Text Analysis with NLTK

Week 1

2 November 2020

CDCS Python Course Series

with Lucy Havens

Course Structure

Anticipate about ~7 hours/week

- 2 course meetings per week
    - 10:00 - 11:00 AM BST Mondays
    - 10:00 - 11:00 AM BST Fridays
- 1 assignment per week ~2 hours
- Office hours on Wednesdays for 30 minutes per participant
- Independent learning ~2 hours

Teams for introductions, meetings, office hours, questions, files

Course Topics

Text analysis

Analyzing unstructured data

Natural Language Toolkit (NLTK)

A Python library for text analysis

More Python Courses

Course 3: Network Analysis and Data Visualization

30th November - 11th December

Course 3 in the Python series will use the same course structure as this course!

Instructor Introduction

- Pursuing a PhD in the School of Informatics ILCC
- AMSc Design Informatics, B.S. Information Systems
- Taught myself programming and data science skills outside courses using online resources
- Please share feedback on the course!

For Participants

- Introduce material for you to review in greater depth on your own
- I’ll direct you to further resources if you’d like to go beyond material covered in each week’s assignment
- Course meetings won’t be recorded
  - Three strike policy
  - Please let me know in advance if you cannot attend!
- Office hours: questions about assignments, your own projects
  - Chat with me on Teams to schedule

For Participants, continued

We’ll be using Jupyter Notebooks

2 options:

A. Use the **Language and Machine Learning Notebook** with Noteable: https://www.ed.ac.uk/information-services/learning-technology/noteable/accessing-noteable
B. Install to your computer to run locally: https://jupyter.org/install

DEMO

Further Resources

- Noteable User Guide: https://noteable.edina.ac.uk/user_guide/#hide_ge_7
- Jupyter Notebooks, Noteable: https://github.com/edina/Exemplars2020/blob/master/TeachingDocs/Tutorials/UsingNoteableBeginner.ipynb
- Jupyter Notebooks: https://glam-workbench.github.io/getting-started/
- Python: https://programminghistorian.org/en/lessons/introduction-and-installation

NLTK

Natural Language Toolkit

Natural language = human language = unstructured data

NLTK

Examples of data sources for natural language:

- Books
- Newspapers
- Magazines
- Websites
- Transcriptions of audio (i.e. interview, movie dialogue)
- Social media

Always read the licensing/copyright information and terms of use!

What kinds of questions can you ask when you can use a programming language to study hundreds, thousands, or even millions of pages of digital text?

What kinds of questions can you ask when you can physically hold and look at a printed text, be it an original publication or later edition of the text?

Distant Reading

vs.

Close Reading

NLTK

Built-in methods for getting familiar with a text:

```
.concordance(“word”, lines=20)
.similar(“word”)
.common_contexts([“list”, “of”, “words”])
.dispersion_plot([“list”, “of”, “words”])
```

