

12th–19th February 2025


Centre for Data, Culture & Society



INTRODUCTION TO TEXT ANALYSIS WITH PYTHON

Instructor: Xandra Dave Cochran

Course Topics

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

Sir

8th Milda Jan: 20 1743

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 he
him, a Pattern for on in your office. I leave 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 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

Sir

8th Milda Jan: 20 1743

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 help
him, a Pattern for on in your office. I leave 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

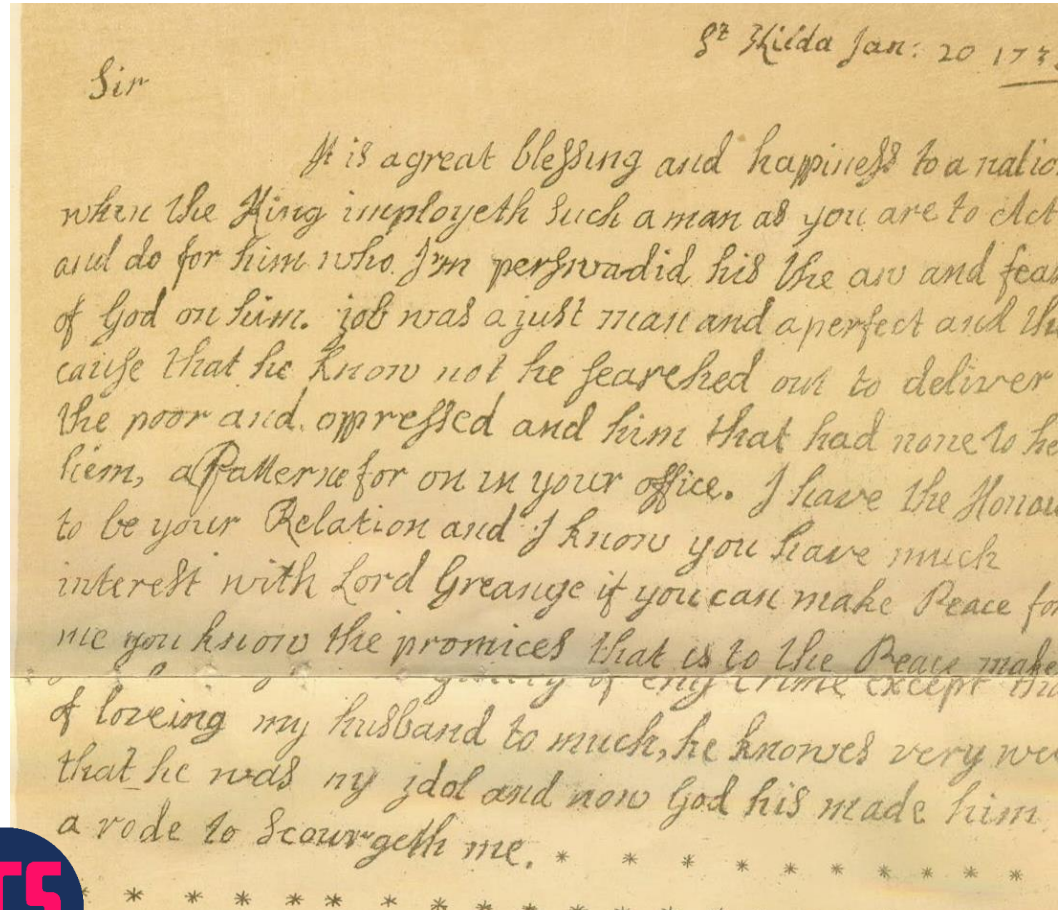
Wednesday classes are introductions to material

