

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)

## Course Topics

- Text Analysis – analysing unstructured data
-  Python
- Regular Expressions
- Natural Language Toolkit (NLTK)

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  
you much more to tell then this when this comes to you if you have





## Course Structure

Anticipate about ~7 hours/week

- 2 hour course meeting, 2-4 Wednesday
- 1 assignment per week, ~2 hours
- Office hours on request
- Independent learning, ~2 hours
- Teams for introductions, meetings, office hours, questions, files

8<sup>th</sup> Minda Jan: 20 17<sup>th</sup>

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 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 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 made him  
a rode to scourgeth me. \* \* \* \* \*

# INTRODUCTIONS

- Why are you interested in text analysis?
- Have you used Python before?
- Have you used Jupyter Notebooks before?
- Have you used Regular Expressions before?
- Have you used NLTK before?





## Participant Expectations

Wednesday classes are introductions to material

Assignments will be given on Thursday

Classes are not recorded but all class materials will be uploaded to Teams

Please let me know in advance if you cannot attend!

Message me on Teams to schedule office hours for questions

8<sup>th</sup> Minda Jan: 20 17<sup>th</sup>

Sir

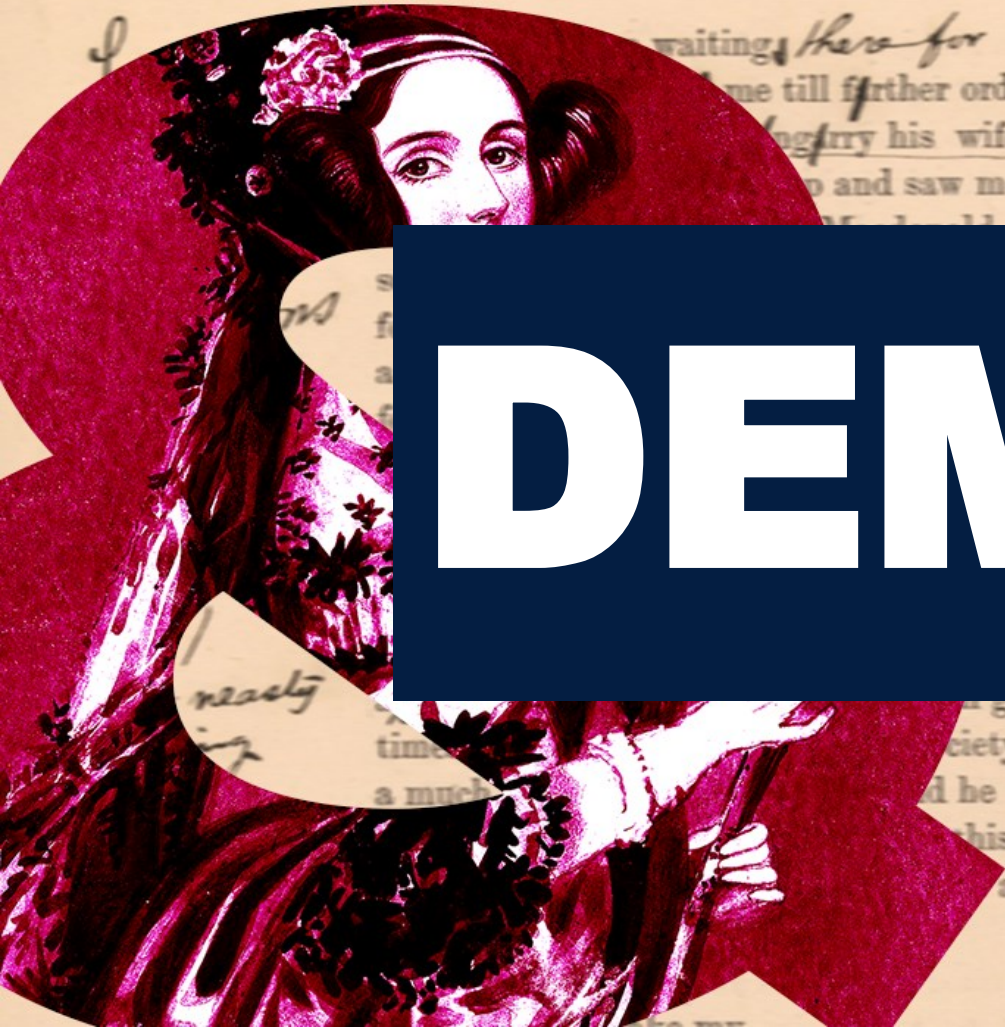
It is a great blessing and happiness to a nation when the King employeth such a man as you are to act and do for him who I'm perswaded his the awe and fear of God on him. Job was a just man and a perfect and the cause that he know not he searched out to deliver the poor and oppressed and him that had none to help 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 made him a rode to scourgeth me. \* \* \* \* \*