Reference: https://www.nltk.org/book/ch01.html

NLTK

Tokens vs. words

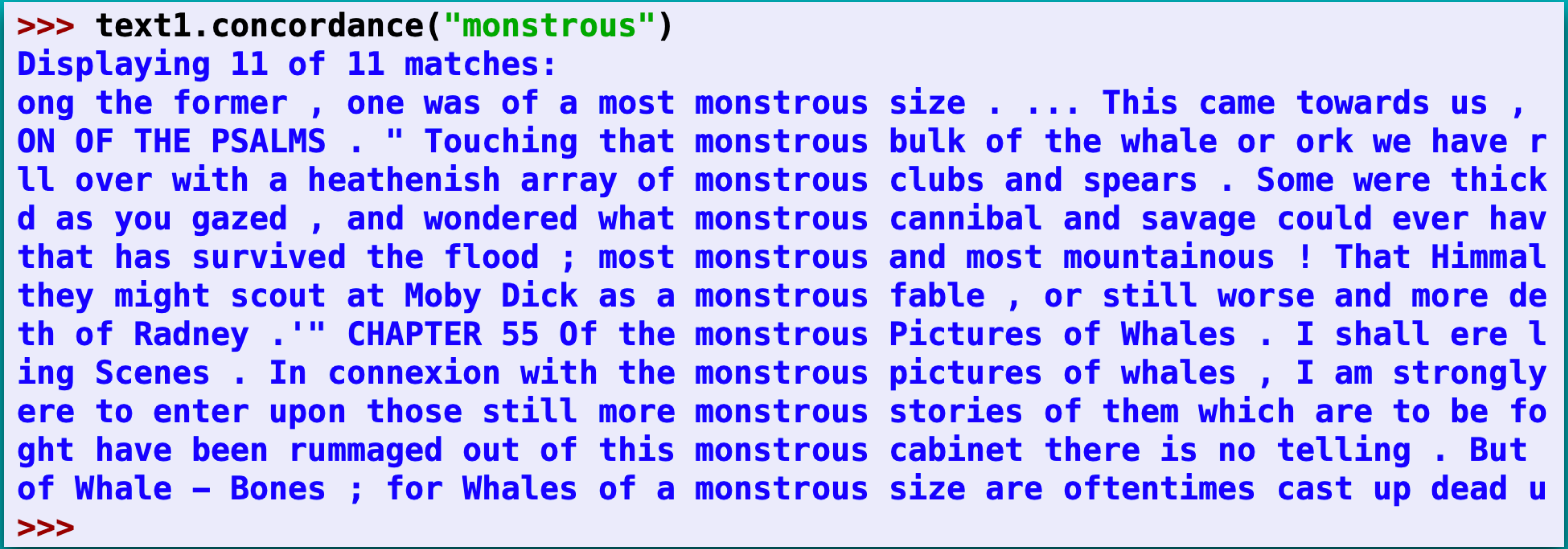
Digitized vs. digital

Normalisation for what?

NLTK

See the word in context with `.concordanace()`

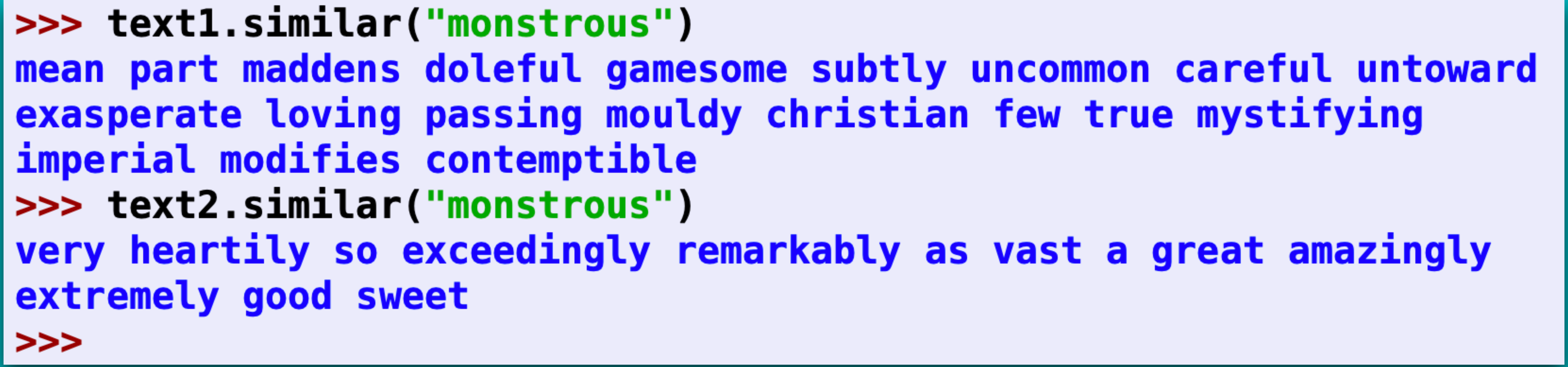
Note: the `lines=__` *parameter is optional input*



```
>>> 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
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 Himmal
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 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

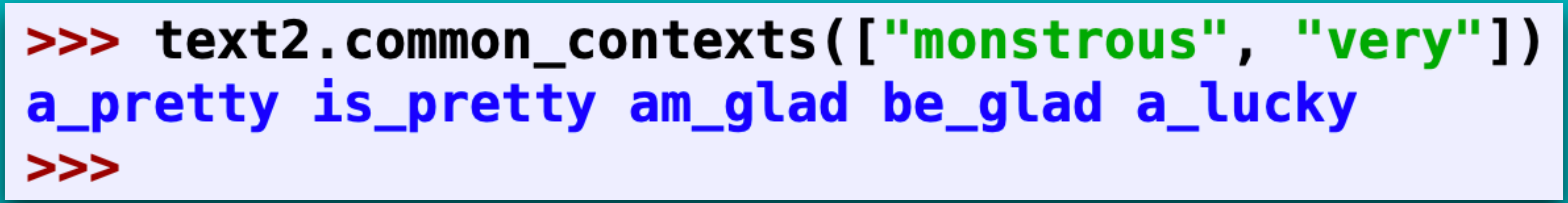
See what appears in similar contexts as the input with `.similar()`