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

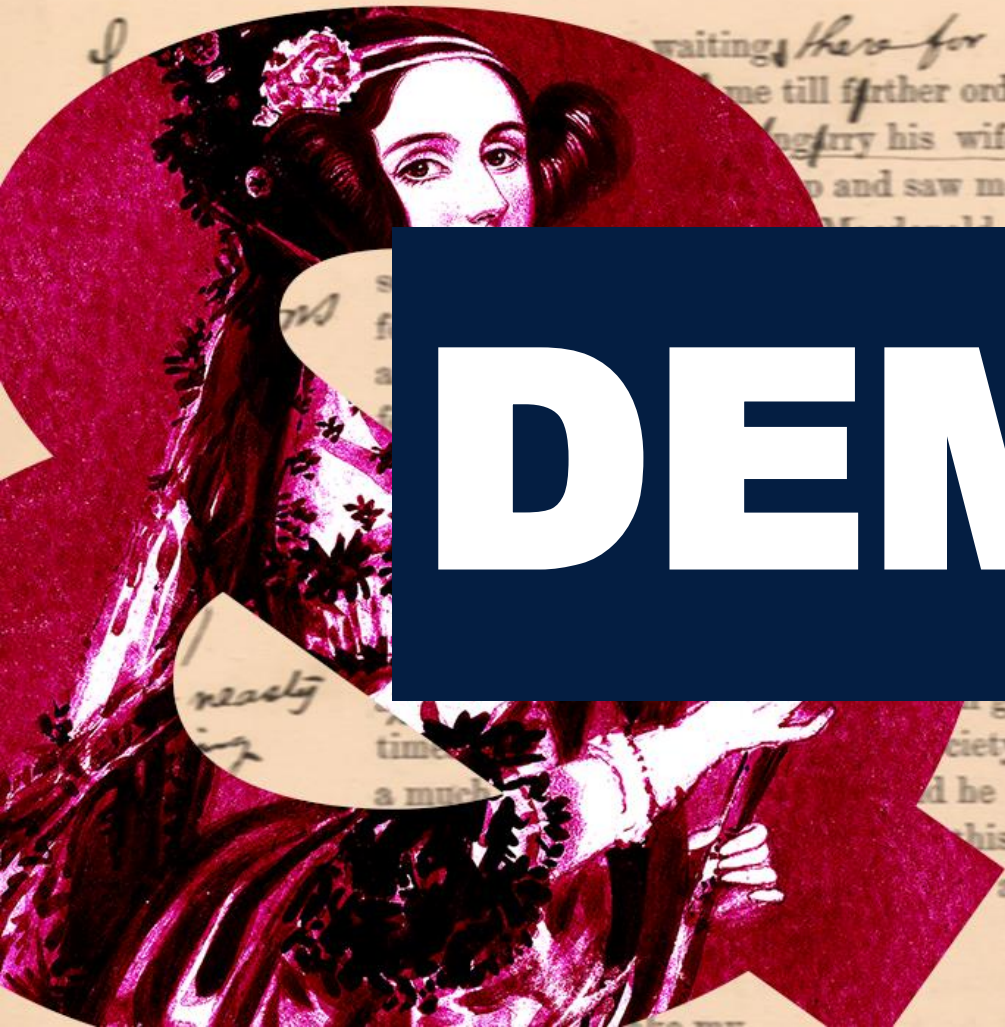


Course Software

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
his wife A ~~George~~ Ronald with
and saw me on Sep 30 we came to the Isle of Macleod. he
ordered him
take me
come to his

great interest in the...
society sent a minister here I have given him
and he wrote it down, you may be sure I have
this comes to you if you hear I'm alive do me
all hast but if you hear I'm dead do what

