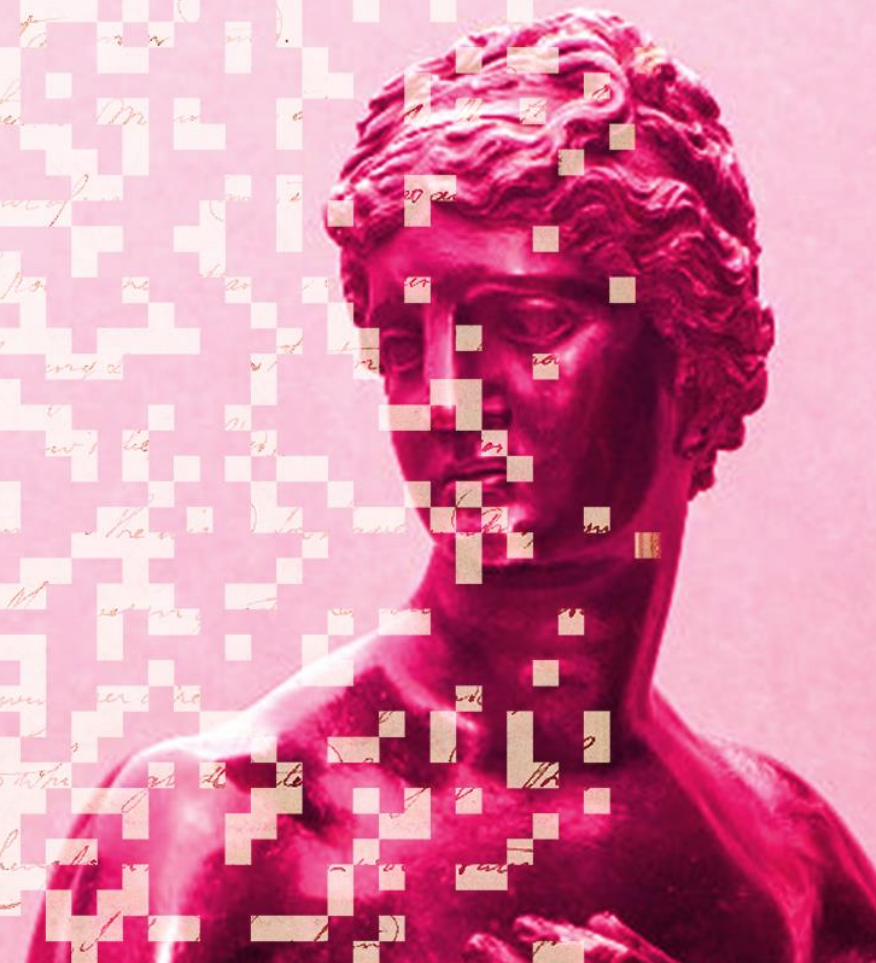
THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

Data Culture & society



@edCDCS

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

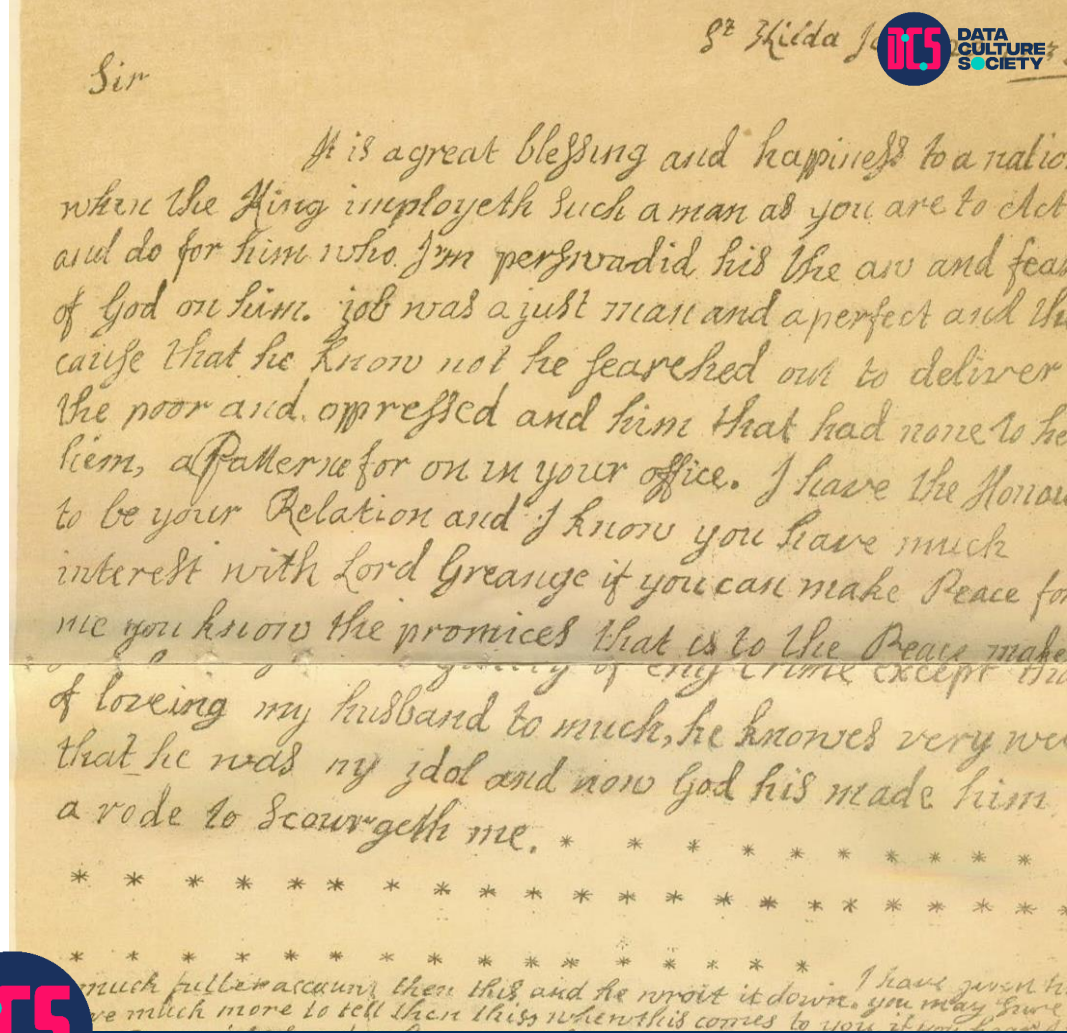
Introduction to Text Analysis With Python

