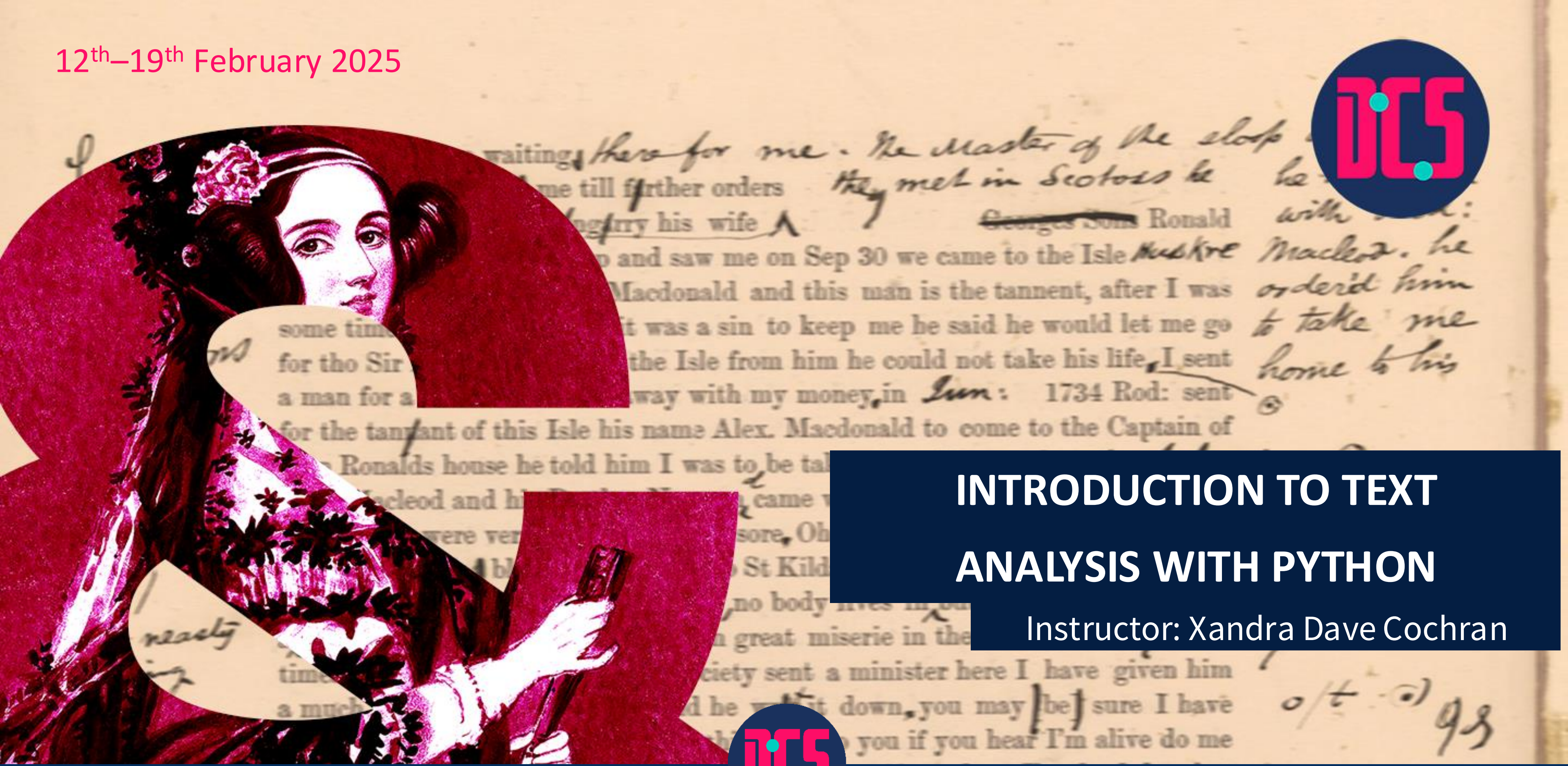
12th–19th February 2025



