

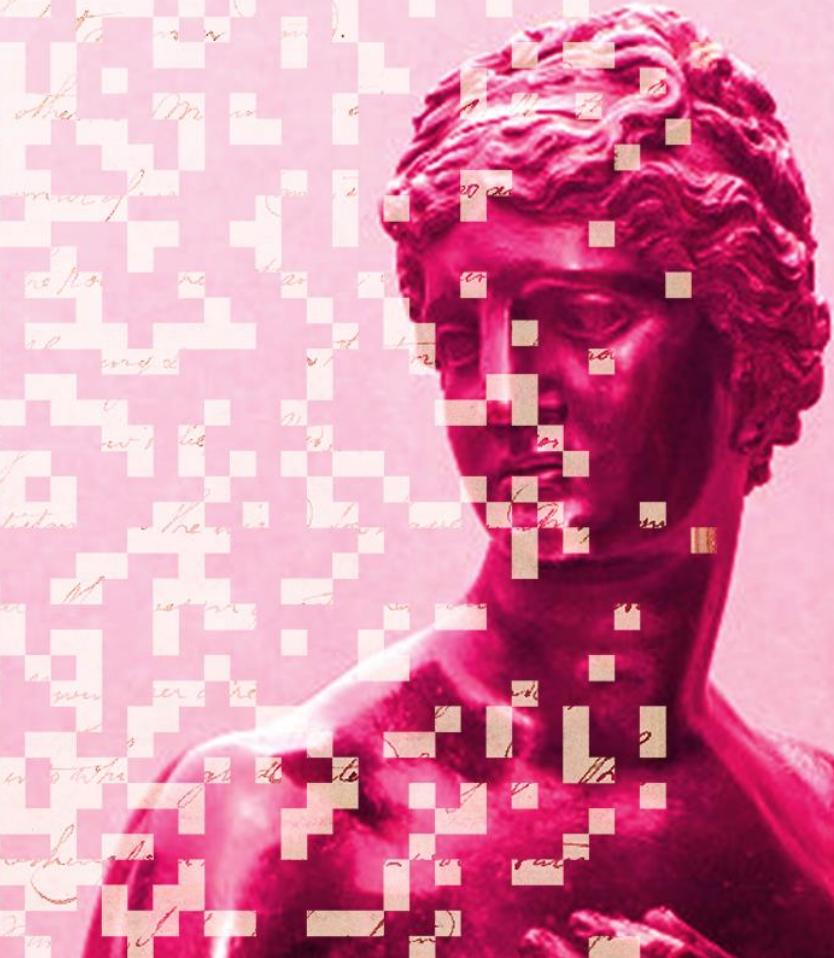


THE UNIVERSITY OF EDINBURGH  
Centre for Data, Culture & Society



# Data Culture & Society

@edCDCS





THE UNIVERSITY of EDINBURGH  
Centre for Data, Culture & Society



# Introduction to Text Analysis With Python

16 & 25 Feb 2026

Instructor: Joy Lan

Adapted from materials made by Xandra Dave Cochran

# Recap



- What is NLTK for?
- How many tokens are in the following sentence?  
I ate the cake and she ate the cake too.
- Can you name a form of normalisation of text for NLP? What does it do?





# 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 act  
and do for him who I'm perswadid his the aw and fear  
of God on him. Job was a just man and a perfect and blis-  
sful cause that he know not he feareched out to deliver  
the poor and oppressed and him that had none to helpe  
him, a patterne 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 loving my husband to much, he knowes very well  
that he was my idol and now God his meade him  
a rode to scourgeth me. \*



# Research with NLTK

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

Please Answer the following 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?
- Note: Lexical diversity = count of unique words / count of all words



# 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

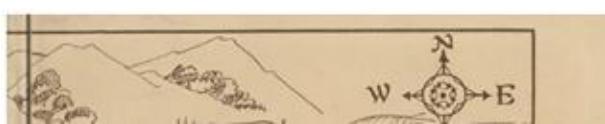


145,457 lines and  
1,237,615 words



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





# 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):

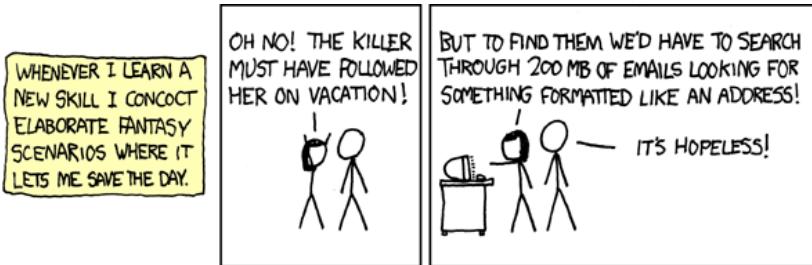
```
re.findall("regex_pattern", "string_to_search")
```

