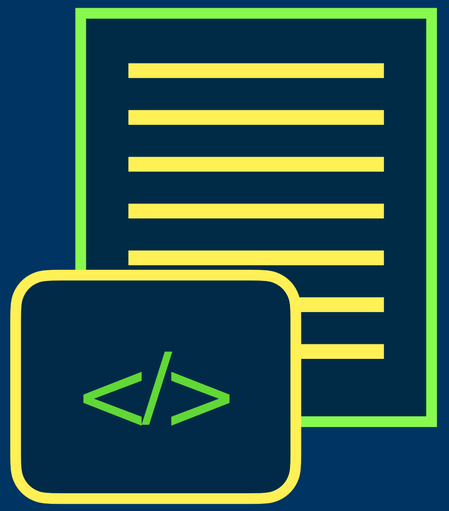

Text Analysis with NLTK

March 30-April 8, 2022

Week 1

Centre for Data, Culture & Society

Instructor: Lucy Havens

Quick Recap

NLTK: Python library for text analysis

Text = natural (human) language = unstructured data

Built-in functions and methods for getting familiar with & summarizing a text

Assignment

How did it go?

Questions?

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!)


DEMO!

Important Considerations

In what context what your document/corpus produced?

Who is represented in the document/corpus? Who is not represented?

Text data, especially digitized text data, is messy. Do not be afraid!


EXAMPLE!

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

More Research with NLTK

What are 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

Use NLTK on a text of your choice!

Thanks everyone!

Next class: Wednesday, 10-11 AM

Please message me on Teams for office hours!

Further Resources

CDCS

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