16 & 25 Feb 2026

Instructor: Joy Lan
Adapted from materials made by Xandra Dave Cochran

Course Topic

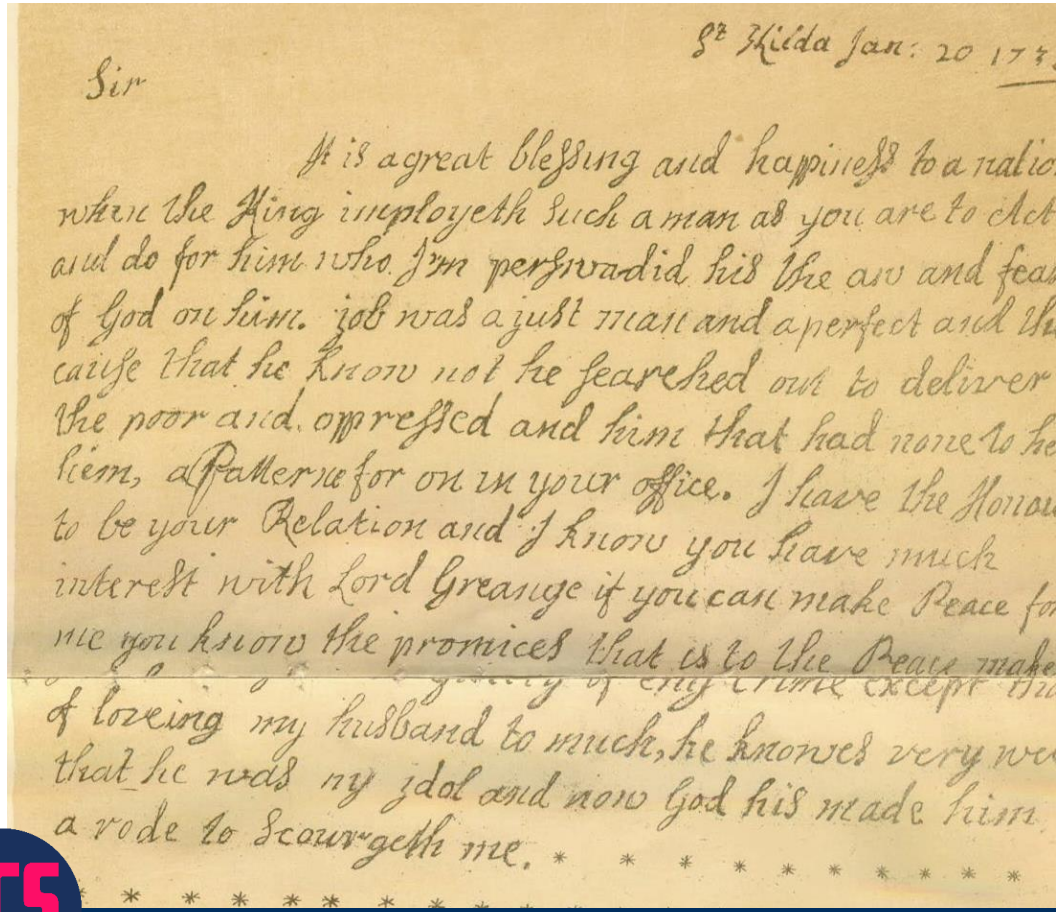
- Analysing unstructured data with Python
- Natural Language Toolkit (NLTK)
- Regular Expressions



Course Structure

Anticipate about ~7 hours/week

- 2 hour course, 2-4 pm on Monday
- Independent learning, ~2 hours
- Office hours on request on Teams
- All materials will be uploaded to the course page on Github



Introduction

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

Set up the environment

Jupyter Notebooks / Jupyterlabs

- With Google Colab https://colab.research.google.com
  - Go to File > Open Notebook > Github> paste the Github link to the notebook
- Locally
  - Install jupyter notebook with pip/pip3 or conda
  - Download and open the notebook file
  - You'll also need to install the libraries (nltk, etc.) locally

Python Refresher (Opt.)

