Supporting, delivering and facilitating high quality opportunities and training by engaging with students, staff and partners to positively impact the social science community in Scotland

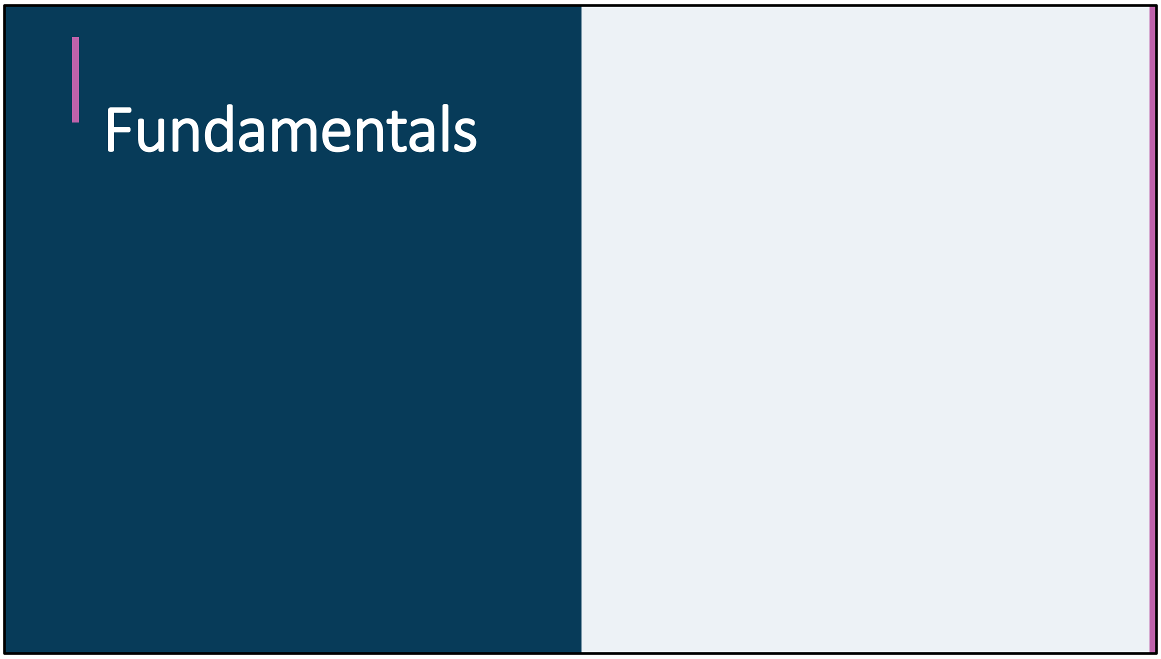
Text Analysis

2025-06-18
Dr Diarmuid McDonnell

Scottish Graduate School of Social Science
Sgoil Cheumnaichean Saidheans Sòisealta na h-Alba

UKRI Economic and Social Research Council

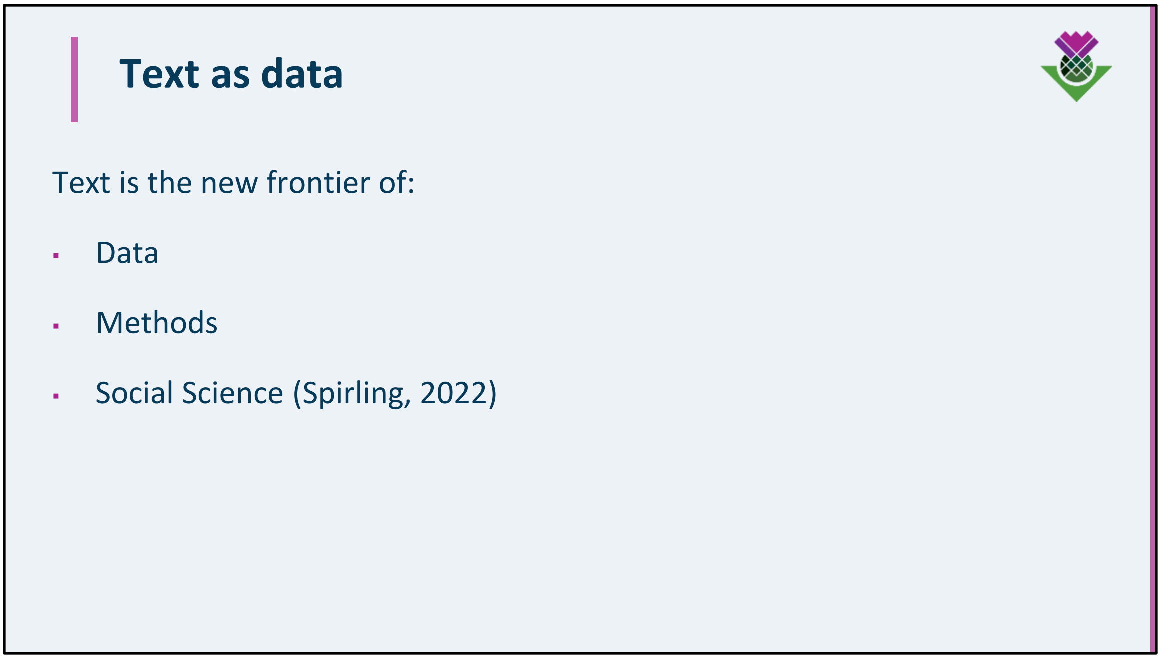
Scottish Funding Council
Promoting further and higher education