```
>>> text1.similar("monstrous")
mean part maddens doleful gamesome subtly uncommon careful untoward
exasperate loving passing mouldy christian few true mystifying
imperial modifies contemptible
>>> text2.similar("monstrous")
very heartily so exceedingly remarkably as vast a great amazingly
extremely good sweet
>>>
```

NLTK

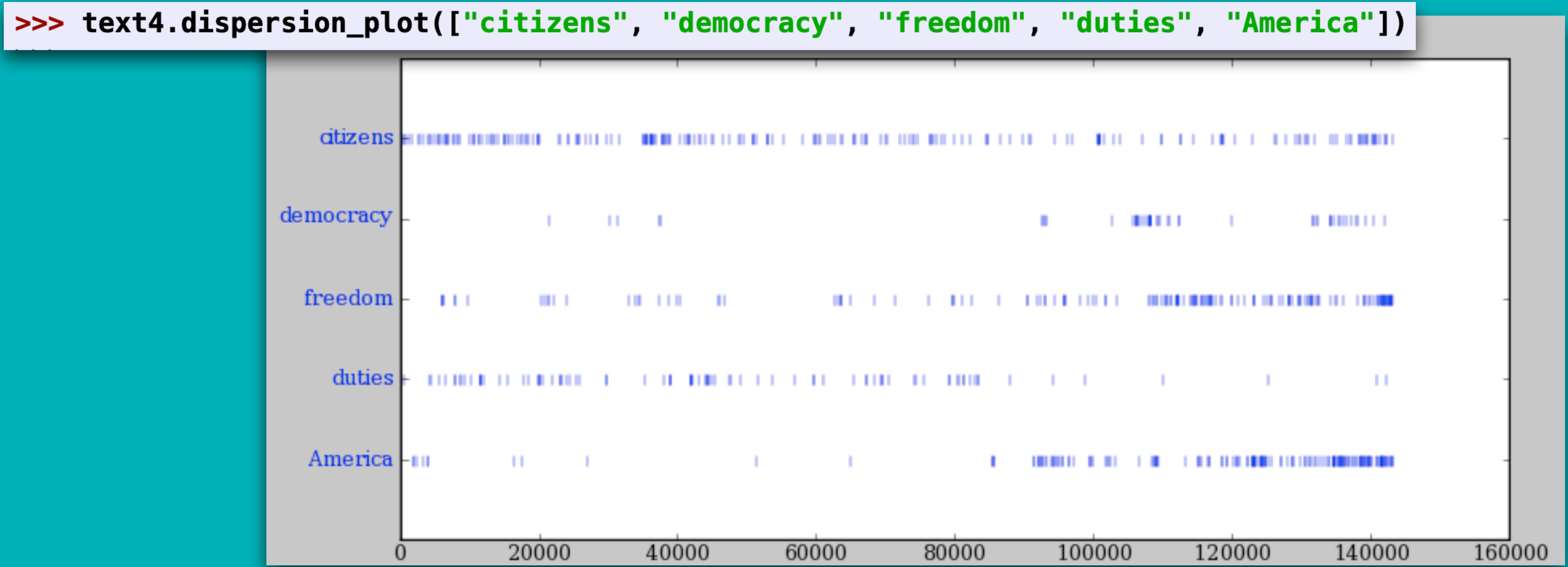
See the context **multiple** words share with `.common_contexts()`



```
>>> text2.common_contexts(["monstrous", "very"])
a_pretty is_pretty am_glad be_glad a_lucky
>>>
```

NLTK

See words’ frequency across a text with `.dispersion_plot()`



```
>>> text4.dispersion_plot(["citizens", "democracy", "freedom", "duties", "America"])
```

DEMO

NLTK

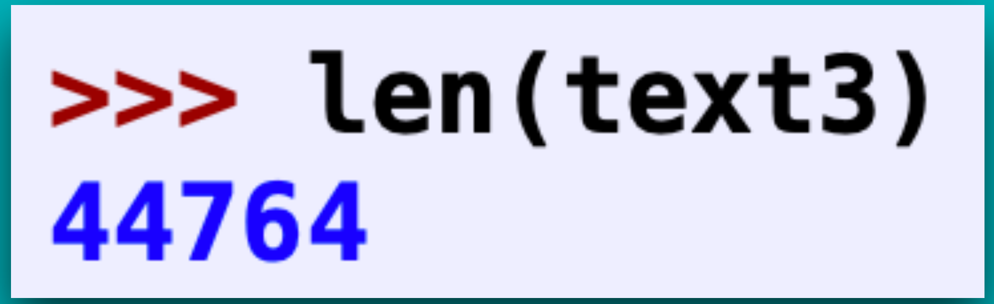
Built-in functions and methods for summarizing a text:

