

Week 2 Topics

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



It is agreat blessing and happiness to a natio when the Hing imployeth such a man as you are to chet and do for him who I'm personadid his the are and fear of God on Sum. job was a just man and a perfect and the cause that he know not he fearehed out to deliver The poor and oppressed and fine that had none to he Riem, affallerne for on in your office. I have the Honor to be your Relation and I know you have much interest with Lord Greange if you can make Peace for of loveing my husband to much, he knowed very we that he was my jobol and now god his made him

a vode lo Scourgeth me. * * * * * * * * *

Assignment

- How's it going?
- Any questions?

82 Hilda Julia Science

Sim

It is agreat blessing and happiness to a natio where the Hing imployeth such a man as you are to Act and do for him who I'm personadid his the are and fear of God on Sum. job was a just man and aperfect and the cause that he know not he fearehed out to deliver the poor and oppressed and him that had none to he liem, affallerne for on in your office. I have the Honor to be your Relation and I know you have much interest with Lord Greange if you can make Peace for of loveing my husband to much, he knowed very we that he was my job and now god his made him a vode lo Scourgeth me. * * * * * * * * *

much fullet account then this and he wroit it down you may four



Research with NLTK

Corpus: Lewis Grassic 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)
- 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!)



TOOLS









4,685 ALTO XML





4,685 image files



METS metadata files at item level



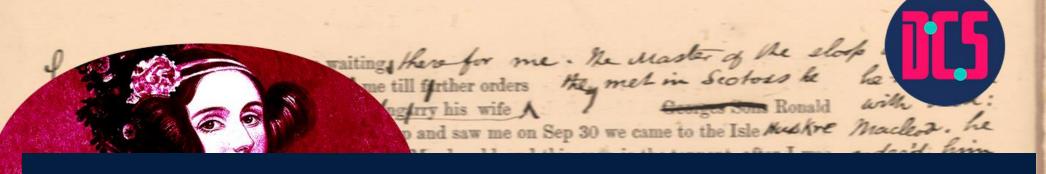
145,457 lines and 1,237,615 words



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!

ciety sent a minister here I have given him d he writ it down, you may be sure I have this come to you if you hear I'm alive do me all hast but if you hear I'm dead do what

0/t 0) 93





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!

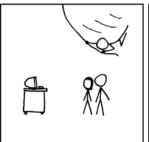




















Regular Expressions (RegEx)

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"]





BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!













Cleaning Messy Text

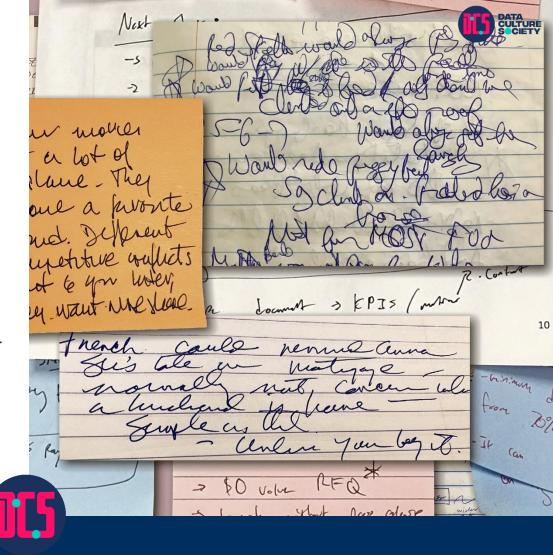
Remember, digitization is imperfect

- 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





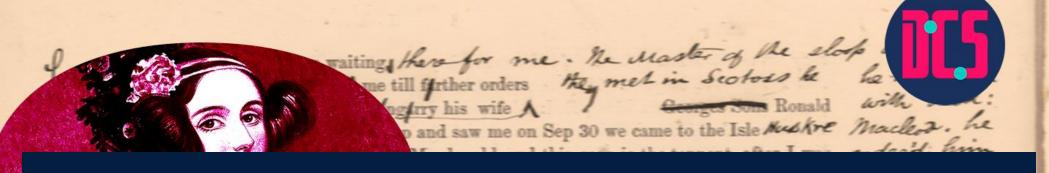




Beware!

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!

ciety sent a minister here I have given him d he writit down, you may be sure I have this come to you if you hear I'm alive do me all hast but if you hear I'm dead do what

0/t 0) 93





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 28th Feb-7th Marthis semester
- https://www.cdcs.ed.ac.uk/events/intro-topic-modelling-feb25

W3Schools - Python, RegEx, and much more!

https://www.w3schools.com/





Thanks Everyone

a man for a way with my money in Lim: 1734 Rod: sent of for the tangant of this Isle his name Alex. Macdonald to come to the Captain of

Please message me on Teams for office hours!

d he writ 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