Fundamentals

Outline

1. Fundamental concepts of text analysis:
    1. Text as data
    2. Text as social science data
    3. Demystifying concepts I
    4. Demystifying concepts II
2. Exercise



Text as data

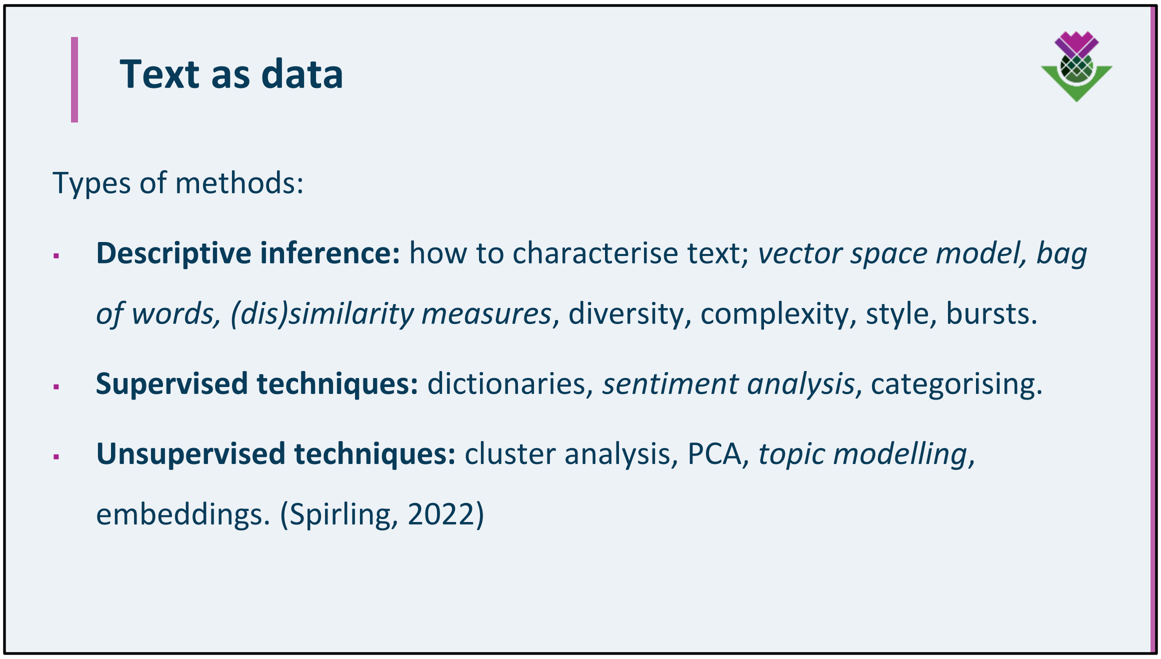
Text is the new frontier of:

- Data
- Methods
- Social Science (Spirling, 2022)

In terms of data, we inhabit an era of voluminous, readily-accessible text.

In terms of methods, we can harness a range of mature (e.g., sentiment analysis, topic modelling) and novel (e.g., word embeddings, LLMs) methods that can make sense of voluminous, unstructured text.

