**CHAPTER 2**

2)

1945-Truman.txt : men: 2 women: 2 people: 10

1946-Truman.txt : men: 12 women: 7 people: 49

1947-Truman.txt : men: 7 women: 2 people: 12

...

2004-GWBush.txt : men: 6 women: 8 people: 21

2005-GWBush.txt : men: 8 women: 11 people: 18

2006-GWBush.txt : men: 7 women: 7 people: 22

The word 'women' hasn't really been used until the 1970s, which was also a time for dramatic social change for them. After this, the word 'women' has been used around the same number of times as the word 'men', but the most frequent word is still 'people'.

5 )

>>> from nltk.corpus import wordnet as wn

>>> animal = wn.synsets('animal')[0]

>>> animal.member\_meronyms()

[]

>>> animal.part\_meronyms()

[Synset('face.n.07'), Synset('head.n.01')]

>>> animal.substance\_meronyms()

[Synset('animal\_tissue.n.01')]

>>> animal.member\_holonyms()

[Synset('animalia.n.01')]

>>> animal.part\_holonyms()

[]

>>> animal.substance\_holonyms()

[]

7 )

>>> text1.concordance('however')

Displaying 25 of 95 matches:

gledy - piggledy whale statements , however authentic , in these extracts , for

lave ? Tell me that . Well , then , however the old sea - captains may order me

ea - captains may order me about -- however they may thump and punch me about ,

...

isely -- who knows ? Certain I am , however , that a king ' s head is solemnly

o scientific description . As yet , however , the sperm whale , scientific or p

IZONTAL TAIL . There you have him . However contracted , that definition is the

several varieties , most of which , however , are little known . Broad - nosed

9 )

>>> caesar = nltk.Text(nltk.corpus.gutenberg.words('shakespeare-caesar.txt'))

>>> bible = nltk.Text(nltk.corpus.gutenberg.words('bible-kjv.txt'))

>>> len(set(caesar)) / len(caesar)

>>> caesar.concordance('blood')

Displaying 24 of 24 matches:

That comes in Triumph ouer Pompeyes blood ? Be gone , Runne to your houses , f

d neede an Oath . When euery drop of blood That euery Roman beares , and Nobly

d in the Spirit of men , there is no blood : O that we then could come by Caesa

...

hter to his Brutus , When greefe and blood ill temper ' d , vexeth him ? Bru .

, or some Diuell , That mak ' st my blood cold , and my haire to stare ? Speak

doest sinke to night ; So in his red blood Cassius day is set . The Sunne of Ro

>>> bible.concordance('blood')

Displaying 25 of 447 matches:

done ? the voice of thy brother ' s blood crieth unto me from the ground . 4 :

her mouth to receive thy brother ' s blood from thy hand ; 4 : 12 When thou til

with the life thereof , which is the blood thereof , shall ye not eat . 9 : 5 A

...

he Egyptians ; and when he seeth the blood upon the lintel , and on the two sid

smitten that he die , there shall no blood be shed for him . 22 : 3 If the sun

n be risen upon him , there shall be blood shed for him ; for he should make fu

12 )

0.15813859300959865

17 )

>>> highest\_freq(text1)

['whale', 'one', 'like', 'upon', 'man', 'ship', 'Ahab', 'ye', 'old', 'sea', 'would', 'head', 'though', 'boat', 'time', 'long', 'said', 'yet', 'still', 'great', 'two', 'seemed', 'must', 'Whale', 'last', 'way', 'Stubb', 'see', 'Queequeg', 'little', 'round', 'whales', 'say', 'three', 'men', 'thou', 'may', 'us', 'every', 'much', 'could', 'Captain', 'first', 'side', 'hand', 'ever', 'Starbuck', 'never', 'good', 'white']

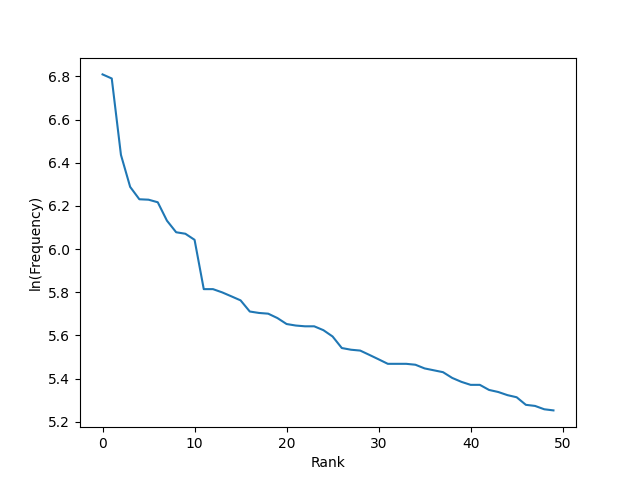
18 )

>>> highest\_freq(nltk.Text(nltk.corpus.gutenberg.words('blake-poems.txt')))