## Course Software

### Jupyter Notebooks / Jupyterlabs

- With Notable <https://www.ed.ac.uk/information-services/learning-technology/noteable/accessing-noteable>
  - After logging into MyEd: <https://noteable.edina.ac.uk/launch>
- With Google Colab <https://colab.research.google.com>
- Locally (install with pip/pip3 or conda)





# DEMO 1

waiting here for me. The master of the sloop  
till further orders they met in Scotos he  
logfury his wife A Georges Sons Ronald with  
and saw me on Sep 30 we came to the Isle Muskre Macleod. he  
ordered him  
take me  
come to his  
It  
o/t 93  
society sent a minister here I have given him  
d 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

## Further Resources

- Noteable User Guide:  
[https://noteable.edina.ac.uk/user\\_guide/#hide\\_ge\\_7](https://noteable.edina.ac.uk/user_guide/#hide_ge_7)
- Jupyter Notebooks in Noteable:  
<https://github.com/edina/Exemplars2020/blob/master/TeachingDocs/Tutorials/UsingNoteableBeginner.ipynb>
- Jupyter Notebooks: <https://glam-workbench.github.io/getting-started/>
- Python: <https://programminghistorian.org/en/lessons/introduction-and-installation>



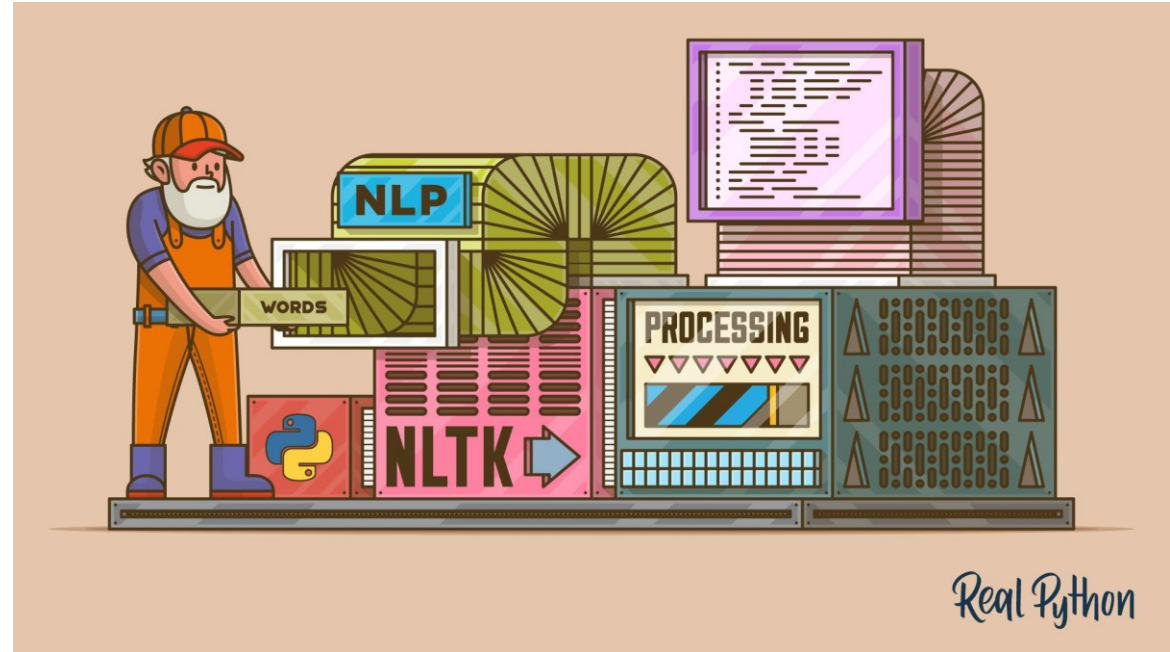


# NLTK

Natural Language Toolkit

Natural language = human language

= “unstructured” data

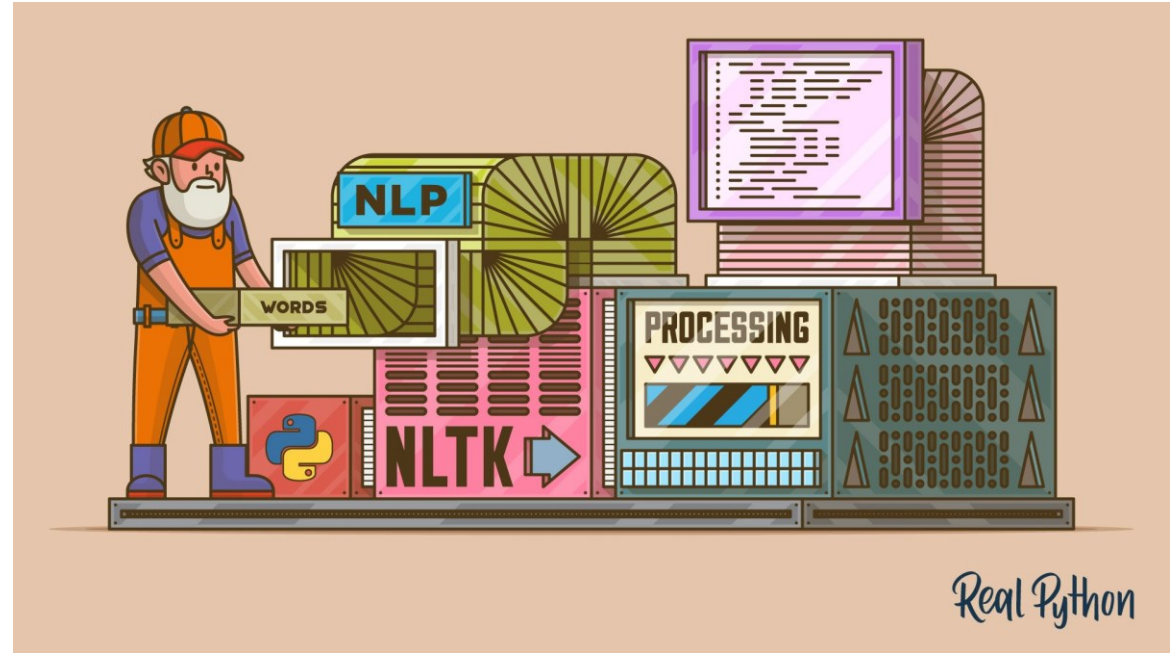


## NLTK

Examples of data sources for natural language:

- Books
- Newspapers
- Magazines
- Websites
- Transcriptions of audio (i.e. interview, movie dialogue)
- Social media

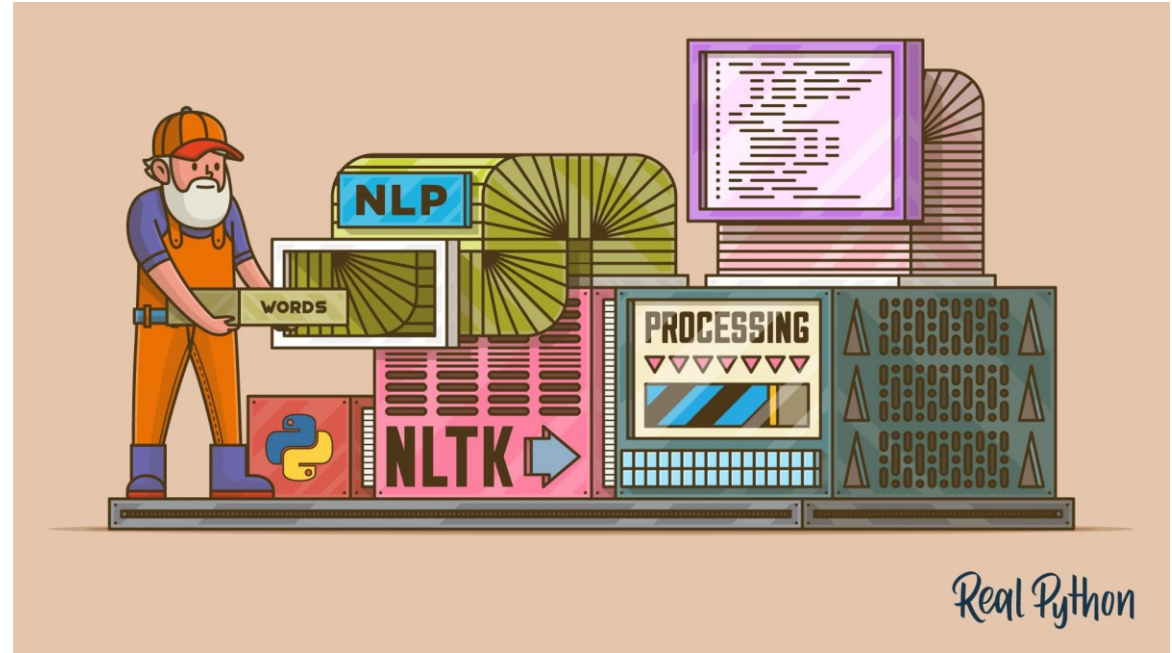
Always read the licensing/copyright information and terms of use!



