CSCI-8450

2018 Spring

Yuchen Wang

1. Review [1](http://www.nltk.org/book/ch01.html#sec-computing-with-language-texts-and-words) on computing with language. How many words are there in text2? How many distinct words are there?

len(text2)

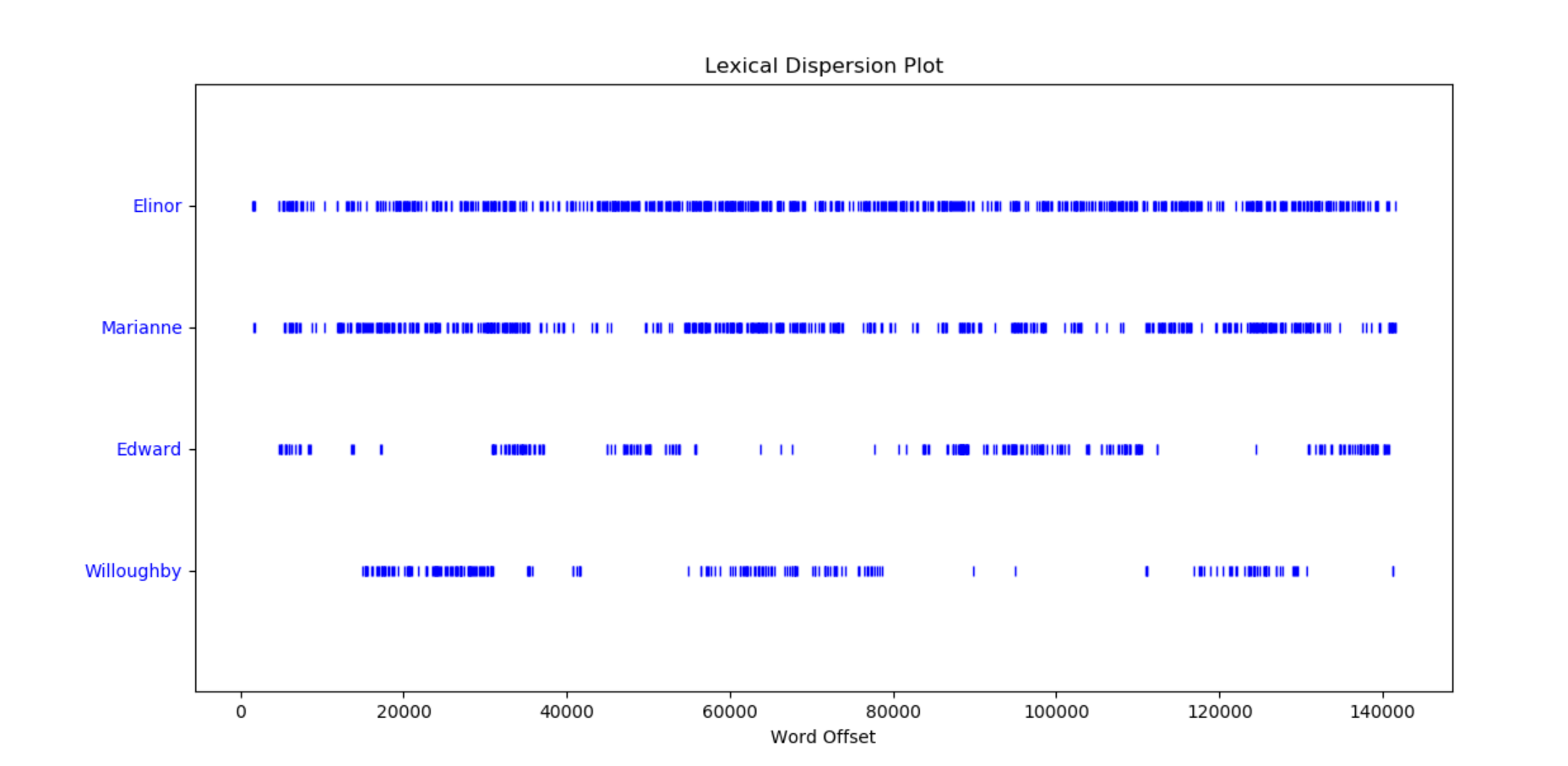
len(set(text2))

141576, 6833

1. Compare the lexical diversity scores for humor and romance fiction in [1.1](http://www.nltk.org/book/ch01.html#tab-brown-types). Which genre is more lexically diverse?

The lexical diversity of humor is 0.231, while romance is 0.121. Therefore, humor is more

lexically diverse.