1/ It
0/t 93

Further Resources

- Noteable User Guide:
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>

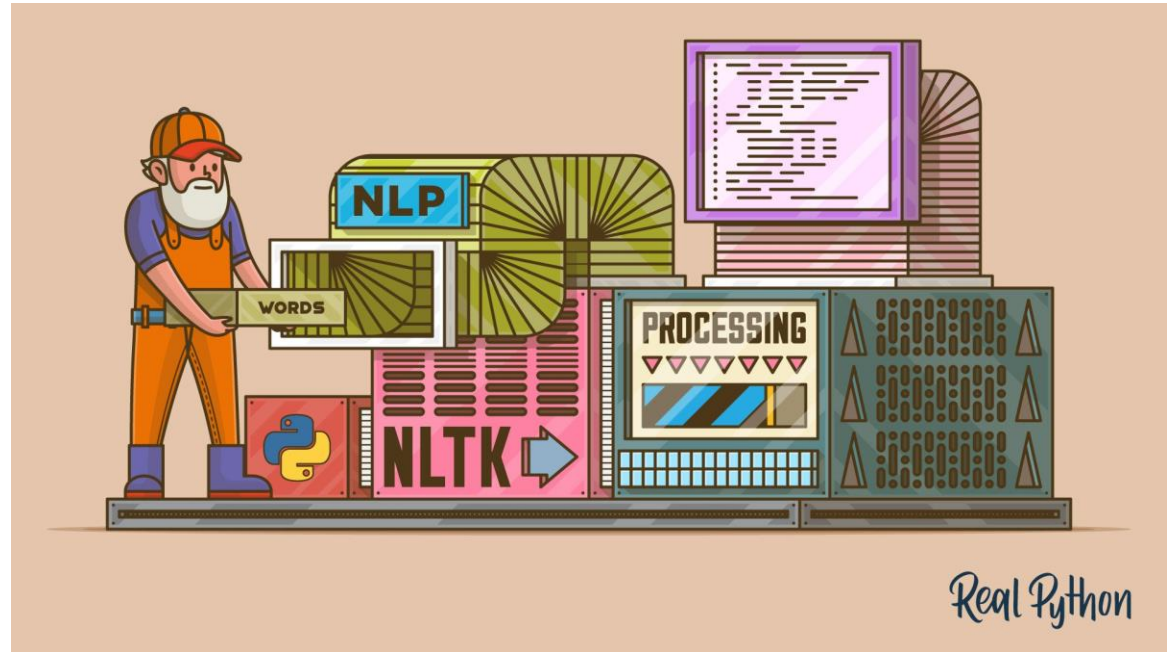


NLTK

Natural Language Toolkit

Natural language = human language

= “unstructured” data

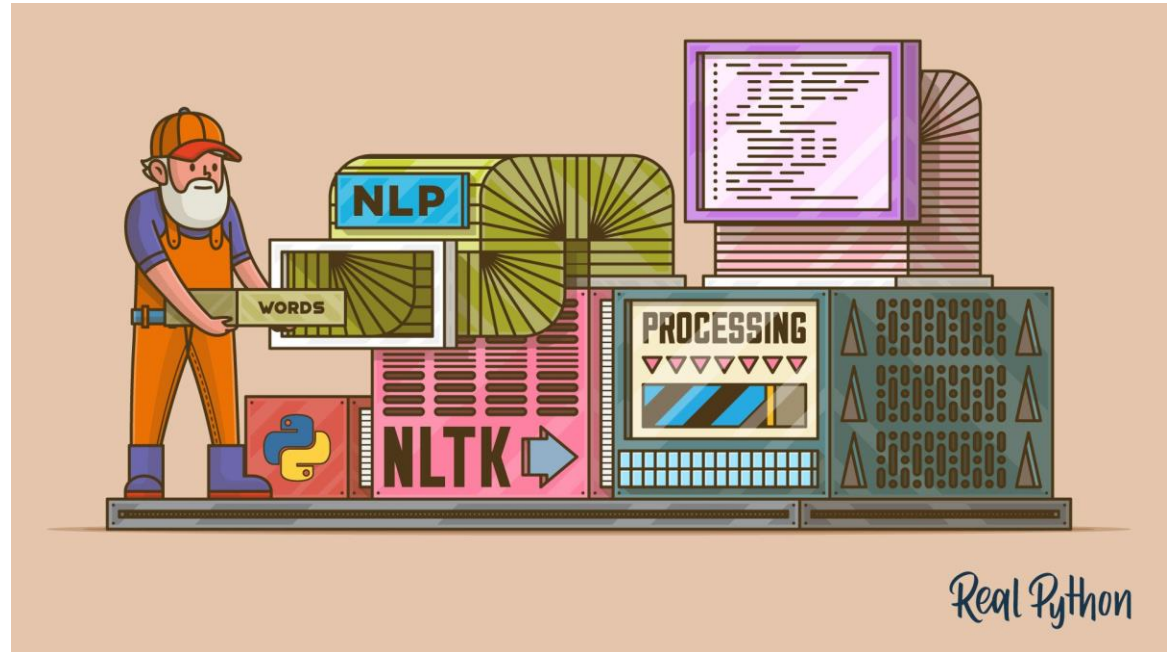


NLTK

Examples of data sources for natural language:

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

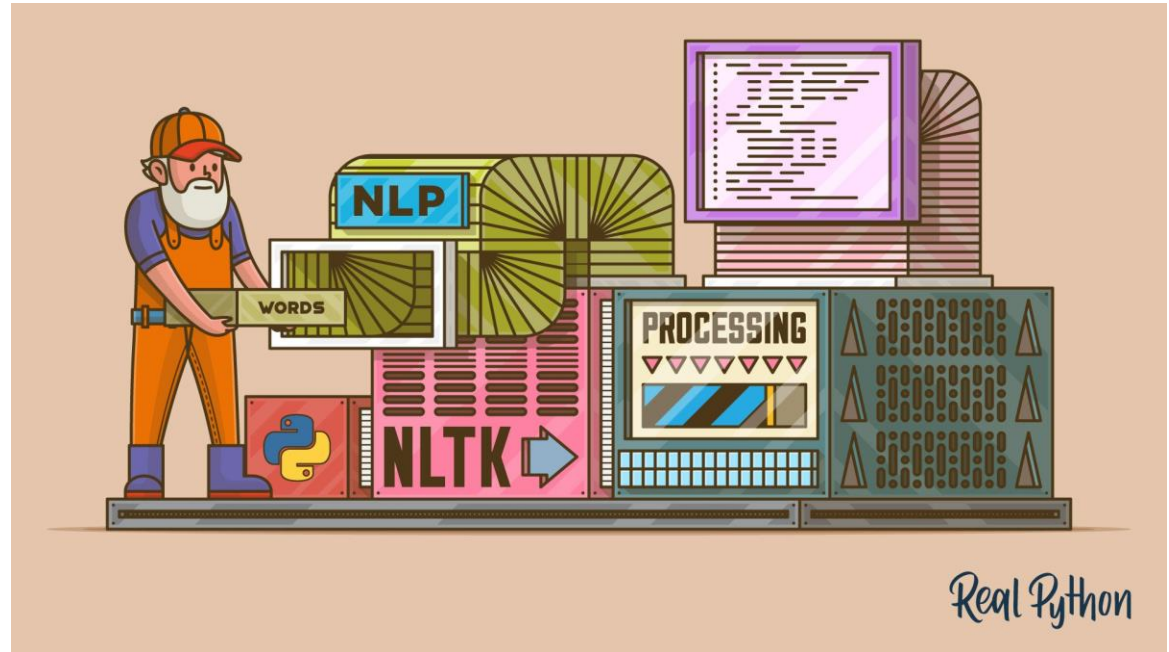
Always read the licensing/copyright information and terms of use!