## Why use NLTK

What kinds of questions can you ask when you can use a programming language to study hundreds, thousands, or even millions of pages of digital text?

“Distant reading”





## Why use NLTK

What kinds of questions can you ask when you can use a programming language to study hundreds, thousands, or even millions of pages of digital text?

“Distant reading”

## NLTK Isn't everything

What kinds of questions can you ask when you can physically hold and look at a printed text, be it an original publication or later edition of the text?

“Close reading”

Book history



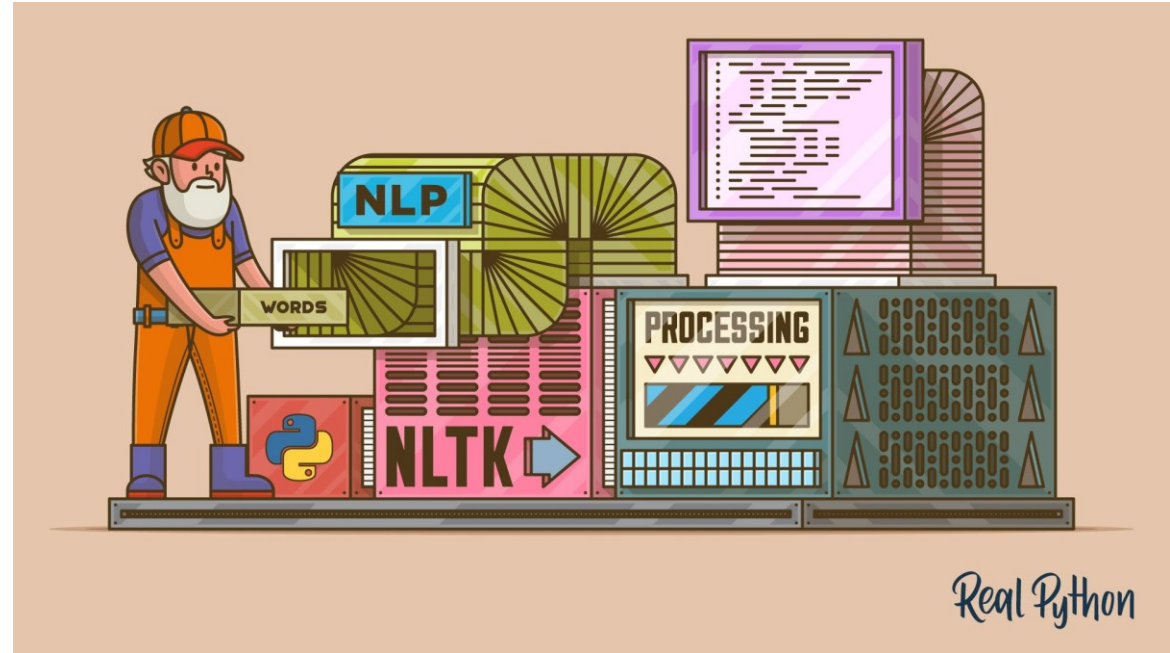
## NLTK Terminology

Tokens vs. words

Digitized vs. digital

Normalization (a.k.a. standardization)