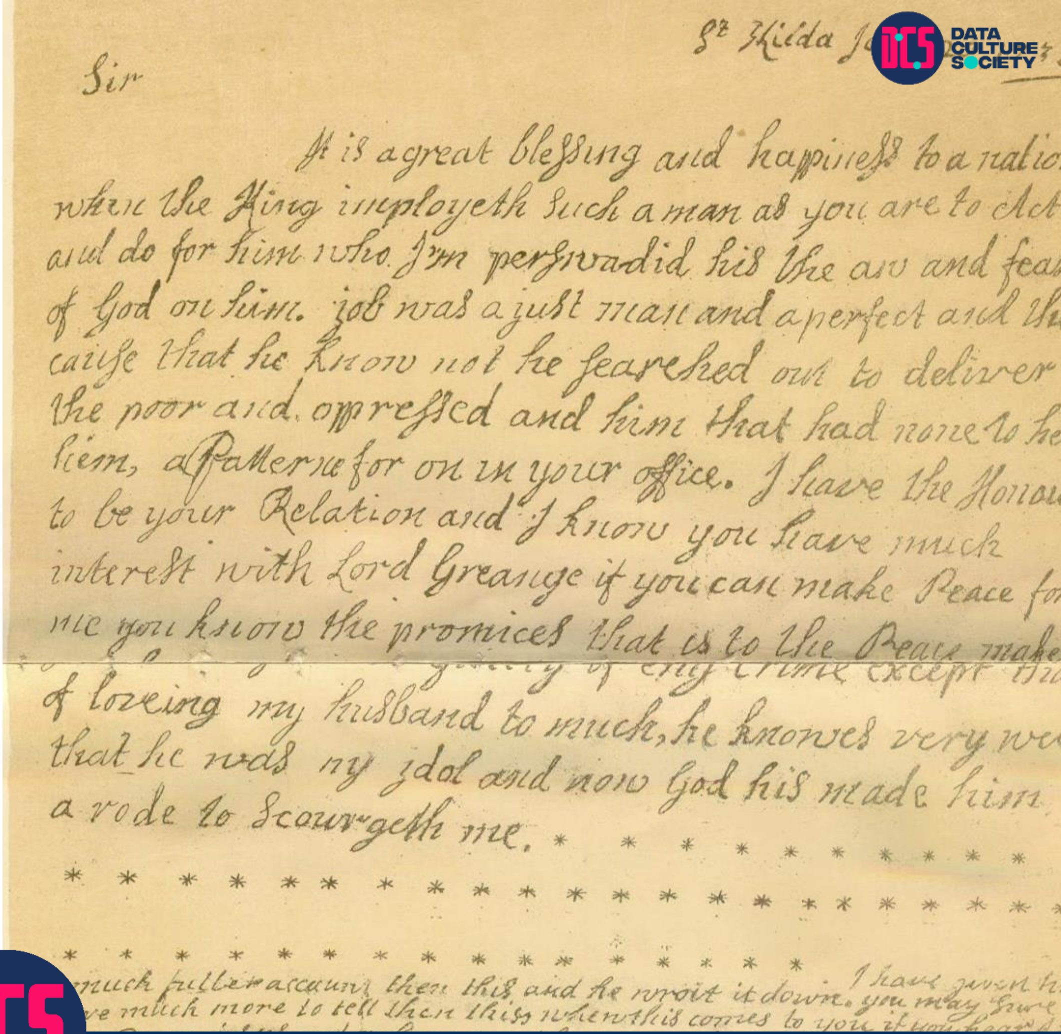
INTRODUCTION TO TEXT ANALYSIS WITH PYTHON