```
len(text)
sorted(vocab_of_text)
.count(“word”)
```

Reference: https://www.nltk.org/book/ch01.html

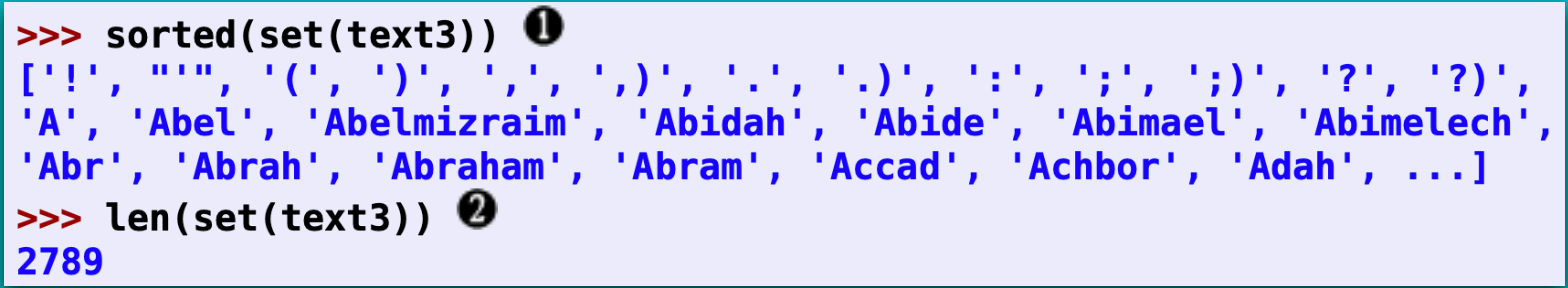
NLTK

Length of a text - total number of tokens



```
>>> len(text3)
44764
```

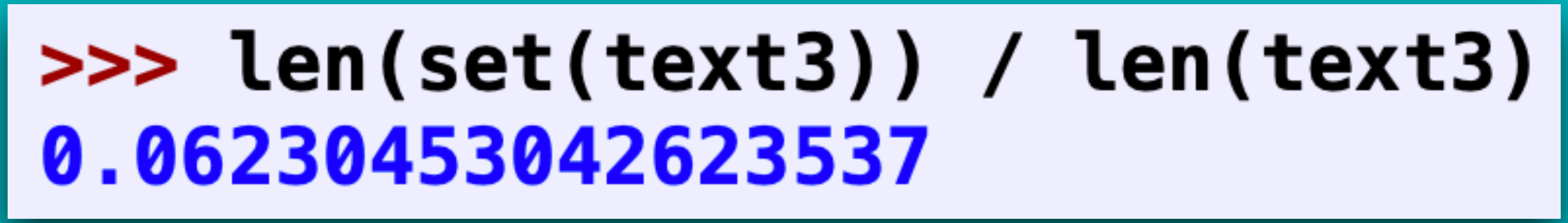
Length of a text’s **vocabulary** - total number of **unique** tokens



```
>>> sorted(set(text3)) ❶
['!', "'", '(', ')', ',', ',)', '.', '.)', ':', ';', ';)', '?', '?)',
'A', 'Abel', 'Abelmizraim', 'Abidah', 'Abide', 'Abimael', 'Abimelech',
'Abr', 'Abrah', 'Abraham', 'Abram', 'Accad', 'Achbor', 'Adah', ...]
>>> len(set(text3)) ❷
2789
```

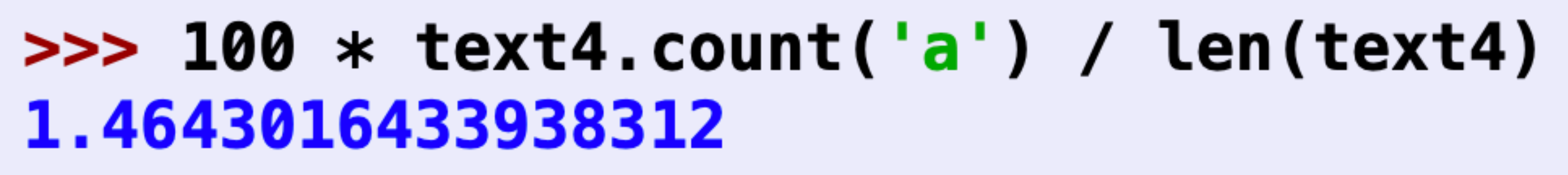
NLTK

Lexical diversity - ratio of unique tokens to total tokens



```
>>> len(set(text3)) / len(text3)
0.06230453042623537
```

Percentage of a text that a token accounts for



```
>>> 100 * text4.count('a') / len(text4)
1.4643016433938312
```

DEMO

Assignment

Prework for those who haven’t used Python or Jupyter Notebooks before

Helpful Resources - **for reference**, not required reading

Steps 1-5: Independent learning
Step 6: A tutorial to complete in your own Jupyter Notebook

Go Further - **optional** reading assignment

Course2Week1 PDF file uploaded to the “Week 1 Assignment” channel of Teams

Thanks everyone!

Next course meeting: Friday, 10:00-11:00 AM BST

Office hours available on Wednesday (30 minutes)

To schedule, please message me on Teams!