . A spokesman for the state , however , calls the idea `` not effective o

e considerable skill \* to resolve . However , neither Mr. Baum nor Mr. Harper h

sly give the president such power . However , the president does have a duty \*

r. Riese says 0 \*T\*-1 . Until now , however , buyers who \*T\*-53 wanted \*-1 to f

all this \*-2 lying down for long , however . Its new products and trading tech

ock-selection products . They may , however , risk \*-1 bringing some damaging i

nment or private watchdogs insist , however , on \* introducing greater friction

rm 's equity . '' CS First Boston , however , benefits from the backing of its

the defendants . The big problem , however , is that there 's no guarantee tha

Of the 25 uses of “however” found in the Wall Street Journal text, 7 of them were used to begin the sentence whereas 18 of them were used in other contexts. In the different contexts, however was used to mean “nevertheless” or “nonetheless” or “in whatever manner”. Of the 18 other uses, however was either used as an aside, meaning an interruption, or a conjunctive verb, joining two sentences together. The majority of the sentences, it was used to either begin a sentence or as an aside.

**Exercise 9**

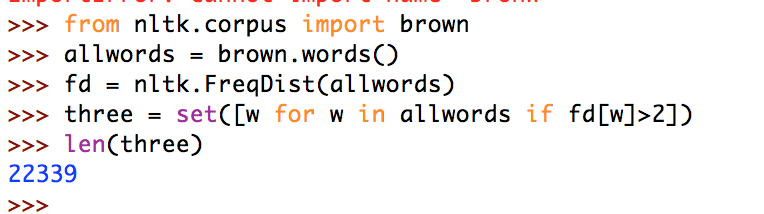
|  |  |  |
| --- | --- | --- |
|  | **Text 3: Book of Genesis** | **Text 7: Wall Street Journal** |
| **Distinct Number of Words**  ***(Created a set of all words)*** | 2789 | 12408 |
| **Lexical Diversity**  ***(Created set of words to find unique words and divided by total number of words)*** | 0.06230453042623537 | 0.12324685128531129 |
| **Collocations**  ***(Determined by using text#.collocations())*** | said unto; pray thee; thou shalt; thou hast; thy seed; years old;  spake unto; thou art; LORD God; every living; God hath; begat sons;  seven years; shalt thou; little ones; living creature; creeping thing;  savoury meat; thirty years; every beast | million \*U\*; New York; billion \*U\*; Wall Street; program trading; Mrs.  Yeargin; vice president; Stock Exchange; Big Board; Georgia Gulf;  chief executive; Dow Jones; S&P 500; says \*T\*-1; York Stock; last  year; Sea Containers; South Korea; American Express; San Francisco |

The task for exercise 9 was to analyze the differences between two texts. We chose to analyze the Book of Genesis and the Wall Street Journal since they take place in completely different time periods. By using lexical diversity, we found that the Book of Genesis repeats words more often than the Wall Street Journal. Using collocations, we found that the Book of Genesis used words in a religious context while the Wall Street Journal used words in a financial context. We analyzed the word “company” and found that in the Book of Genesis, the word company is relating to a groups of people while the Wall Street Journal utilizes it to mean business.

**Exercise 12**

The number of distinct words in the CMU Pronunciation dictionary is 123455 words, found by len(cmu.dict()). Since the code took too long to run through the entire dictionary, we selection 100 random words from the text and found that 13 words had more than one possible pronunciation.