- Name these data types:
  - ”Hello!”
  - [1, 2, 3]
  - {”A”:1, “B”:2}
- Functions:
  - print()
  - while:
  - for i in list:
  - def function_name()

Download and Import nltk packages

- Import ntlk
- Use nltk.download() for downloading the packages, once they are downloaded, you can call these packages using "import”
- We will use nltk ‘gutenberg' and 'book’ as example corpus (textual data)

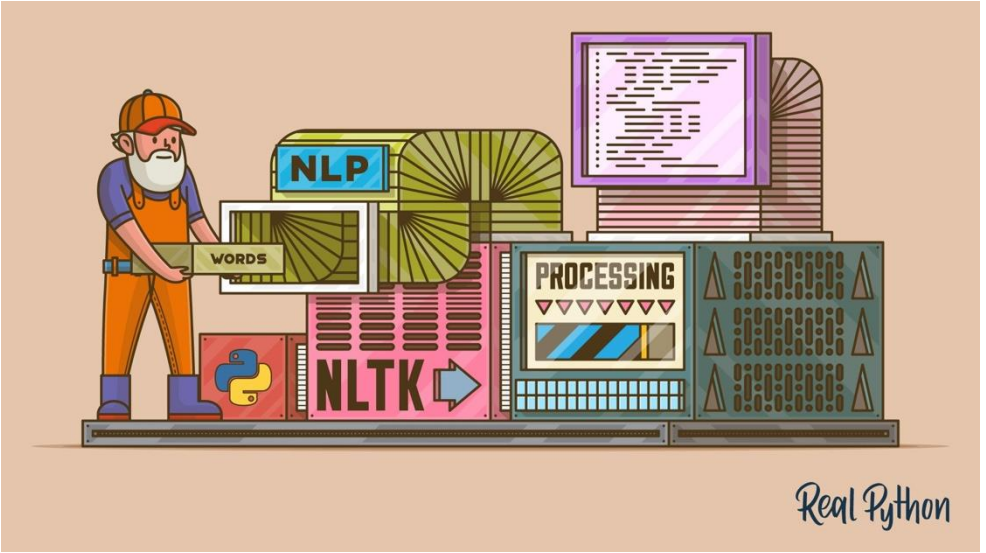
Getting to Know a Text

Data Source

Examples of data sources for natural language:

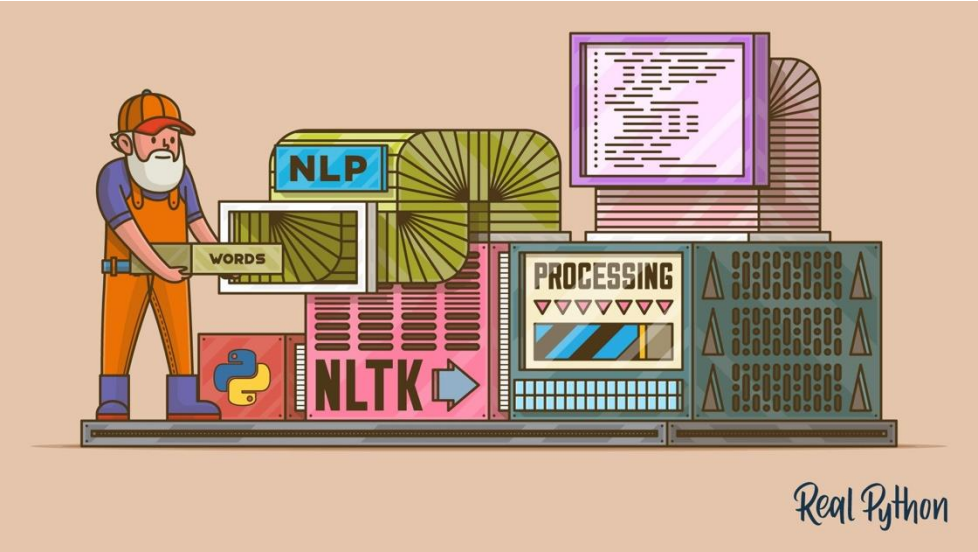
- Books
- Newspapers
- Magazines
- Websites
- Transcriptions of audio (i.e. interview, movie dialogue)
- Social media (memes)

