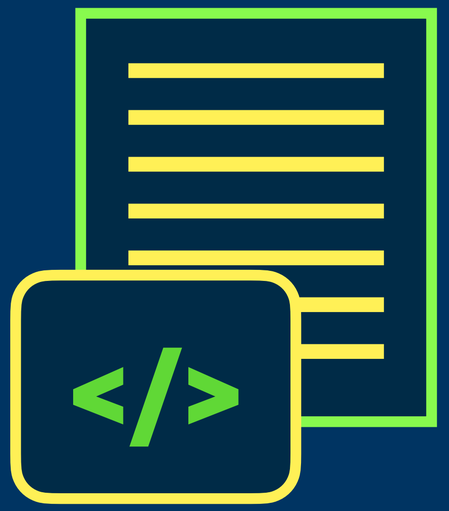


Text Analysis with NLTK

March 30-April 8, 2022

Week 2

Centre for Data, Culture & Society

Instructor: Lucy Havens

Week 2 Topics

Class 3 (Today):

Research with NLTK on a corpus

- NLTK with pandas (for tabular data)
- NLTK with Altair (for data visualization)

Class 4:

- Regular Expression practice
- Cleaning messy text
- Resources for more text analysis practice

Assignment

How is it going?

Questions?

Research with NLTK

Corpus: Lewis Grassic Gibbon First Editions

Questions:

What are the most common words in the corpus?

What are the most common words in one book from the corpus?

How does the word choice of the author change from one book to another?

Lexical diversity = count of unique words / count of all words

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!)

Lewis Grassic Gibbon First Editions

Original OCR: no clean-up

4,685 ALTO XML files at page level

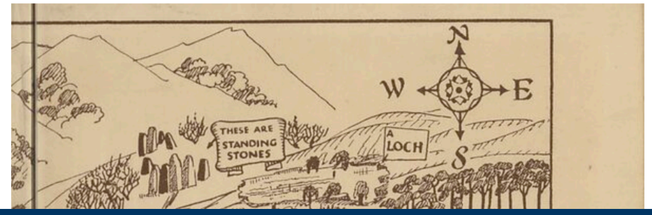
4,685 image files

METS metadata files at item level

145,457 lines and 1,237,615 words

Covers years 1928-1934

The dataset consists of the first editions of sixteen books published by James Leslie Mitchell (1901-1935) during his lifetime under his birth name Mitchell and the pseudonym Lewis Grassic Gibbon. The books were published between 1928 and 1934 and include novels, collections of short stories, biographies and accounts of exploration. Among the titles is the trilogy 'A Scots Quair' which consists of the novels 'Sunset Song' (1932), 'Cloud Howe' (1933) and 'Grey Granite' (1934). Published



Reference: https://data.nls.uk/data/digitised-collections/lewis-grassic-gibbon-first-editions/

Exploring Lewis Grassic Gibbon First Editions

Get started with 'Lewis Grassic Gibbon First Editions' with this Jupyter Notebook.

Whether or not you have experience programming or working with data, this Notebook will give you a starting point for analysing digitised text. Using Python and several of its libraries, including Pandas and Natural Language Toolkit (NLTK), the Notebook demonstrates how to:

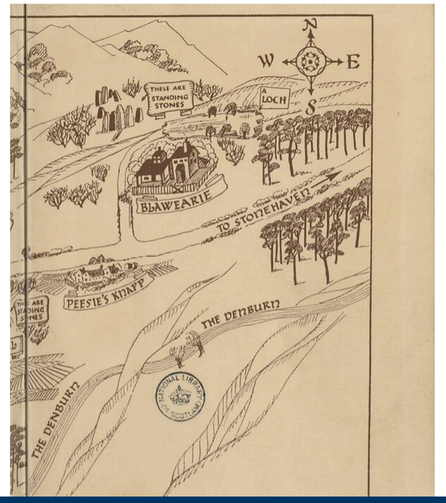
– Load a folder of text (.TXT) files as a corpus

– Ask questions about the words and sentence structure of all of Gibbon's works in the collection

– Group Gibbon's works into subsets, such as the three books in the trilogy, A Scot's Quair

Questions this Notebook can help you begin investigating include:

– What are the most common words in Gibbon's works?



Reference: https://data.nls.uk/tools/jupyter-notebooks/exploring-lewis-grassic-gibbon-first-editions/

LET'S CODE!

Next Class: More Research with NLTK

Regular Expression practice

Cleaning messy text

Resources for more text analysis practice

Thanks everyone!

Next (last!) class: Friday, 10-11 AM

Message on Teams for if you'd like to schedule office hours