Document vs. corpus vs. corpora



# Summarising a Text

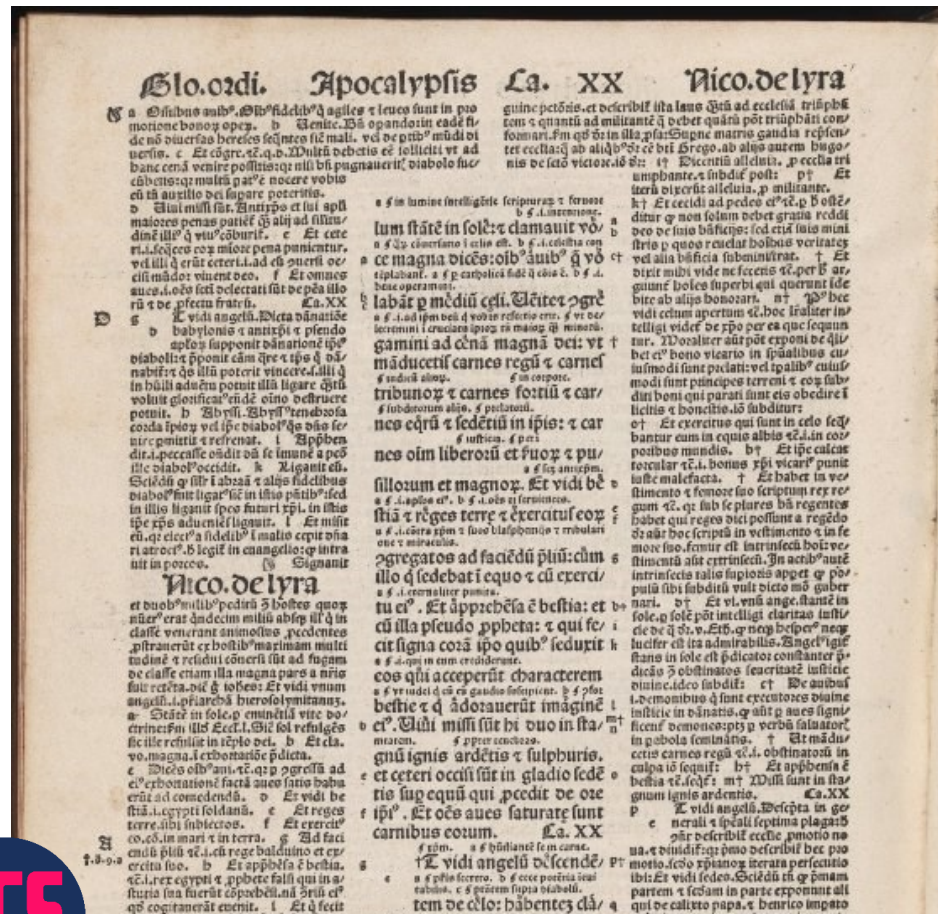
Built-in functions include:

`len(text)`

`sorted(vocabulary_of_text)`

NLTK Text methods include:

`Text.count(word)`

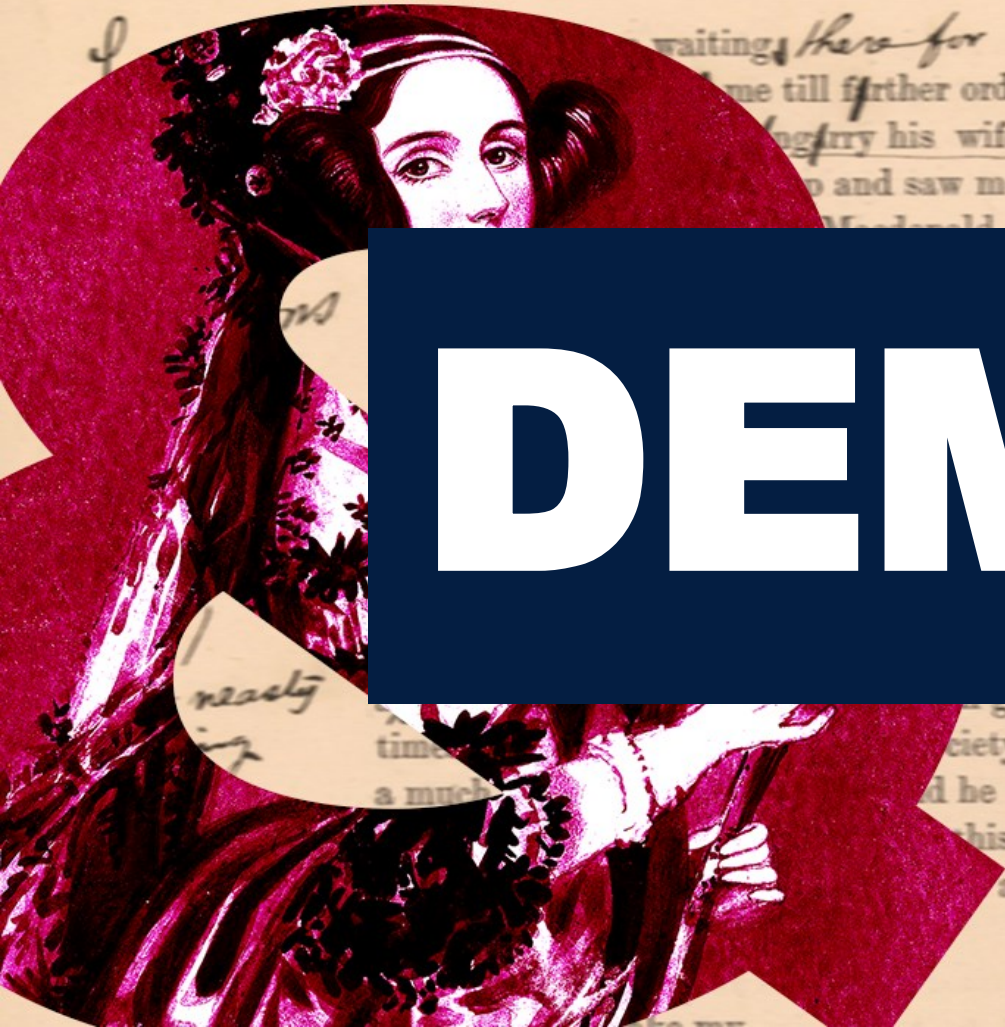


Reference: <https://www.nltk.org/book/ch01.html>

Biblia Sacra by N/A - University of Edinburgh, United Kingdom - CC BY.  
[https://www.europeana.eu/item/9200261/BibliographicResource\\_3000058482943](https://www.europeana.eu/item/9200261/BibliographicResource_3000058482943)

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





# DEMO 2

waiting here for me. The master of the sloop  
me till further orders they met in Scotos he  
logfury his wife A Georges Sons Ronald with  
and saw me on Sep 30 we came to the Isle Huskre Macleod. he  
Macleod and this was in the tennent after I was  
ordered him  
take me  
me to his  
It  
o/t 93  
Society sent a minister here I have given him  
d he write 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

# Getting to know a text

NLTK Text methods include:

`Text.concordance ("word",`

`lines=20)`

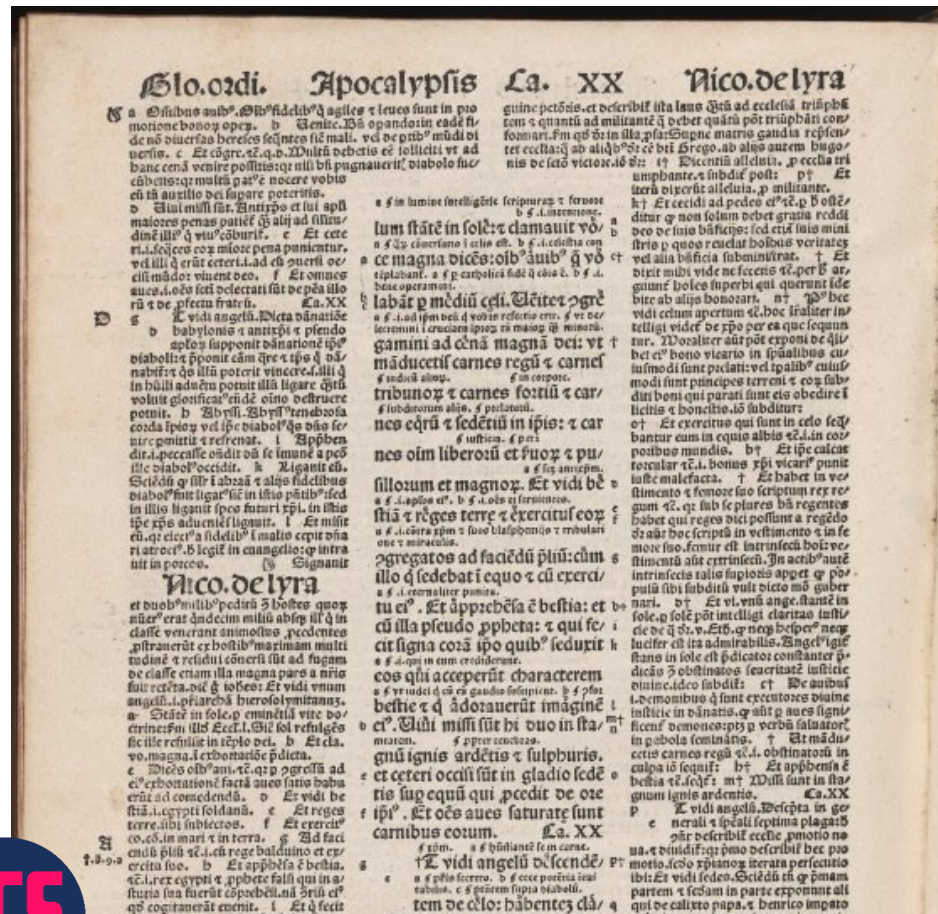
`Text.similar ("word")`

`Text.common_contexts (["list",`

`"of", "words"])`

`Text.dispersion_plot (["list",`

`"of", "words"])`





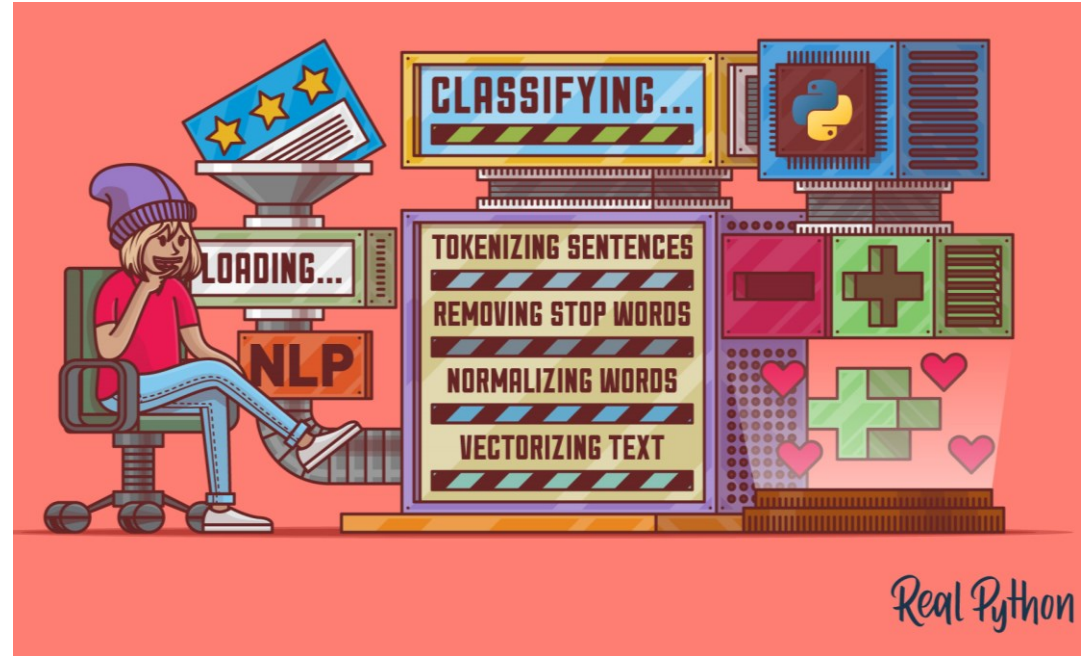


# DEMO 3-5



## The Building Blocks

- Tokenization - words/punctuation, sentences
- Normalization
- Stemming and lemmatizing
- Frequency counts
- Part-of-speech tagging



# Tokenisation

- Tokenisation involves breaking down a piece of text into smaller units called tokens.
- Tokens can be individual words, sentences, or even characters, depending on the level of granularity desired.
- Tokenisation helps in standardizing and organizing text data, making it easier to analyse and process.
- Word-based tokenisation breaks down text into individual words, treating each word as a separate token.