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



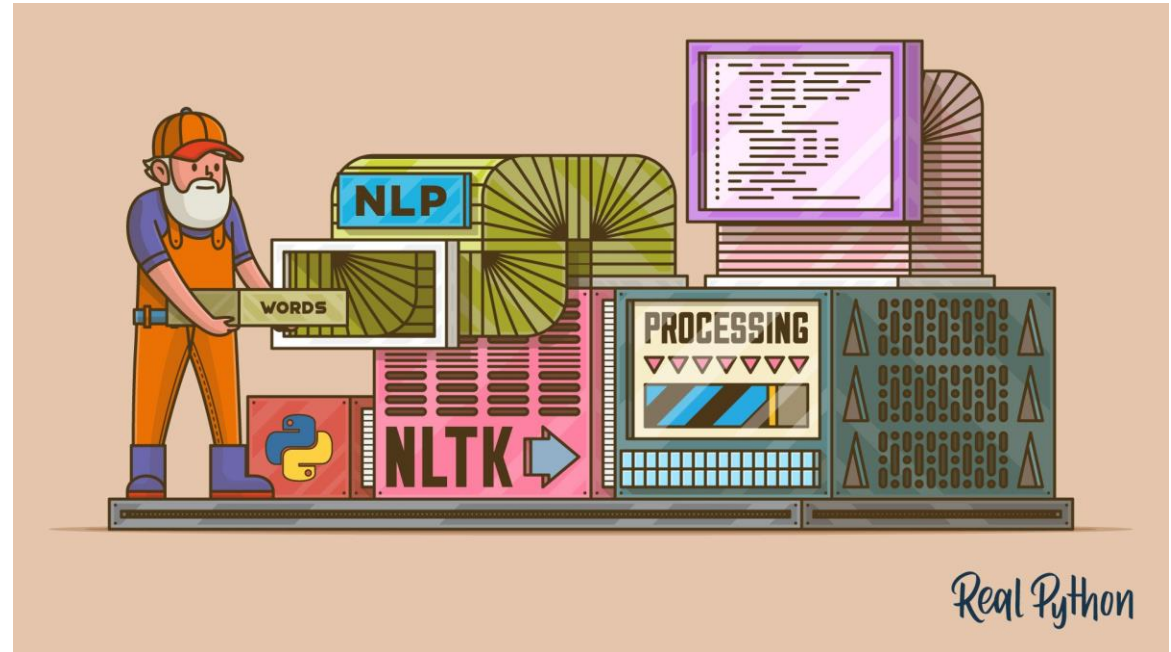
NLTK Terminology

Tokens vs. words

Digitized vs. digital

Normalization (a.k.a. standardization)

Document vs. corpus vs. corpora



Summarising a Text

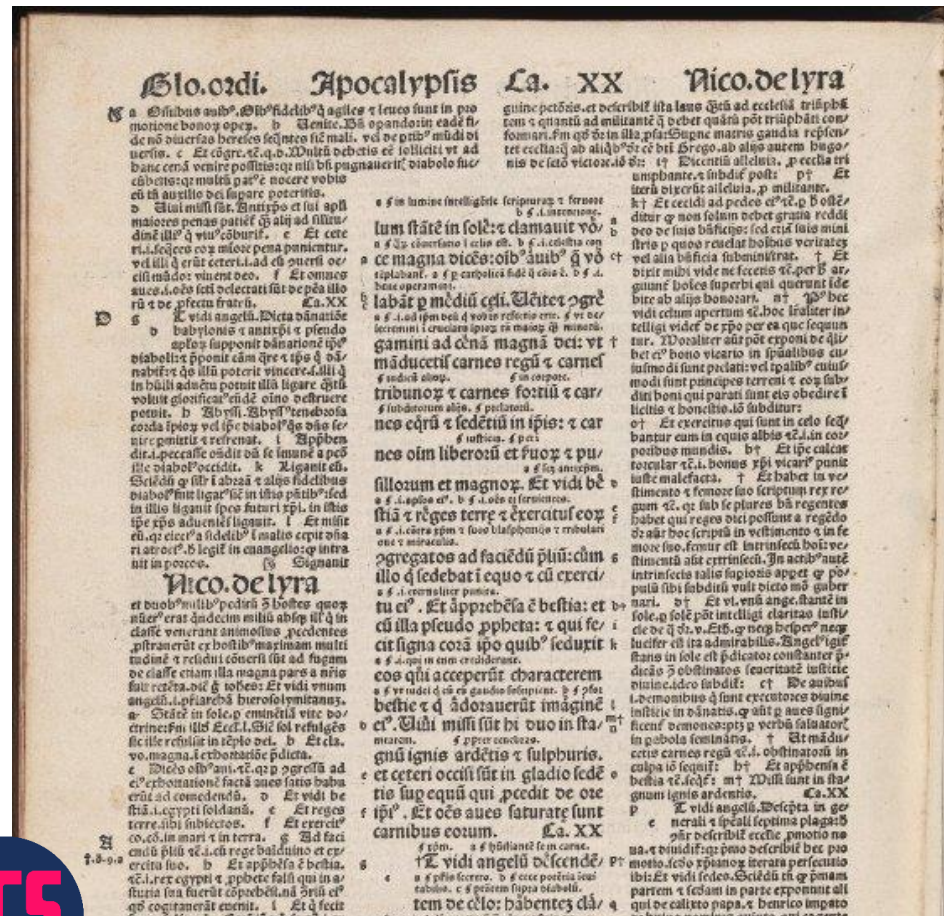
Built-in functions include:

