

12<sup>th</sup>–19<sup>th</sup> February 2025



# INTRODUCTION TO TEXT ANALYSIS WITH PYTHON

Instructor: Xandra Dave Cochran



[www.cdcs.ed.ac.uk](http://www.cdcs.ed.ac.uk)

## Week 2 Topics

- Corpus Research with NLTK
- Data Visualisation
- Regular Expressions
- Data Cleaning

Sir

It is a great blessing and happiness to a nation  
when the King employeth such a man as you are to do  
and do for him who I am perswaded his the awe and fear  
of God on him. Job was a just man and a perfect and the  
cause that he know not he feared out to deliver  
the poor and oppressed and him that had none to help  
him, a pattern for on in your office. I have the Honour  
to be your Relation and I know you have much  
interest with Lord Greange if you can make Peace for  
me you know the promises that is to the Peace make  
of losing my husband to much, he knowes very well  
that he was my idol and now God his made him  
a rode to scourgeth me. \* \* \* \* \*

\* \* \* \* \*

\* \* \* \* \* much fuller account then this and he wrote it down. I have given to  
you much more to tell then this when this comes to you if you hear of  
it.





## Assignment

- How's it going?
- Any questions?

Sir

It is a great blessing and happiness to a nation  
when the King employeth such a man as you are to do  
and do for him who I'm perswaded his the awe and fear  
of God on him. Job was a just man and a perfect and the  
cause that he know not he feared out to deliver  
the poor and oppressed and him that had none to help  
him, a pattern for on in your office. I have the Honour  
to be your Relation and I know you have much  
interest with Lord Greange if you can make Peace for  
me you know the promises that is to the Peace make  
of losing my husband to much, he knowes very well  
that he was my idol and now God his made him  
a rode to scourgeth me. \* \* \* \* \*

\* \* \* \* \*

\* \* \* \* \* much fuller account then this and he wrote it down. I have given to  
ve much more to tell then this when this comes to you if you hear of it



## Research with NLTK

Corpus: Lewis Grassie Gibbon First Editions (National Library of Scotland)

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](http://data.nls.uk))
- Project Gutenberg ([gutenberg.org](http://gutenberg.org))
- Hathi Trust Digital Library ([hathitrust.org](http://hathitrust.org))
- Websites - Internet Archive ([archive.org](http://archive.org))'s Wayback Machine, UK Web archive ([webarchive.org.uk](http://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



The background is a collage of historical elements. It features a circular portrait of a woman with dark hair and a pink flower in her hair, set against a red background. Below this, there is a circular image of a person in a red garment, possibly a historical figure, also set against a red background. The entire collage is overlaid on a parchment-like texture with handwritten text in cursive and printed text in a serif font. The text includes phrases like 'waiting here for me', 'the master of the sloop', 'they met in Scotos', 'George Som Ronald', 'with', 'and saw me on Sep 30 we came to the Isle', 'MacLeod', 'he', 'I', 'a much', 'time', 'a great miserie in the', 'society sent a minister here I have given him', 'and he wrt it down, you may [be] sure I have', 'his come to you if you hear I'm alive do me', 'all hast but if you hear I'm dead do what', and 'o/t 93'.

# Let's Code!

# Regular Expressions (RegEx)

Pattern matching for the string (`str`) data type

Documentation:

- Intro:
  - <https://docs.python.org/3/howto/regex.html#regex-howto>
- Python re module:
  - <https://docs.python.org/3/library/re.html>
- For practice:
  - <https://regex101.com/>
  - <http://pythex.org/>
    - Check out the Pythex cheat sheet!





## Regular Expressions (RegEx)

To use in a Jupyter Notebook:

```
import re
```

To find patterns (2 ways):

```
re.findall("regex_pattern", "string_to_search")
```

```
["list", "of", "all", "matches", "found"]
```

```
p = re.compile("regex_pattern")
```

```
p.findall("string_to_search")
```

```
["list", "of", "all", "matches", "found"]
```



`s.replace('a', 'b')` - replace **a** with **b** in string **s**





## Beware!

Be careful not to spend all your time cleaning data! It may be useful to time-box this task so that you do not run out of time for the actual analysis work.

Alternatively, you could consider investing in or applying for funds to manually correct your text. This will yield more accurate results than programmatic methods.





The background is a collage of historical elements. It features a circular portrait of a woman with dark hair and a pink flower in her hair, set against a red background. Below this, there is a circular image of a person in a red garment, possibly a historical figure, also set against a red background. The entire collage is overlaid on a parchment-like texture with handwritten text in cursive and printed text in a serif font. The text includes phrases like 'waiting here for me', 'the master of the sloop', 'they met in Scotos', 'George Som Ronald', 'with', 'and saw me on Sep 30 we came to the Isle', 'MacLeod', 'he', 'I', 'a much', 'time', 'a great miserie in the', 'society sent a minister here I have given him', 'and he wrt it down, you may [be] sure I have', 'his come to you if you hear I'm alive do me', 'all hast but if you hear I'm dead do what', and 'o/t 93'.

# Let's Code!

# Text Analysis: NLTK & Beyond

## Training Pathway for Text Analysis

- <https://www.cdcs.ed.ac.uk/training/training-pathways/text-analysis-pathway>

## Sentiment Analysis Tutorial in Jupyter Notebooks

- Includes cleaning up text! Uses an NLTK tool called VADER for sentiment analysis
- <https://github.com/DCS-training/SentimentAnalysistimes>

## Text Analysis with Constellate

- Uses Jupyter Notebooks and libraries other than NLTK
- <https://github.com/ithaka/constellate-notebooks>

## Topic Modelling with BERT

- Uses Jupyter Notebooks and BERT, a transformer-based LLM. Will be delivered by me on 28<sup>th</sup> Feb–7<sup>th</sup> Mar this semester
- <https://www.cdcs.ed.ac.uk/events/intro-topic-modelling-feb25>

## W3Schools - Python, RegEx, and much more!

- <https://www.w3schools.com/>





# Thanks Everyone

**Please message me on Teams  
for office hours!**