**Exercise 15**



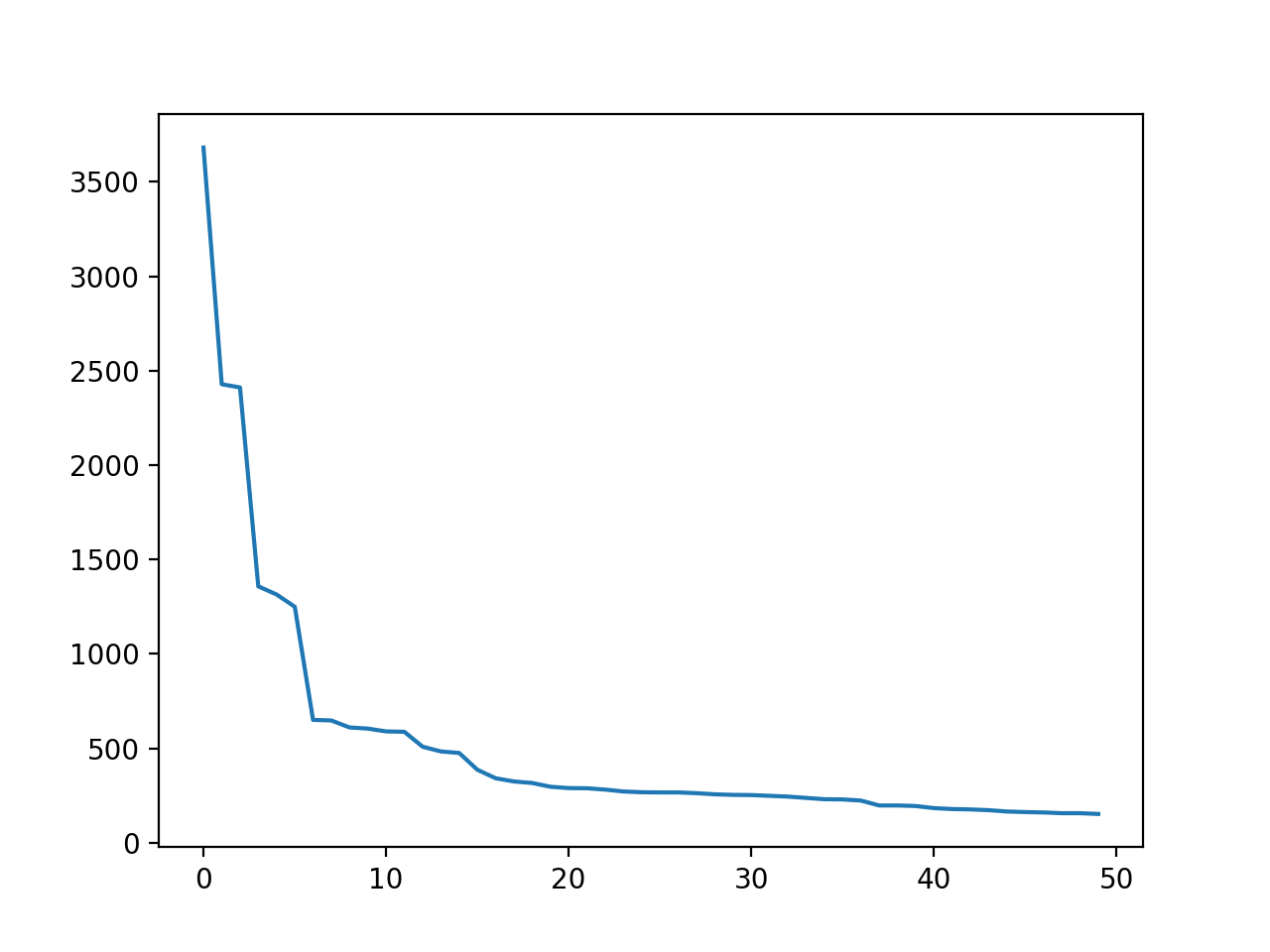
In exercise 15, the task was to find all words in the Brown Corpus that occurred more than three times. We did this by creating a frequency distribution of all the words in the Brown Corpus and then placing them into a list if their frequency was more than 2. The set was used to make sure that there were no duplicates of words. The program found that there were 22,339 words that occurred more than 3 times in the Brown Corpus. We simply printed the number because there were too many words to print.