`len(text)`

`sorted(vocabulary_of_text)`

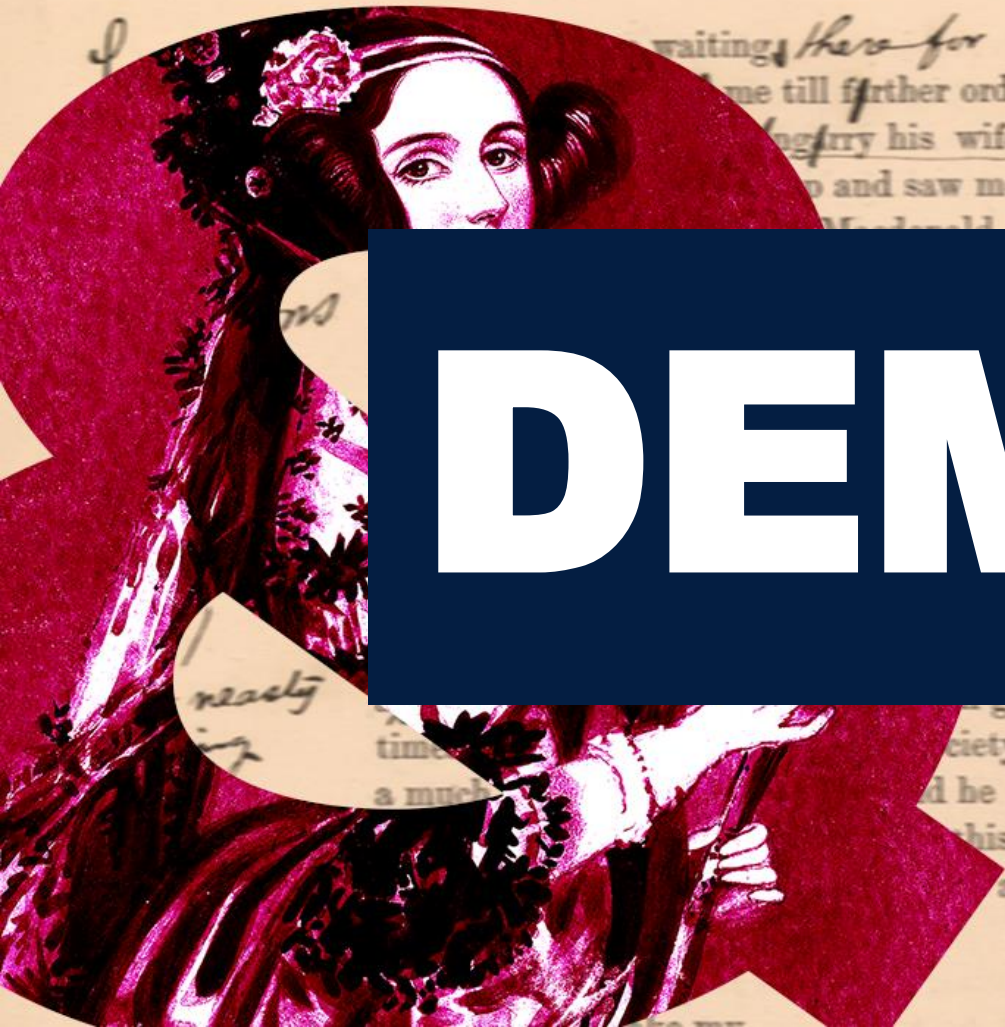
NLTK Text methods include:

`Text.count(word)`





DEMO 2



waiting there for me. The Master of the sloop
till further orders they met in Scotos he
George's Son Ronald with
and saw me on Sep 30 we came to the Isle MacLeod. he
MacLeod and this was the tenner 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 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