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



Demo 1

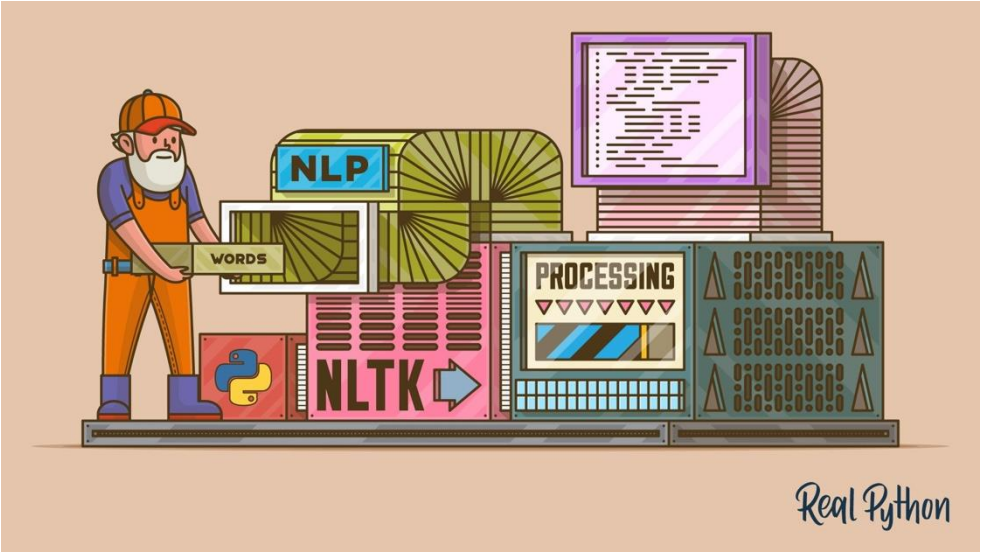
- Use built-in functions to print out and summarise text1, and text2



NLP Natural Language Processing

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

Text Represented by ?

In human language, text is represented by a collections of words, or sentences.

In NLTK, similarly, text is represented by tokens, unique word.

Each token is link to a numerical representation, such as a number.

Example:

the dog is eating, and the cat is eating. (9 words)

➔ Tokens:

0:the, 1:dog, 2:is, 3:eating, 4:”,” , 5:and, 6:cat, 7:”.”

(8 tokens)

Tokenisation

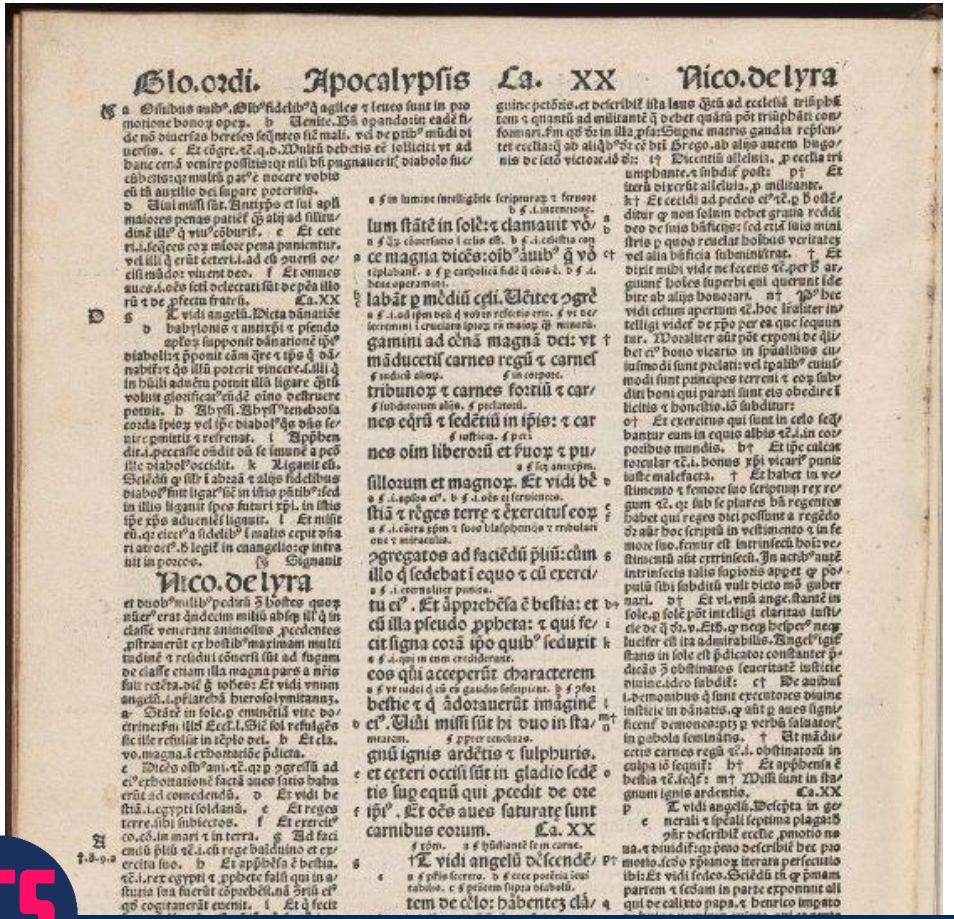
- Tokenisation involves breaking down a piece of text into smaller units called tokens.
- Tokens can be individual words, sentences, or even characters, depending on the level of granularity desired.
- Tokenisation helps in standardizing and organizing text data, making it easier to analyse and process.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

Getting to Know a Text

We can use Python built-in functions such as

```
len(text)
```

```
set(text)
```