[('.', '"'), ("'", 'er'), ('night', ','), ('day', ','), ('Love', ','), ('delight', ','), ('thee', '?'), ('father', ','), (':', '"'), ('Little', 'Lamb'), ('Lamb', ','), ('thee', ','), ('voice', ','), ('away', ','), ('deep', ','), ('joy', ','), ('!', 'Sweet'), ('sleep', ','), ('weep', '.'), (',', 'Till'), ('?', '"'), ('child', ','), (',', '"'), ('thee', '!'), ('Mercy', ','), ('Pity', ','), ('Peace', ','), ('human', 'form'), (',', 'like'), ('care', ','), ('play', ','), (',', 'Like'), ('tell', 'thee'), ('lamb', ','), ('light', ','), ('!', 'weep'), ('dark', ','), ('infant', "'"), ('Sleep', ','), (',', 'Pity'), (',', 'Peace'), ('man', ','), ('care', '.'), ('snow', ','), ('delight', '.'), ('year', '.'), ('children', ','), (',', '--'), ('mother', ','), ('away', '.')]

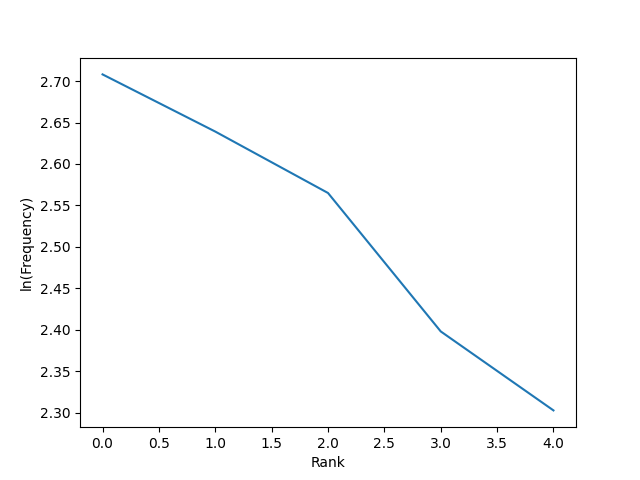
23 )

a)



While the linear relationship isn't perfect, the most common word is clearly significantly more frequent than the 50th most common word. At the ends, the line tapers off.

b )



Zipf's Law seems to be more applicable when the text is randomly generated.

27 )

Noun Average : 1.3972549720138145

Verb Average : 2.6370967741935485

Adjective Average : 1.5458440445586976

Adverb Average : 1.3291803278688525

**CHAPTER 3**

20 )

'65°'

22 )

r"/live/i"

**CHAPTER 6**

4 )

>>> classifier.show\_most\_informative\_features(30)

Most Informative Features

contains(outstanding) = True pos : neg = 13.5 : 1.0

contains(mulan) = True pos : neg = 9.2 : 1.0

contains(seagal) = True neg : pos = 8.0 : 1.0

contains(wonderfully) = True pos : neg = 6.5 : 1.0

contains(damon) = True pos : neg = 6.0 : 1.0

contains(awful) = True neg : pos = 5.5 : 1.0

contains(wasted) = True neg : pos = 5.4 : 1.0

contains(poorly) = True neg : pos = 5.3 : 1.0

contains(lame) = True neg : pos = 5.2 : 1.0

contains(ridiculous) = True neg : pos = 5.2 : 1.0

contains(jedi) = True pos : neg = 5.1 : 1.0

contains(waste) = True neg : pos = 5.0 : 1.0

contains(era) = True pos : neg = 4.6 : 1.0

contains(laughable) = True neg : pos = 4.6 : 1.0

contains(flynt) = True pos : neg = 4.4 : 1.0

contains(stupid) = True neg : pos = 4.4 : 1.0

contains(pointless) = True neg : pos = 4.4 : 1.0

contains(hanks) = True pos : neg = 4.4 : 1.0

contains(portrayal) = True pos : neg = 4.3 : 1.0

contains(worst) = True neg : pos = 4.3 : 1.0

contains(dull) = True neg : pos = 4.2 : 1.0

contains(bland) = True neg : pos = 4.2 : 1.0

contains(allows) = True pos : neg = 4.0 : 1.0

contains(unfunny) = True neg : pos = 3.9 : 1.0

contains(damme) = True neg : pos = 3.8 : 1.0

contains(memorable) = True pos : neg = 3.8 : 1.0

contains(terrific) = True pos : neg = 3.6 : 1.0

contains(superb) = True pos : neg = 3.6 : 1.0

contains(fantastic) = True pos : neg = 3.5 : 1.0

contains(mess) = True neg : pos = 3.5 : 1.0

These features are particularly informative because they have higher positive to negative correlations for movie reviews. For instance, the words "outstanding" and "wonderfully" both have positive connotations and are also highly ranked in how informative they are in predicting a movie review. I found it surprising that the word "damon", which I assume has to do with the actor Matt Damon, is so highly ranked in how informative the presence of the word is. I think it might be a good example of how there might be a correlation between good movie reviews and certain words, but that does not equate to correlation.