

NOTE: skipped chapter 2 exercise 23 and chapter 3 exercise 22

Chapter 2

4. Read in the texts of the *State of the Union* addresses, using the `state_union` corpus reader. Count occurrences of men, women, and people in each document. What has happened to the usage of these words over time?

1945 men: 2 women: 2 people: 2

1946 men: 16 women: 7 people: 7

...

2005 men: 8 women: 11 people: 11

2006 men: 7 women: 7 people: 7

Overtime, the number of occurrences of “women” increased, which makes sense with the civil rights movements. The number of occurrences of “people” tends to always be the highest frequency out of the three categories

5. Investigate the holonym-meronym relations for some nouns. Remember that there are three kinds of holonym-meronym relation, so you need to use: `member_meronyms()`, `part_meronyms()`, `substance_meronyms()`, `member_holonyms()`, `part_holonyms()`, and `substance_holonyms()`.

Member meronyms for “school”: [Synset('schoolteacher.n.01'), Synset('staff.n.03')]

Part meronyms for “tree”: [Synset('burl.n.02'), Synset('crown.n.07'), Synset('limb.n.02'), Synset('stump.n.01'), Synset('trunk.n.01')]

Substance meronyms for “tree”: [Synset('heartwood.n.01'), Synset('sapwood.n.01')]

Member holonyms for “tree”: [Synset('forest.n.01')]

Part holonyms for “tree”: [Synset('dwelling.n.01')]

Substance holonyms for “ice”: [Synset('glacier.n.01'), Synset('ice_cube.n.01')]

7. According to Strunk and White's *Elements of Style*, the word *however*, used at the start of a sentence, means "in whatever way" or "to whatever extent", and not "nevertheless". They give this example of correct usage: *However you advise him, he will probably do as he thinks best.* (<http://www.bartleby.com/141/strunk3.html>) Use the concordance tool to study actual usage of this word in the various texts we have been considering. See also the *LanguageLog* posting "Fossilized prejudices about 'however'" at <http://itre.cis.upenn.edu/~myl/languagelog/archives/001913.html>

inctly , and was sadly alarmed at . However , when I read on , I found it was n
hough she speaks rather too quick . However , she is very agreeable , and Mrs .
at I would not expose any body to . However , as they are so very desirous to h

m often astonished at what I hear . However , this does make a difference ; and
“However” is rather used correctly when used at the start of a sentence!

9. Pick a pair of texts and study the differences between them, in terms of vocabulary, vocabulary richness, genre, etc. Can you find pairs of words which have quite different meanings across the two texts, such as *monstrous* in *Moby Dick* and in *Sense and Sensibility*?

“date” in romance:

hee boy-furiendo ? ? Maybe you likee date with me " ? ? `` I beg your pardon
fore and her gay chatter about their date lightened my mood a little . But one

“date” in news:

bling funds and re-set the effective date so that an orderly implementation of
duled to go up by 1 per cent on that date to pay for other social security cos
ll to be taken on Nov. 15 , the same date as a \$581,000 bond election for the

12. The CMU Pronouncing Dictionary contains multiple pronunciations for certain words. How many distinct words does it contain? What fraction of words in this dictionary have more than one possible pronunciation?

123455 distinct words

9241/123455, or approx 7.49% have more than one possible pronunciation

17. Write a function that finds the 50 most frequently occurring words of a text that are not stopwords.

['I', 'He', 'said', 'would', 'She', 'The', 'could', 'like', 'It', 'But', 'And', 'back', 'thought', 'You', 'little',
'time', 'get', 'got', 'know', 'never', 'way', 'went', 'eyes', 'go', 'came', 'see', 'old', 'even', 'looked',
'They', 'knew', 'much', 'around', 'There', 'long', 'good', 'away', 'felt', 'day', 'still', 'made', 'say',
'going', 'something', 'think', 'In', 'right', 'night', 'look', 'thing']

18. Write a program to print the 50 most frequent bigrams (pairs of adjacent words) of a text, omitting bigrams that contain stopwords.

[('New', 'York'), ('per', 'cent'), ('United', 'States'), ('last', 'year'), ('White', 'House'), ('last', 'week'),
('He', 'said'), ('home', 'runs'), ('President', 'Kennedy'), ('The', 'President'), ('last', 'night'), ('years',
'ago'), ('I', 'think'), ('San', 'Francisco'), ('Premier', 'Khrushchev'), ('Kansas', 'City'), ('high',
'school'), ('Los', 'Angeles'), ('United', 'Nations'), ('New', 'Orleans'), ('High', 'School'), ('vice',
'president'), ('sales', 'tax'), ('The', 'first'), ('American', 'Catholic'), ('The', 'jury'), ('He', 'also'), ('El',
'Paso'), ('In', 'addition'), ('first', 'time'), ('Rules', 'Committee'), ('Air', 'Force'), ('tax', 'bill'), ('two',
'years'), ('would', 'like'), ('The', 'new'), ('Rhode', 'Island'), ('social', 'security'), ('weeks', 'ago'),
('home', 'run'), ('police', 'said'), ('higher', 'education'), ('collective', 'bargaining'), ('The',
'company'), ('farm', 'equipment'), ('jury', 'said'), ('three', 'years'), ('million', 'dollars'), ('We', 'must'),
('President', 'said')]

27. The polysemy of a word is the number of senses it has. Using WordNet, we can determine that the noun *dog* has 7 senses with: `len(wn.synsets('dog', 'n'))`. Compute the average polysemy of nouns, verbs, adjectives and adverbs according to WordNet.

average polysemy of:

nouns: 1.2833560159282222

verbs: 2.1866273523545225

adjectives: 1.4104837960813446

adverbs: 1.2532916759651864

Chapter 3

20. Write code to access a favorite webpage and extract some text from it. For example, access a weather site and extract the forecast top temperature for your town or city today.

"detailedForecast": "Mostly clear, with a low around 36. Northwest wind around 8 mph."

Chapter 6

4. Using the movie review document classifier discussed in this chapter, generate a list of the 30 features that the classifier finds to be most informative. Can you explain why these particular features are informative? Do you find any of them surprising?

Most Informative Features

contains(outstanding) = True	pos : neg = 10.4 : 1.0
contains(seagal) = True	neg : pos = 8.7 : 1.0
contains(mulan) = True	pos : neg = 8.1 : 1.0
...	
contains(boring) = True	neg : pos = 3.5 : 1.0
contains(portrayal) = True	pos : neg = 3.5 : 1.0
contains(memorable) = True	pos : neg = 3.4 : 1.0

Most of these features have exaggerated connotated adjectives so they are good indicators of whether the review was positive or negative (instead of nouns which tell us meaning but rarely sentiment). I found “seagal” surprising because I have never heard of that word before.