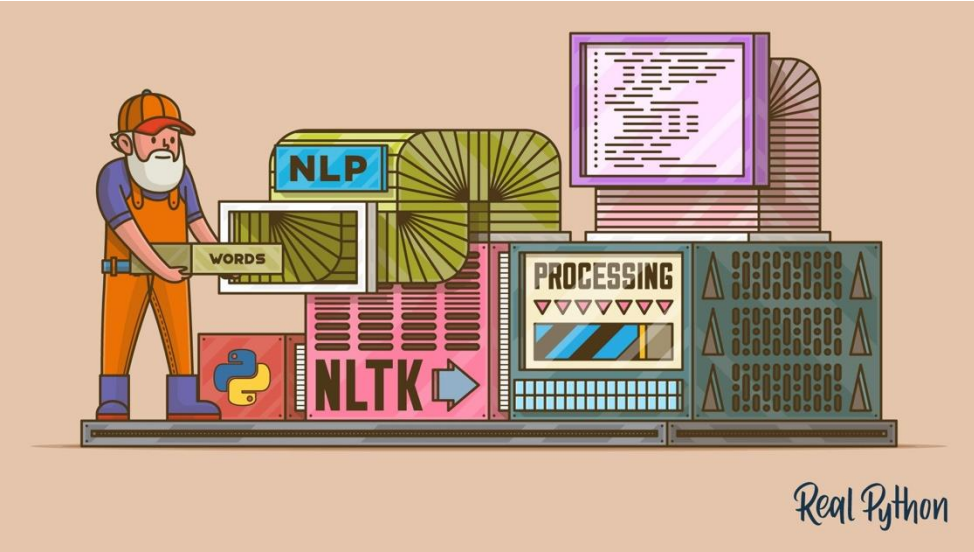


Demo 2

Calculate the following measures of text1 and text 2

- length of vocabulary
- lexical diversity

  (the number of tokens/the length of text)



Getting to Know a Given Word in a Text

NLTK Text methods include:

```
Text.concordance("word", lines=20)
```

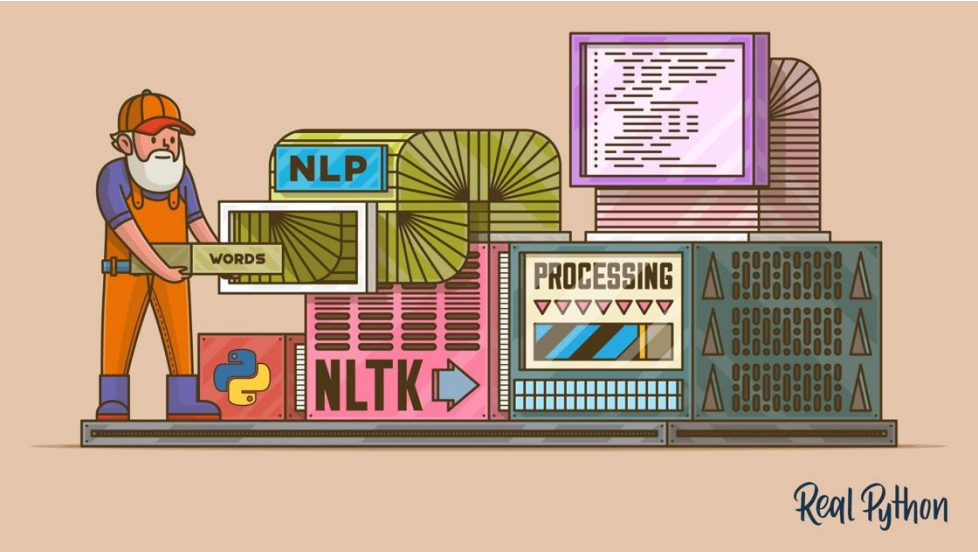
```
Text.similar("word")
```

```
Text.common_contexts(["list", "of", "words"])
```

```
Text.dispersion_plot(["list", "of", "words"])
```