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

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

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

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



## Cleaning Messy Text

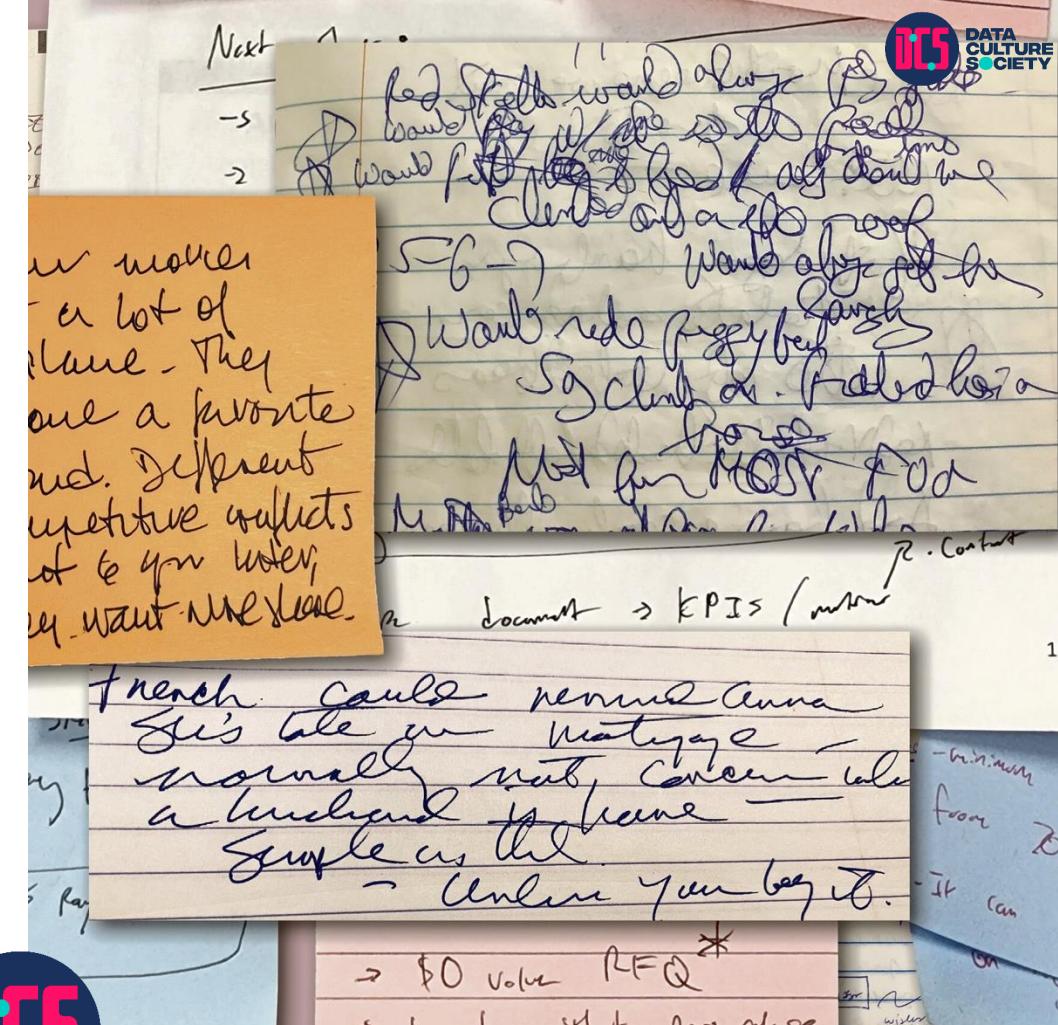
Remember, digitization is imperfect

- Includes OCR (optical character recognition)
- Includes HWT or HRT (handwriting recognition)

Besides Regular Expressions, you can use...

**s.strip()** - remove leading & trailing whitespace or  
input characters in string **s**

**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.





# 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!