Instructor: Xandra Dave Cochran

www.cdcs.ed.ac.uk

Week 2 Topics

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



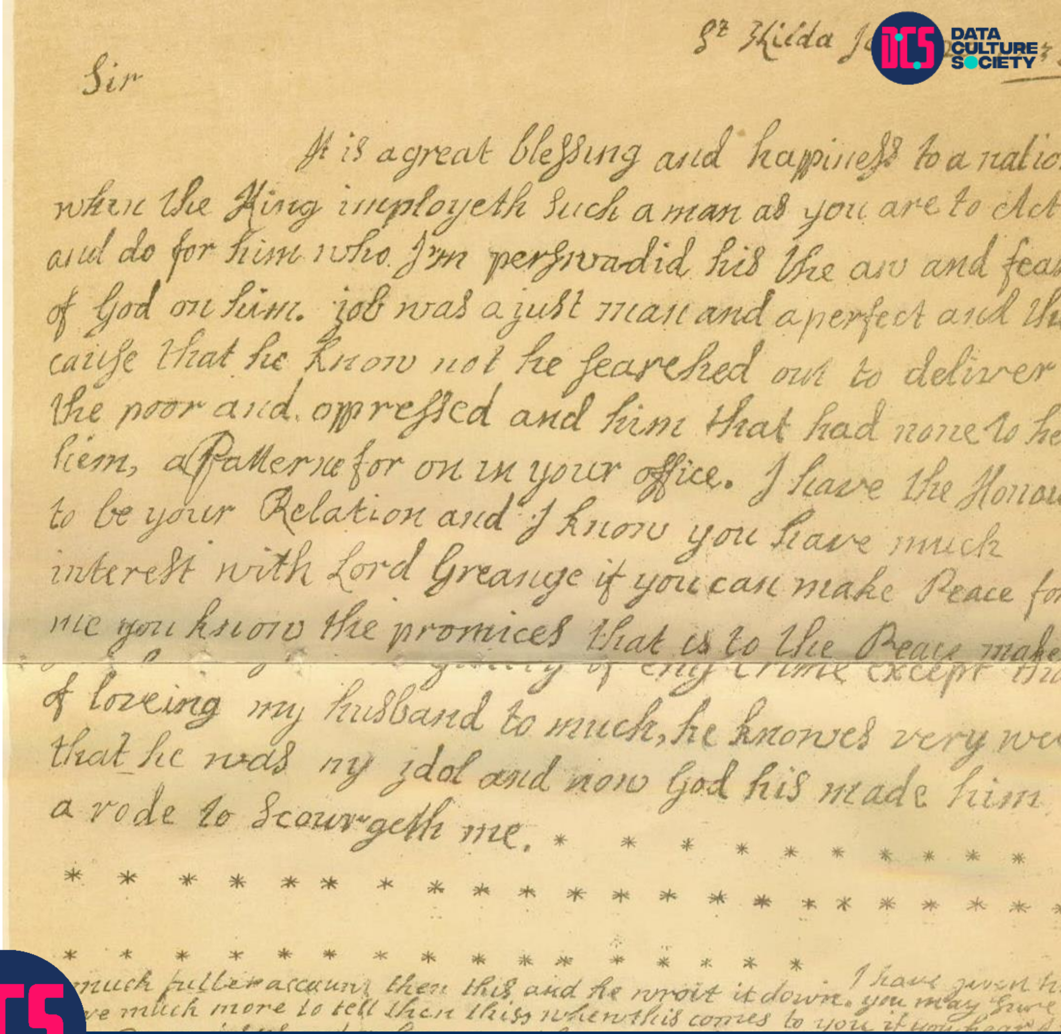
St Kilda Ja

Sir

It is agreat blessing and happiness to a natio
when the King imployeth such a man as you are to Act
and do for him who I'm perswadid his the aw and fea
of God on him. job was a just man and a perfect and th
caise that he know not he searched out to deliver
the poor and oppressed and him that had none to he
liem, a Patterne for on in your office. I have the Honou
to be your Relation and I know you have much
interest with Lord Greange if you can make Peace fo
me you know the promices that is to the Peace make
guilty of eny Crime except tha
of loveing my husband to much, he knowes very we
that he was my idol and now God his made him
a rode to Scourgeth me. * * * * * * * * * *
* * * * * * * * * * * * * * * * * * * *
* * * * * * * * * * * * * * * *
much fuller accaunt then this and he wroit it down. I have given hi
ve much more to tell then thiss when this comes to you you may sure

Assignment

- How’s it going?
- Any questions?