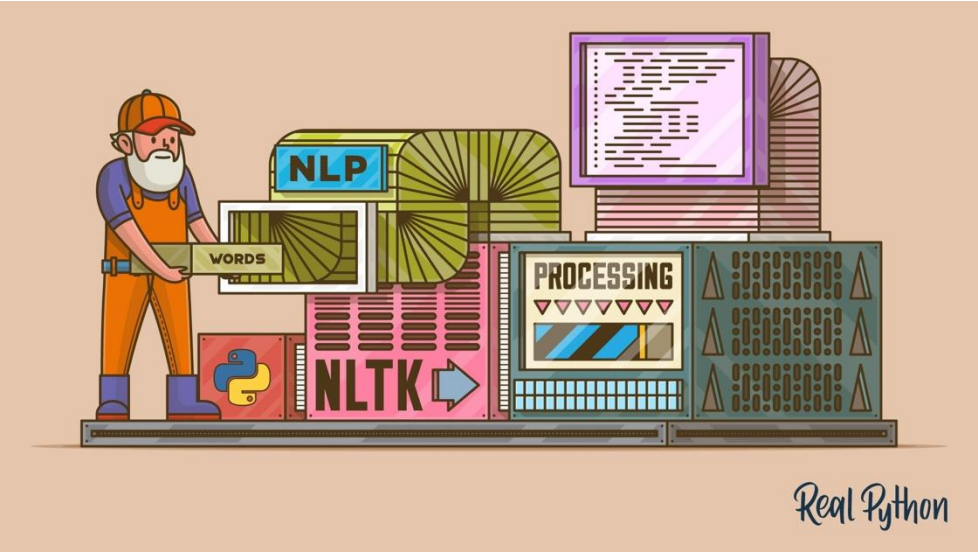
Demo 3-4

- Get the context of a given word
- Get similar words of a given word
- Get the common context of a list of words
- Plot the appearance of a given word across the document



Demo 5

- Using context of words and similar words to gain insight to the the relation between given words. (ex. good, opinion)
- Bigram and N-gram

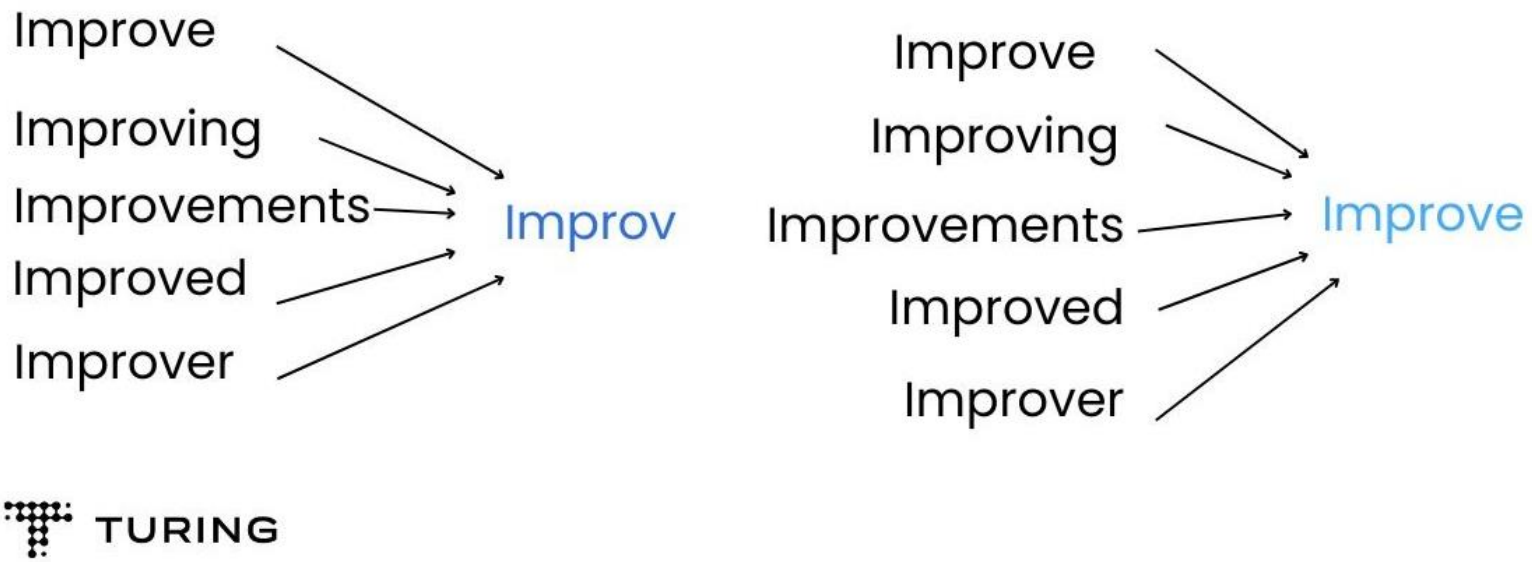


Customise Tokenization

Text Cleaning and Pre-processing

- Text cleaning and pre-processing:
  - remove unwanted characters (https:// , emojis, etc.)
  - Normalisation:
    - Lowercasing
    - Stemming and lemmatizing
  - Stopword removal: remove words that don't add meaning to the data
- Tokenization: split documents into words/punctuation, or sentences
- POS Tagging (Part-of-Speech Tagging): assigns grammatical labels to each token.
- Bigram, N-gram