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"])`

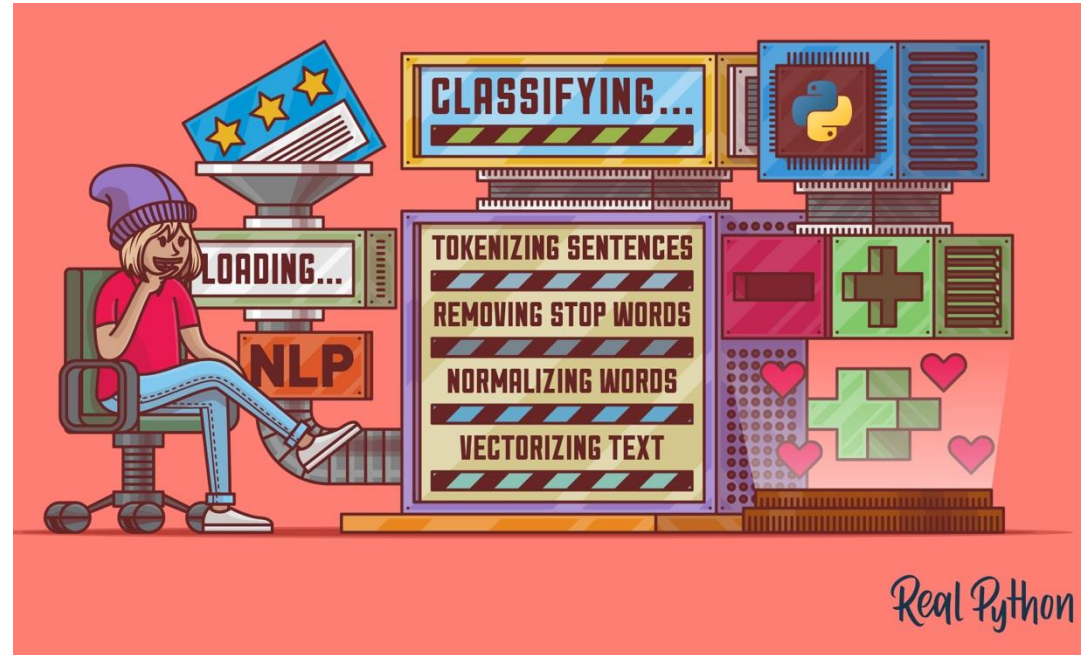




DEMO 3-5

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.

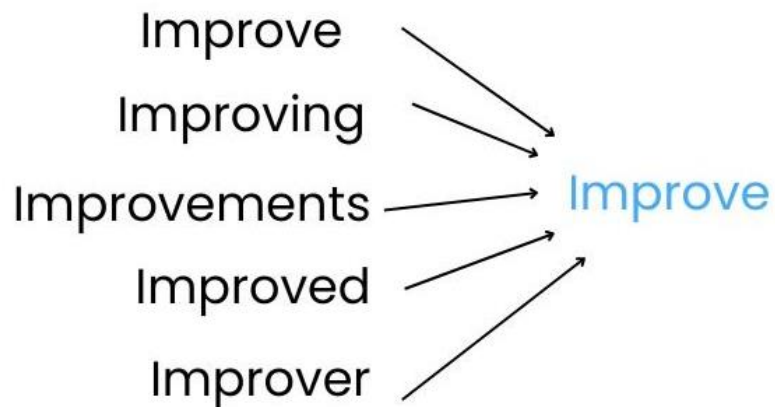
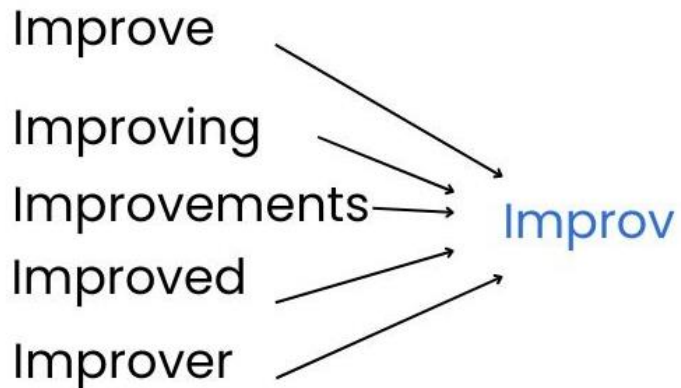


Text cleaning & pre-processing

- 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





DEMO 6 & 7

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!)



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

Sir

8th Milda Jan: 20 1743

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 he
him, a Pattern for on in your office. I leave 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

8th Milda Jan: 20 1743

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 help
him, a Pattern for on in your office. I leave 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', 'he', 'with', 'for the tenant of this Isle his name Alex. Macdonald to come to the Captain of', 'a much', 'a great miserie in the husker but I am ten', 'ciety sent a minister here I have given him', 'd he wr[it] 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'.

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