St Kilda Ja

Sir

It is agreat bleſsing and happineſs to a natio
when the King imployeth such a man as you are to Act
and do for him who I'm perſwadid his the aw and fea
of God on him. job was a juſt man and a perfect and th
caiſe that he know not he ſearched out to deliver
the poor and oppreſsed and him that had none to he
liem, a Patterne for on in your office. I have the Honou
to be your Relation and I know you have much
intereſt with Lord Greange if you can make Peace fo
me you know the promices that is to the Peace make
[illegible] gilty of eny crime except tha
of loveing my husband to much, he knowes very we
that he was my idol and now God his made him
a rode to Scourgeth me. * * * * * * * * * * *
* *
* * * * * * * * * * * * * * * * * *

much fuller accaunt then this and he wroit it down. I have given h
ve much more to tell then thiss when this comes to you you may Suwe if you hear

Research with NLTK

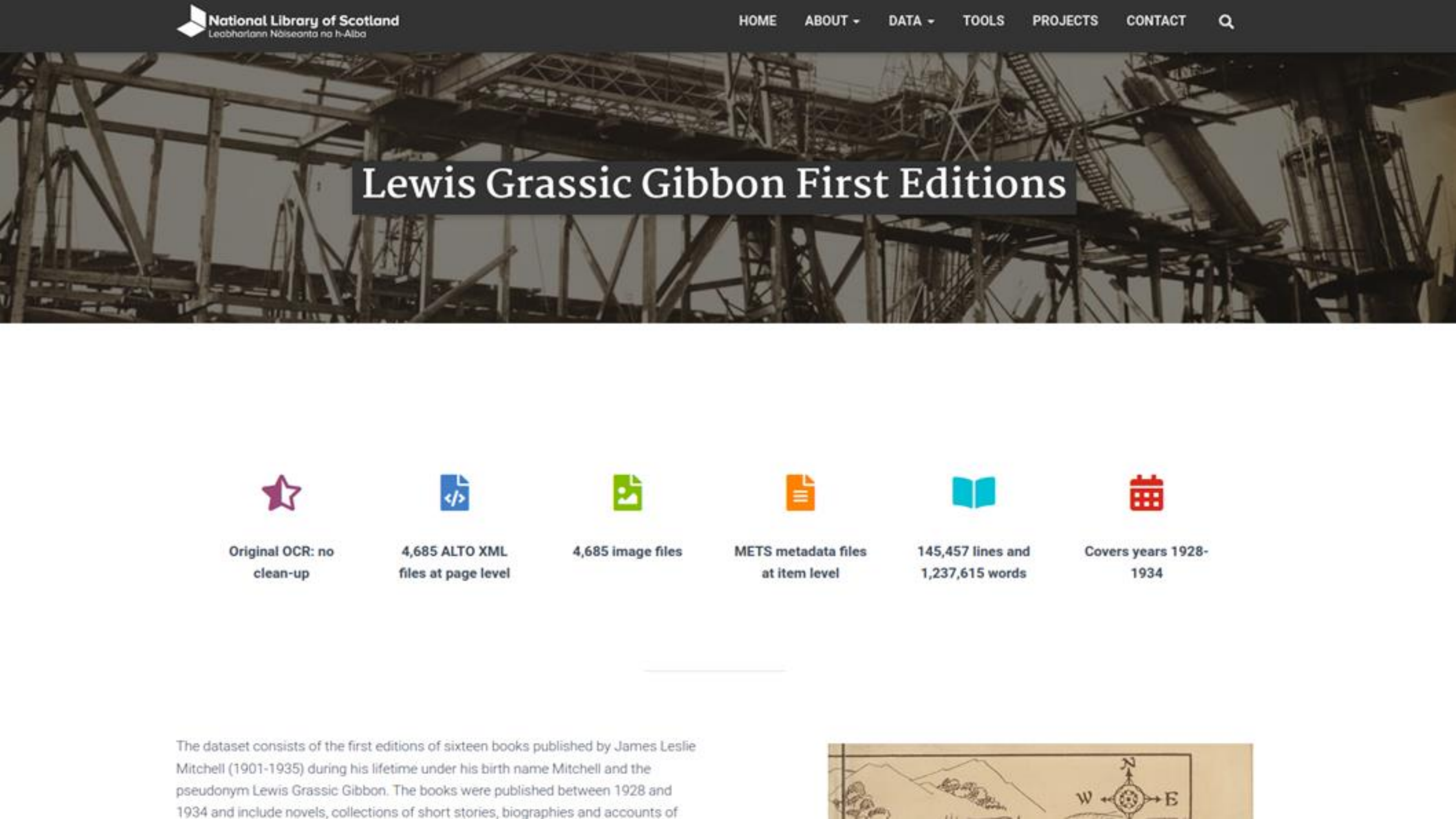
Corpus: Lewis Grassic Gibbon First Editions (National Library of Scotland)

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!)