1. Produce a dispersion plot of the four main protagonists in Sense and Sensibility: Elinor, Marianne, Edward, and Willoughby. What can you observe about the different roles played by the males and females in this novel? Can you identify the couples?

text2.dispersion\_plot(["Elinor","Marianne","Edward","Willoughby"])

Elinor and Marianne show a lot in the book. When Edward shows, Willoughby does not show. Elinor and Marianne might be the couple.

1. Find the collocations in text5.

text5.collocations()

wanna chat; PART JOIN; MODE #14-19teens; JOIN PART; PART PART;

cute.-ass MP3; MP3 player; JOIN JOIN; times .. .; ACTION watches; guys

wanna; song lasts; last night; ACTION sits; -...)...- S.M.R.; Lime

Player; Player 12%; dont know; lez gurls; long time

1. Use text9.index() to find the index of the word *sunset*. You'll need to insert this word as an argument between the parentheses. By a process of trial and error, find the slice for the complete sentence that contains this word.

text9.index('sunset')

text9[613:644]

629

The complete sentence: ['CHAPTER', 'I', 'THE', 'TWO', 'POETS', 'OF', 'SAFFRON', 'PARK', 'THE', 'suburb', 'of', 'Saffron', 'Park', 'lay', 'on', 'the', 'sunset', 'side', 'of', 'London', ',', 'as', 'red', 'and', 'ragged', 'as', 'a', 'cloud', 'of', 'sunset', '.']

1. (1) Using list addition, and the set and sorted operations, compute the vocabulary of the sentences sent1 ... sent8.

(2) In addition: What is the size of the computed vocabulary? List the vocabulary and its size.

1. >>> len(set(sent1))

4

>>> len(set(sent2))

11

>>> len(set(sent3))

9

>>> len(set(sent4))

10

>>> len(set(sent5))

11

>>> len(set(sent6))

12

>>> len(set(sent7))

17

>>> len(set(sent8))

13

(2)

>>> len(sent1)

4

>>> set(sent1)

{'Ishmael', 'me', '.', 'Call'}

>>> len(sent2)

11

>>> set(sent2)

{'settled', 'had', 'of', 'The', 'been', 'family', 'in', 'Dashwood', '.', 'long', 'Sussex'}

>>> len(sent3)

11

>>> set(sent3)

{'and', 'In', 'earth', 'beginning', 'created', 'the', 'heaven', 'God', '.'}

>>> len(sent4)

13

>>> set(sent4)

{'and', '-', 'Senate', 'of', 'Citizens', 'Representatives', 'Fellow', 'House', 'the', ':'}

>>> len(sent5)

11

>>> set(sent5)

{'have', 'I', 'people', 'a', 'me', 'JOIN', 'lol', 'with', 'to', 'problem', 'PMing'}

>>> len(sent6)

17

>>> set(sent6)

{']', 'wind', 'ARTHUR', 'there', 'SCENE', '[', 'KING', 'Whoa', 'clop', ':', '!', '1'}

>>> len(sent7)

18

>>> set(sent7)

{'as', 'Nov.', ',', '61', 'Vinken', 'director', 'Pierre', 'a', 'nonexecutive', 'will', 'years', 'board', 'the', '29', '.', 'old', 'join'}

>>> len(sent8)

14

>>> set(sent8)

{'discreet', 'lady', 'SEXY', ',', 'encounters', '25', 'attrac', 'for', 'MALE', 'single', '.', 'older', 'seeks'}

1. (1) Find all the four-letter words in the Chat Corpus (text5). With the help of a frequency distribution (FreqDist), show these words in decreasing order of frequency.
2. In addition: What are the five most frequent words of length four? List these words together with their frequencies and counts of the number of times these words occur (see Table 3.1).

(1)

>>> FreqDist([w for w in text5 if len(w) == 4 and w.isalpha()])

FreqDist({'JOIN': 1021, 'PART': 1016, 'that': 274, 'what': 183, 'here': 181, 'have': 164, 'like': 156, 'with': 152, 'chat': 142, 'your': 137, ...})

(2)

>>> FreqDist([w for w in text5 if len(w) == 4 and w.isalpha()]).most\_common(5)

[('JOIN', 1021), ('PART', 1016), ('that', 274), ('what', 183), ('here', 181)]

fdist = FreqDist([w for w in text5 if len(w) == 4 and w.isalpha()])

>>> fdist['JOIN']

1021

>>> fdist.freq('JOIN')

0.10895315334542738

>>> fdist['PART']

1016

>>> fdist.freq('PART')

0.10841959235940668

>>> fdist['that']

274

>>> fdist.freq('that')

0.02923914203393448

>>> fdist['what']

183

>>> fdist.freq('what')

0.0195283320883577

>>> fdist['here']

181

>>> fdist.freq('here')

0.019314907693949418