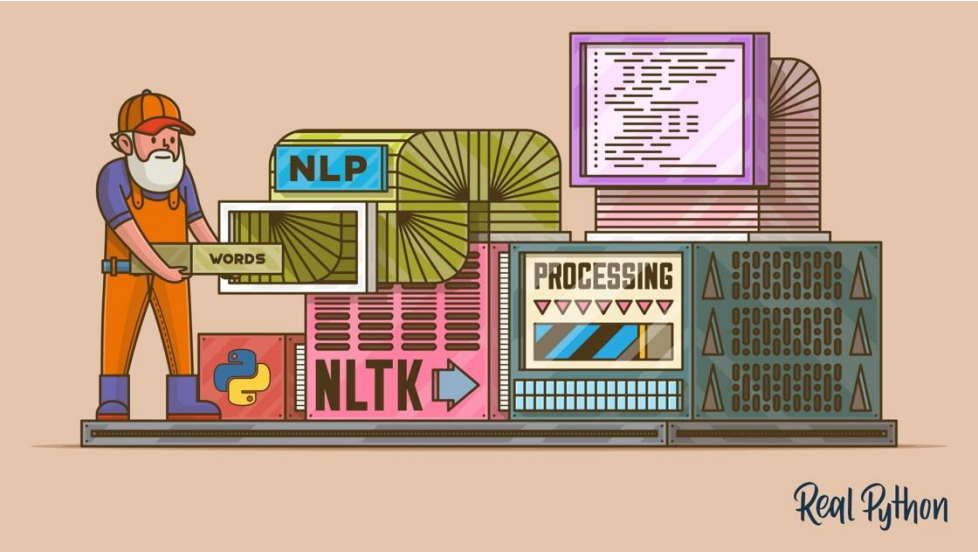
Stemming & Lemmatization



TURING

Demo 6

- Use word and sentence tokenizers
- Normalise text using word.lower()
- Stemming and Lemmatisation

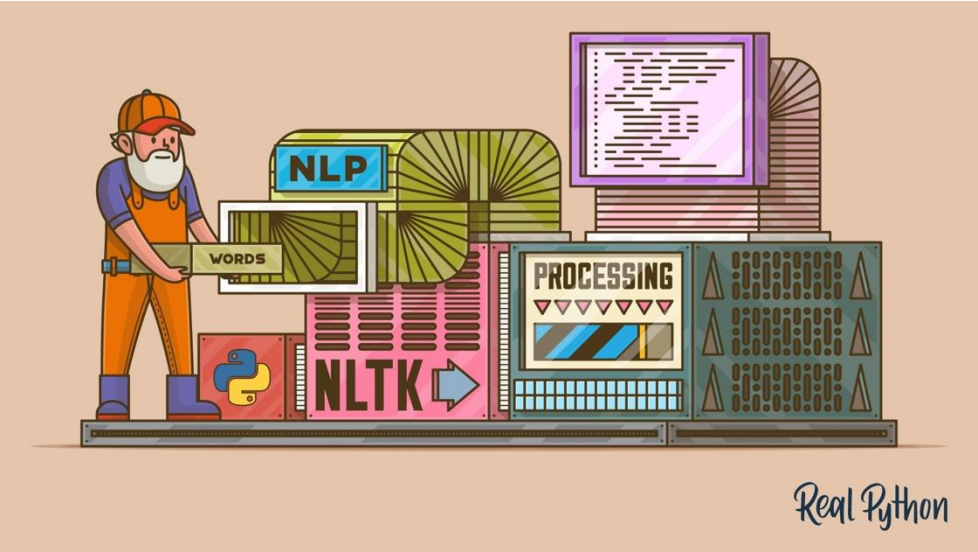


Stopwords

- Stopwords include words like “a,” “the,” “of,” “an” that don’t add meaning to the dataset.
- I am going to the shop. -> “I”, “going”, “shop”

Demo 7

- Apply stopword removal



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

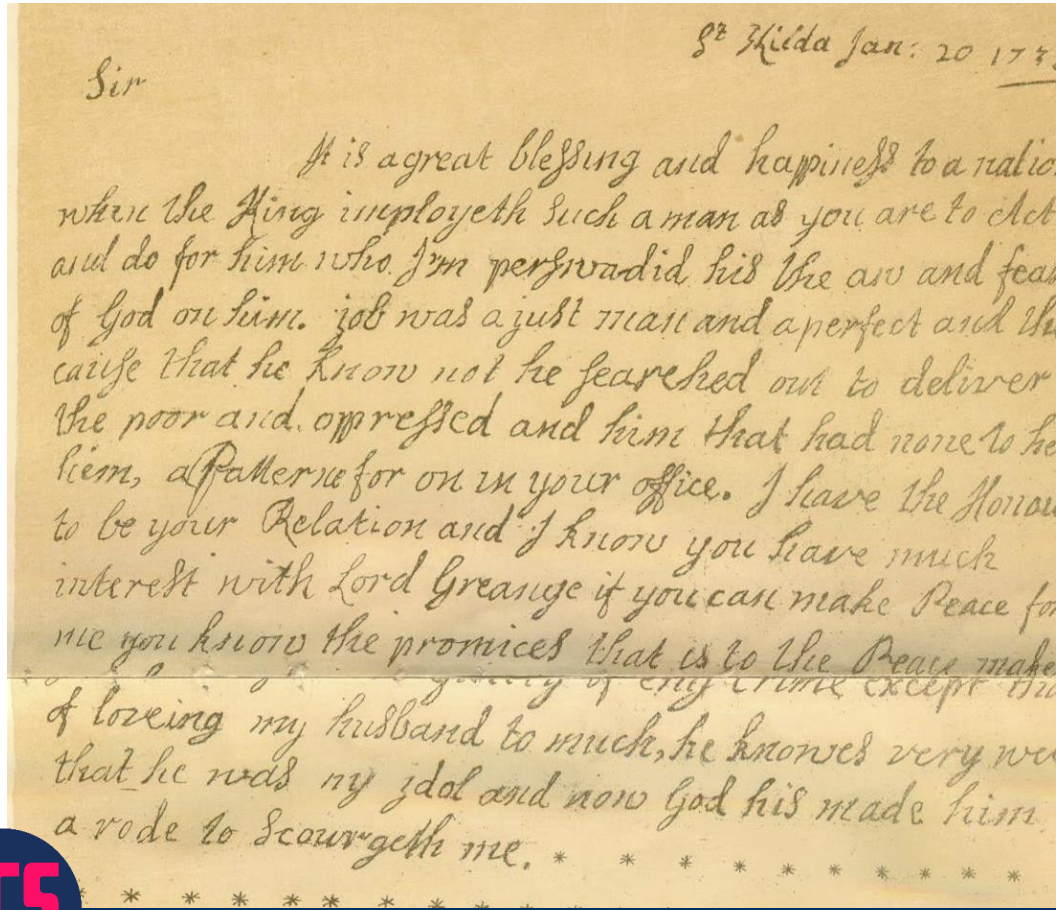
- Who is named in a text?
- What places are named in a text?
  - Chunking and Named Entity Recognition
- How does the vocabulary of an author change over time?
  - Lexical Diversity