**Exercise 18**

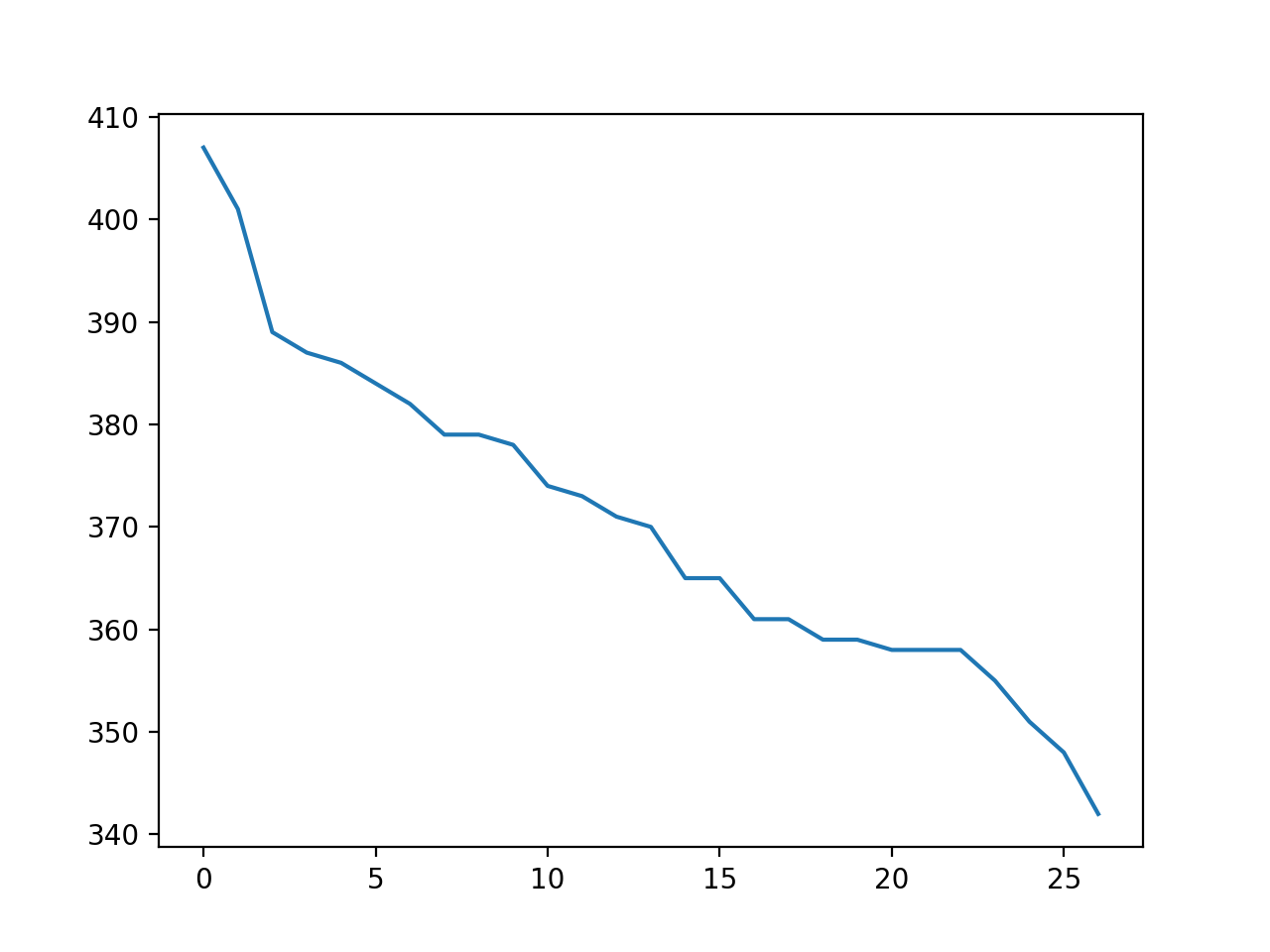
*[(('Sperm', 'Whale'), 119), (('Moby', 'Dick'), 84), (('old', 'man'), 80), (('White', 'Whale'), 74), (('Captain', 'Ahab'), 62), (('sperm', 'whale'), 56), (('mast', 'head'), 45), (('Right', 'Whale'), 39), (('mast', 'heads'), 36), (('ye', 'see'), 36), (('whale', 'ship'), 33), (('cried', 'Ahab'), 33), (('Captain', 'Peleg'), 32), (('white', 'whale'), 31), (('Aye', 'aye'), 31), (('Mr', 'Starbuck'), 29), (('one', 'hand'), 28), (('quarter', 'deck'), 28), (('whale', 'boat'), 26), (('?"', 'said'), 26), (('cried', 'Stubb'), 23), (('one', 'side'), 23), (('every', 'one'), 21), (('never', 'mind'), 20), (('let', 'us'), 20), (('ye', 'ye'), 19), (('chief', 'mate'), 19), (('said', 'Stubb'), 18), (('New', 'Bedford'), 18), (('years', 'ago'), 18), (('?"', 'cried'), 17), (('cried', 'Starbuck'), 17), (('old', 'Ahab'), 16), (('Cape', 'Horn'), 16), (('lower', 'jaw'), 16), (('something', 'like'), 16), (('would', 'seem'), 16), (('try', 'works'), 16), (('Look', 'ye'), 16), ((".'", '"\''), 16), (('thee', 'thou'), 16), (('round', 'round'), 16), (('ivory', 'leg'), 15), (('boat', 'crew'), 15), (('would', 'fain'), 15), (('Dough', 'Boy'), 15), (('well', 'known'), 15), (('Fast', 'Fish'), 15), (('three', 'years'), 15), (('Loose', 'Fish'), 14)]*

The task for Exercise 18 was to find the 50 most common bigrams in a text while excluding stop words and punctuation. A bigram is a pair of words that appear frequently together. To achieve this task, we had to create a list of words in which the stop words, which we got through the NLTK corpus, and punctuation, which we created our own list for, were removed. After doing so, we simply used the NLTK bigrams method and a frequency distribution to find the most common bigrams. I found that the most common bigram in text1, *Moby Dick*, was (‘Sperm’, ‘Whale’), which makes sense because the main villain in Moby Dick is a Sperm Whale.

**Exercise 23**



In our first plot, we plotted text3, The Book of Genesis. In this case the curve shows a relatively logarithmic relationship which is an inverse relationship. On the right end, the graph approaches infinity while on the right it approaches zero, a clear inverse relationship that upholds Zipf’s Law.



In the second plot, we use a randomly generated text with 10000 characters. In this case also, we see an inverse relationship between rank and frequency of word type, upholding Zipf’s Law again.

**Exercise 27**

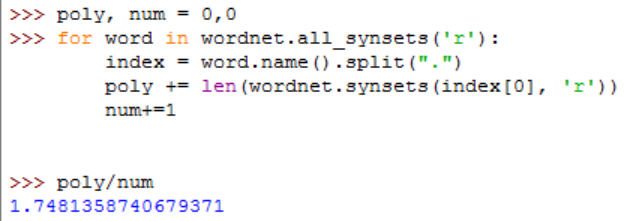
Adjective Polysemy: 2. 3232540207094075

Adverb Polysemy: 1.7481358740679371

Verb Polysemy: 5.366020193215661

Noun Polysemy: 1.9688729221214152

Sample Code for determining Adverb Polysemy:



Exercise 27 had us looking at polysemy in texts, or the senses it is used in. We found that verbs were used in the most senses while Adverbs were used in the least senses.