

February 12-19. 2025

Centre for Data, Culture & Society



INTRODUCTION TO TEXT

ANALYSIS WITH PYTHON

Instructor: Xandra Dave Cochran

Course Structure

Anticipate about ~7 hours/week

- .2 hour course meeting, 2-4 Wednesday

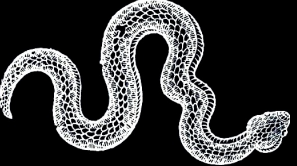
- .1 assignment per week, ~2 hours

- .Office hours on request

- .Independent learning, ~2 hours

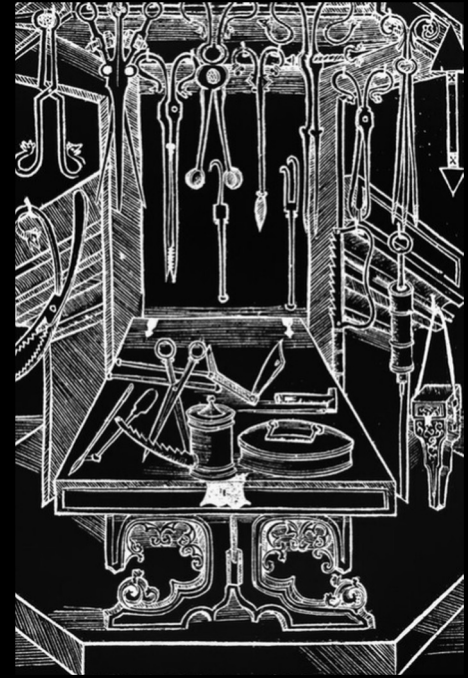
- .Teams for introductions, meetings, office hours, questions, files

- Text Analysis – analysing unstructured data

-  Python

- Regular Expressions

- Natural Language Toolkit (NLTK)



Introductions!

Why are you interested in text analysis?

Have you used Python before?

Have you used Jupyter Notebooks before?

Have you used Regular Expressions before?

Have you used NLTK before?

Participant Expectations

Friday classes are introductions to material

Assignments will be given on Monday

Classes are not recorded but all class materials will be uploaded to Teams

Please let me know in advance if you cannot attend!

Message me on Teams to schedule office hours for questions

Course Software

Jupyter Notebooks / Jupyterlabs

- With Notable

<https://www.ed.ac.uk/information-services/learning-technology/noteable/accessing-noteable>

After logging into MyEd: <https://noteable.edina.ac.uk/launch>

- With Google Colab

<https://colab.research.google.com>

- Locally (install with pip/pip3 or conda)

DEMO



Further Resources

Noteable User Guide

https://noteable.edina.ac.uk/user_guide/#hide_ge_7

Jupyter Notebooks in 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>

Natural Language Toolkit

Natural language = human language =
"unstructured" data

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!

Why use NLTK?

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?

“Distant reading”

NLTK isn't everything

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?

“Close reading”

Book history

NLTK Terminology

Tokens vs. words

Digitized vs. digital

Normalization (a.k.a. standardization)

Document vs. corpus vs. corpora

Summarizing a Text

Built-in functions include:

```
len(text)
```

```
sorted(vocabulary_of_text)
```

NLTK Text methods include:

```
Text.count(word)
```

Reference:

<https://www.nltk.org/book/ch01.html>

DEMO



Getting to know a text

NLTK Text methods include:

```
Text.concordance("word", lines=20)
```

```
Text.similar("word")
```

```
Text.common_contexts(["list", "of", "words"])
```

```
Text.dispersion_plot(["list", "of", "words"])
```

DEMO



The Building Blocks

Tokenization - words/punctuation, sentences

Normalization

Stemming and lemmatizing

Frequency counts

Part-of-speech tagging

DEMO



Finding Text Sources

Libraries - NLS Data Foundry (data.nls.uk)

Project Gutenberg (gutenberg.org)

Hathi Trust Digital Library (hathitrust.org)

Websites - Internet Archive (archive.org)'s
Wayback Machine, UK Web

archive (webarchive.org.uk)

Newspaper archives (universities often
subscribe to them!)

Research with NLTK

Who is named in a text?

What places are named in a text?

Chunking and Named Entity Recognition

How does the vocabulary of an author change over time?

Lexical Diversity

Research with NLTK

Who is named in a text?

What are the common themes throughout a corpus?

Topic Modeling

What attitudes are expressed in a corpus?

Sentiment Analysis

What words occur near each other throughout a corpus? How does the meaning of a word change over time?

Word Embeddings

Next Week

Research with NLTK on a corpus

NLTK with pandas (for tabular data)

NLTK with Altair (for data visualization)

Regular Expression practice

Cleaning messy text

Resources for more text analysis practice

Further Resources

CDCS

- Digital Method of the Month on Text Analysis
- Training Pathway for Text Analysis

Thanks Everyone!

Next class: Wednesday 19th, 2-4PM

Please message me on Teams for office hours!