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

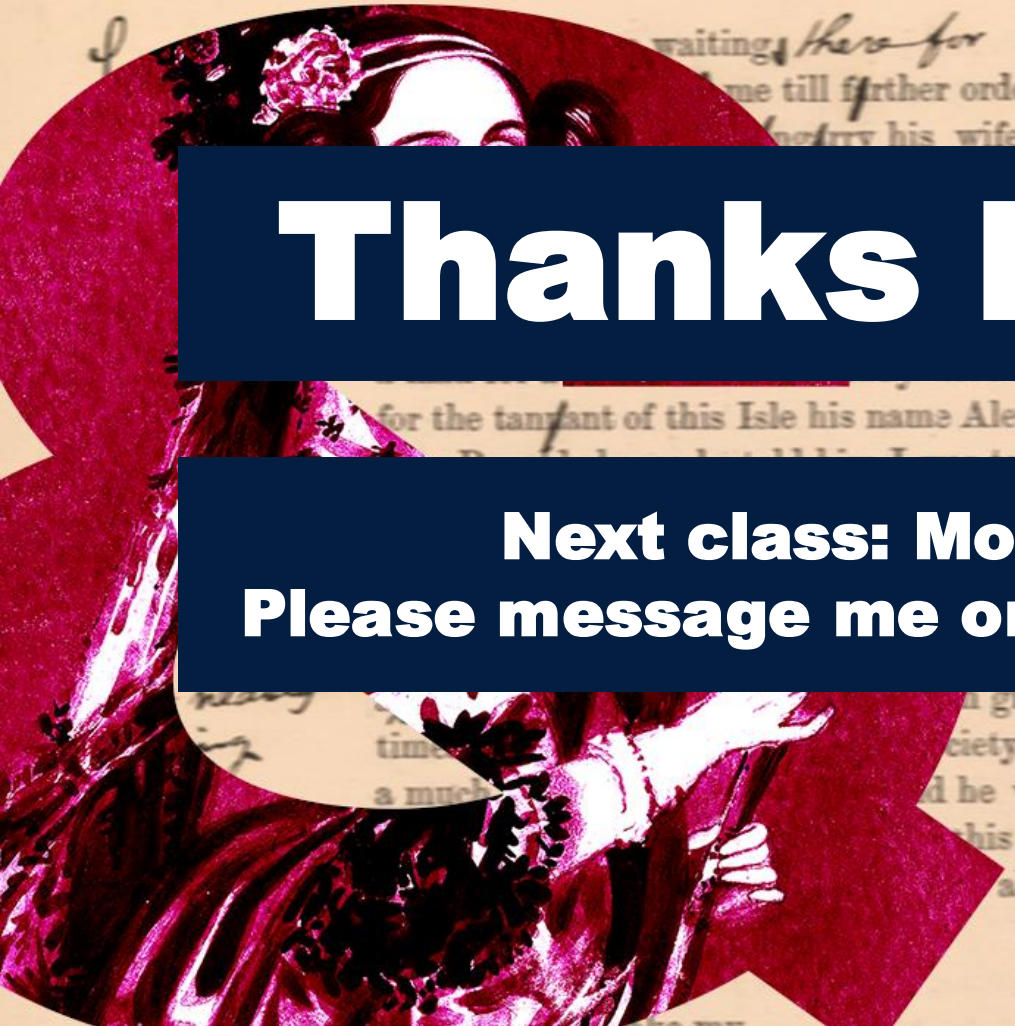
Next Week

- 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


St Kilda Jan: 20 17

Sir

It is a great 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
cause that he Know not he fearched out to deliver
the poor and oppressed and him that had none to he
him, 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
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. * * * * * * * * *



Thanks Everyone!

Next class: Monday 25th, 2-4PM
Please message me on Teams for office hours!