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!

WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

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"]
```

WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

Cleaning Messy Text

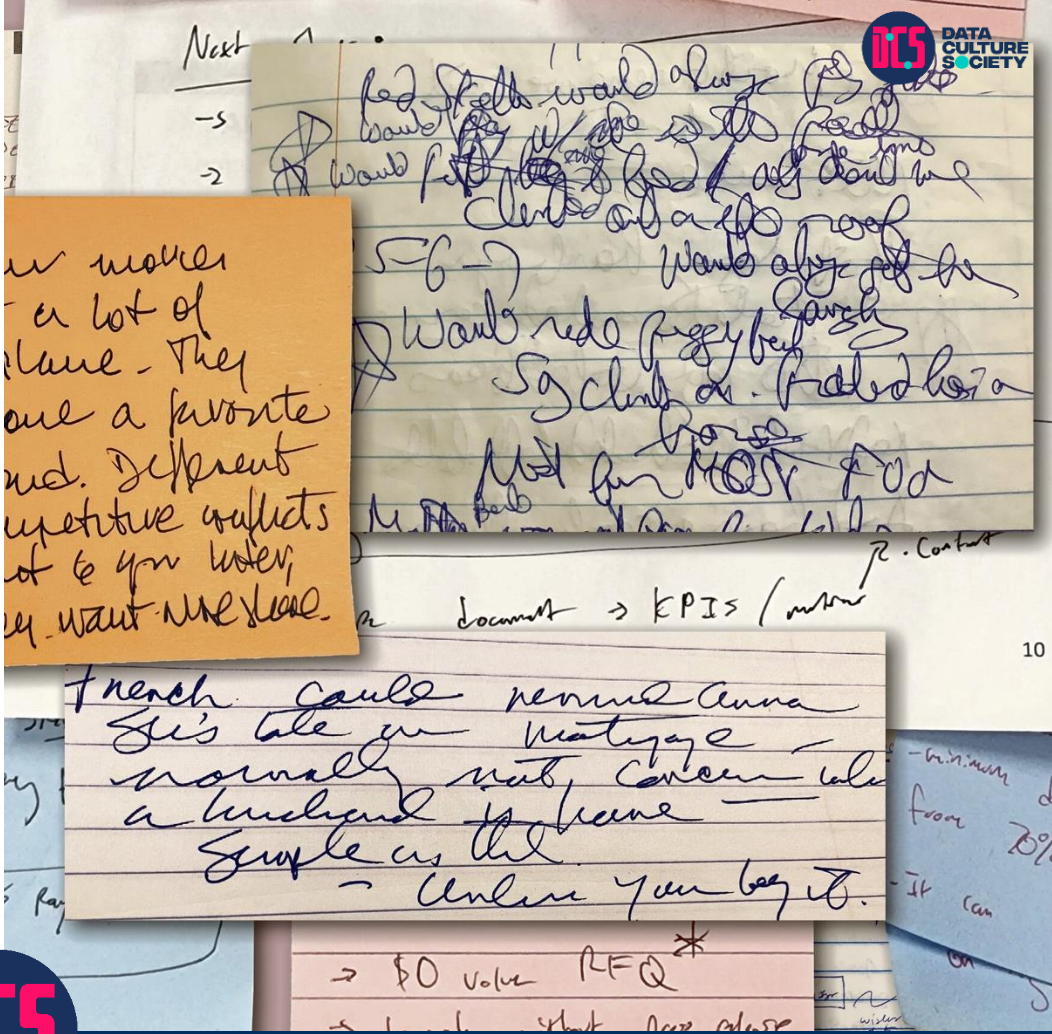
Remember, digitization is imperfect

- 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.

waiting there for me. the Master of the sloop
me till further orders
they met in Scotoss he
garry his wife
~~Georges Sons~~ Ronald
and saw me on Sep 30 we came to the Isle Huskre
Macleod. he
ciety sent a minister here I have given him
d he writ it down, you may be sure I have
his cons to you if you hear I'm alive do me
all hast but if you hear I'm dead do what
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 28th Feb–7th 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!