Administrivia

- Assume you have properly installed nltk and nltk_data
nltk book: preface


• Why Python?
  • "We chose Python because it has a shallow learning curve, its syntax and semantics are transparent, and it has good string-handling functionality."

```python
>>> for line in open("file.txt"):
...     for word in line.split():
...         if word.endswith('ing'):
...             print(word)
```

• Text Processing Services: [https://docs.python.org/3/library/text.html#textservices](https://docs.python.org/3/library/text.html#textservices)
  • `str.split(sep=None, maxsplit=-1)`
  • `str.endswith(suffix[, start[, end]])`
nltk book: Language Processing and Python


```python
>>> from nltk.book import *

*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G. K. Chesterton 1908

>>> text1
<Text: Moby Dick by Herman Melville 1851>

>>> text2
<Text: Sense and Sensibility by Jane Austen 1811>
```
>>> text1.concordance("monstrous")
Displaying 11 of 11 matches:
ong the former , one was of a most monstrous size . .... This came towards us ,
ON OF THE PSALMS . " Touching that monstrousbulk of the whale or ork we have r
ll over with a heathenish array of monstrous clubs and spears . Some were thick
d as you gazed , and wondered what monstrous cannibal and savage could ever hav
that has survived the flood ; most monstrous and most mountainous ! That Himmel
they might scout at Moby Dick as a monstrous fable , or still worse and more de
th of Radney ."" CHAPTER 55 Of the monstrous Pictures of Whales . I shall ere l
ing Scenes . In connexion with the monstrous pictures of whales , I am strongly
er to enter upon those still more monstrous stories of them which are to be fo
ght have been rummaged out of this monstrous cabinet there is no telling . But
of Whale - Bones ; for Whales of a monstrous size are oftentimes cast up dead u
```
monstrous

contemptible

```python
>>> text1.concordance("monstrous")
Displaying 11 of 11 matches:
ong the former, one was of a most monstrous size . ... This came towards us ,
ON OF THE PSALMS . " Touching that monstrous bulk of the whale or ork we have r
il over with a heathenish array of monstrous clubs and spears . Some were thick
ed as you gazed , and wondered what monstrous cannibal and savage could ever hav
that has survived the flood ; most monstrous and most mountainous ! That Himnal
they might scout at Moby Dick as a monstrous fable , or still worse and more de
th of Radney ." CHAPTER 55 Of the Monstrous Pictures of Whales . I shall ere l
ing Scenes . In connexion with the monstrous pictures of whales , I am strongly
ere to enter upon those still more monstrous stories of them which are to be fou
ght have been rummaged out of this monstrous cabinet there is no telling . But
of Whale - Bones ; for Whales of a monstrous size are oftentimes cast up dead u
```

```python
>>> text1.concordance("contemptible")
Displaying 4 of 4 matches:
, we shall find they will appear contemptible in the comparison . The whale is
God ? Miserable man ! Oh ! most contemptible and worthy of all scorn ; with s
these pig - fish are a noisy , contemptible set , mostly lurking in the molt
harm , it may possibly be of no contemptible advantage ; considering that oil
```
monstrous

```python
>>> text1.concordance("monstrous")
Displaying 11 of 11 matches:
ong the former, one was of a most monstrous size . ... This came towards us,
ON OF THE PSALMS . " Touching that
monstrous bulk of the whale or ork we have r
il over with a heathenish array of
monstrous clubs and spears . Some were thick
ed as you gazed , and wondered what
monstrous cannibal and savage could ever hav
that has survived the flood ; most monstrous and most mountainous ! That Himnai
they might scout at Moby Dick as a monstrous fable , or still worse and more de
th of Radney .'" CHAPTER 55 Of the Monstrous Pictures of Whales . I shall ere 1
ing Scenes . In connexion with the
monstrous pictures of whales , I am strongly
er to enter upon those still more monstrous stories of them which are to be fo
ght have been rummaged out of this monstrous cabinet there is no telling . But
of Whale – Bones ; for Whales of a monstrous size are oftentimes cast up dead u
```
nltk book: Language Processing and Python

Inaugural Presidential Addresses

text4.dispersion_plot(['citizens', 'democracy', 'freedom', 'duties', 'America'])
nltk book: Language Processing and Python

```python
>>> text4.generate()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: generate() missing 1 required positional argument: 'words'
>>> text3.generate()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: generate() missing 1 required positional argument: 'words'
```
1.4 Counting Vocabulary

- `<text>` placeholder for some text object
- `<word>` placeholder for a word

1. `len(<text>)`  
   - word count
2. `set(<text>)`  
   - no duplicate words
3. `len(set(<text>))`  
   - no. of different words
4. `len(set(<text>)) / len(<text>)`  
   - lexical diversity
5. `<text>.count(<word>)`  
   - no. of times `<word>` occurs in `<text>`
6. `100 * <text>.count(<word>) / len(<text>)`  
   - % of `<text>` taken up by `<word>`
nltk book: Language Processing and Python

• Defining functions

```python
>>> def lexical_diversity(text):
...     return len(set(text)) / len(text)
...

>>> def percentage(count, total):
...     return 100 * count / total
...

>>> lexical_diversity(text3)
0.06230453042623537
>>> lexical_diversity(text5)
0.13477005109975562
>>> percentage(4, 5)
80.0
>>> percentage(text4.count('a'), len(text4))
1.4643016433938312
>>>```