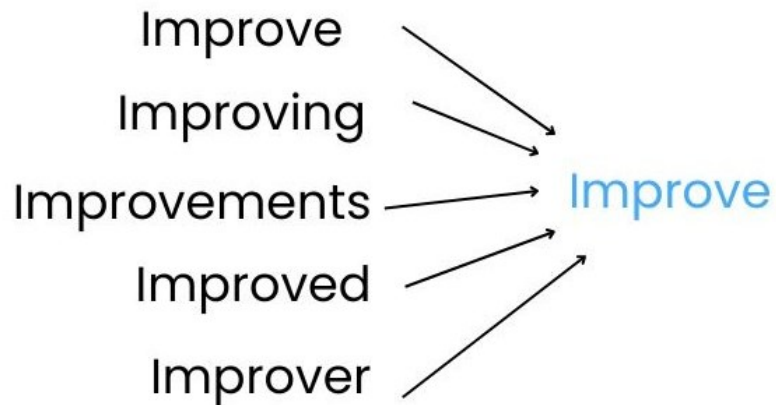
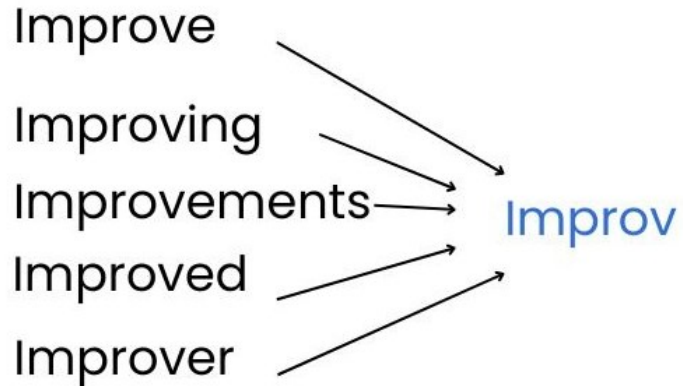
# Text cleaning & pre-processing

- Formatting text for analysis and removing extraneous information Workflows vary depending on research objective, field, and dataset
- Common steps include standardising capitalisation, removing URLs and symbols, stopword removal, tokenisation, stemming, and lemmatization
- Stopwords include words like “a,” “the,” “of,” “an” that don’t add meaning to the dataset





# Stemming & Lemmatization



The background features a collage of historical elements. On the left, there is a circular inset showing a portrait of a woman with dark hair and a pink flower in it. Below this, another circular inset shows a person in a red garment holding a long staff or stick. The background is a textured, aged parchment paper with faint, handwritten text in cursive script. A large, dark blue horizontal band across the center contains the text 'DEMO 6 & 7' in white, bold, sans-serif capital letters.

# DEMO 6 & 7

## 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!)





## Research with NLTK

- Who is named in a text?
- What places are named in a text?
  - Chunking and Named Entity Recognition
- How does the vocabulary of an author change over time?
  - Lexical Diversity



## Research with NLTK

- What are the common themes throughout a corpus?
  - Topic Modeling
- What attitudes are expressed in a corpus?
  - Sentiment Analysis
- What words occur near each other throughout a corpus? How does the meaning of a word change over time?
  - Word Embeddings



## Next Week

- Research with NLTK on a corpus
  - NLTK with pandas (for tabular data)
  - NLTK with Altair (for data visualization)
- Regular Expression practice
- Cleaning messy text
- Resources for more text analysis practice

8<sup>th</sup> Minda Jan: 20 17<sup>th</sup>

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 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 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 made him  
a rode to scourgeth me. \* \* \* \* \*



## Further Resources from CDCS

- Digital Method of the Month on Text Analysis
- Training Pathway for Text Analysis

8<sup>th</sup> Kilda Jan: 20 17<sup>th</sup>

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 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 searched 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 made him  
a rode to scourgeth me. \* \* \* \* \*

The background of the slide is a collage. On the left, there is a circular inset showing a woman in a red dress with a large floral corsage. The rest of the background is a textured, aged paper with faint, handwritten text in cursive. The text is mostly illegible but includes phrases like 'waiting here for me', 'the master of the ship', 'they met in Scotland', 'George's son Ronald', 'with', 'for the tenant of this Isle his name Alex. Macdonald', 'to come to the Captain of', 'a great miserie in the Husker but I am ten', 'society sent a minister here I have given him', 'and he wrote 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', and 'o/t - 93'.

**Next class: Wednesday 19th, 2-4PM**  
**Please message me on Teams for office hours!**