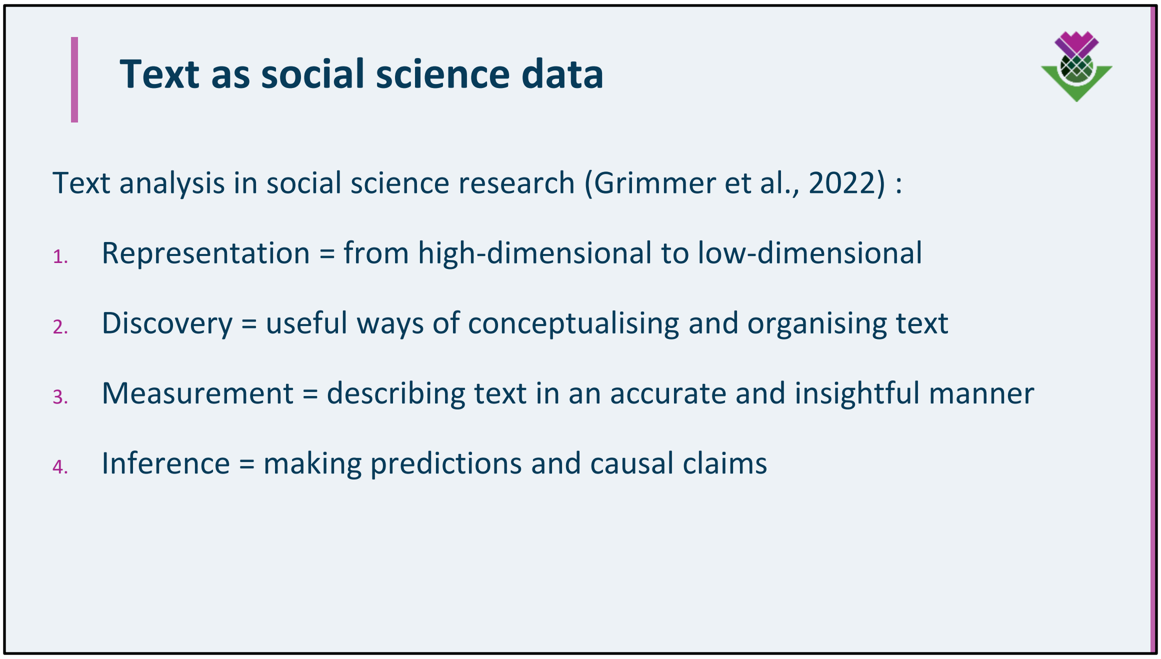
In terms of social science, we can answer existing questions at scale / granularity or new questions for the first time.



Text as data

Types of methods:

- **Descriptive inference:** how to characterise text; *vector space model, bag of words, (dis)similarity measures*, diversity, complexity, style, bursts.
- **Supervised techniques:** dictionaries, *sentiment analysis*, categorising.
- **Unsupervised techniques:** cluster analysis, PCA, *topic modelling*, embeddings. (Spirling, 2022)



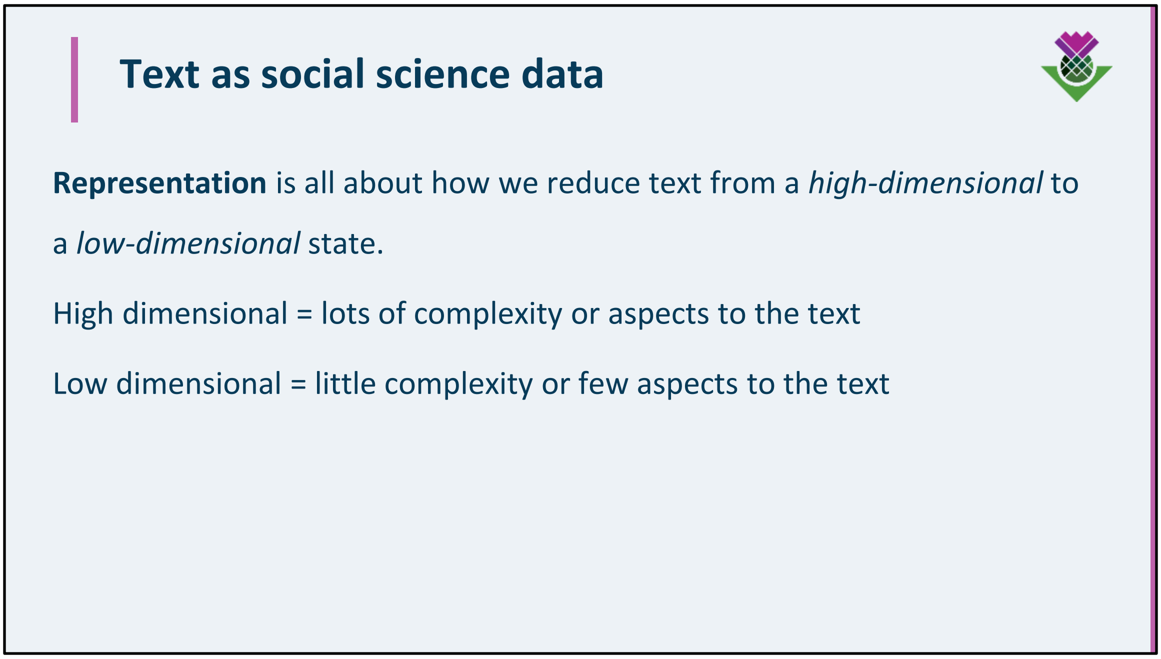
Text as social science data

Text analysis in social science research (Grimmer et al., 2022) :

1. Representation = from high-dimensional to low-dimensional
2. Discovery = useful ways of conceptualising and organising text
3. Measurement = describing text in an accurate and insightful manner
4. Inference = making predictions and causal claims

Grimmer et al. (2022) propose a common research trajectory for text analysis in the social sciences. It is sequential but iterative (e.g., where measures are refined or jettisoned depending on the results of the discovery phase).

All four phases may not be part of every project but 1 and 2 are foundational.

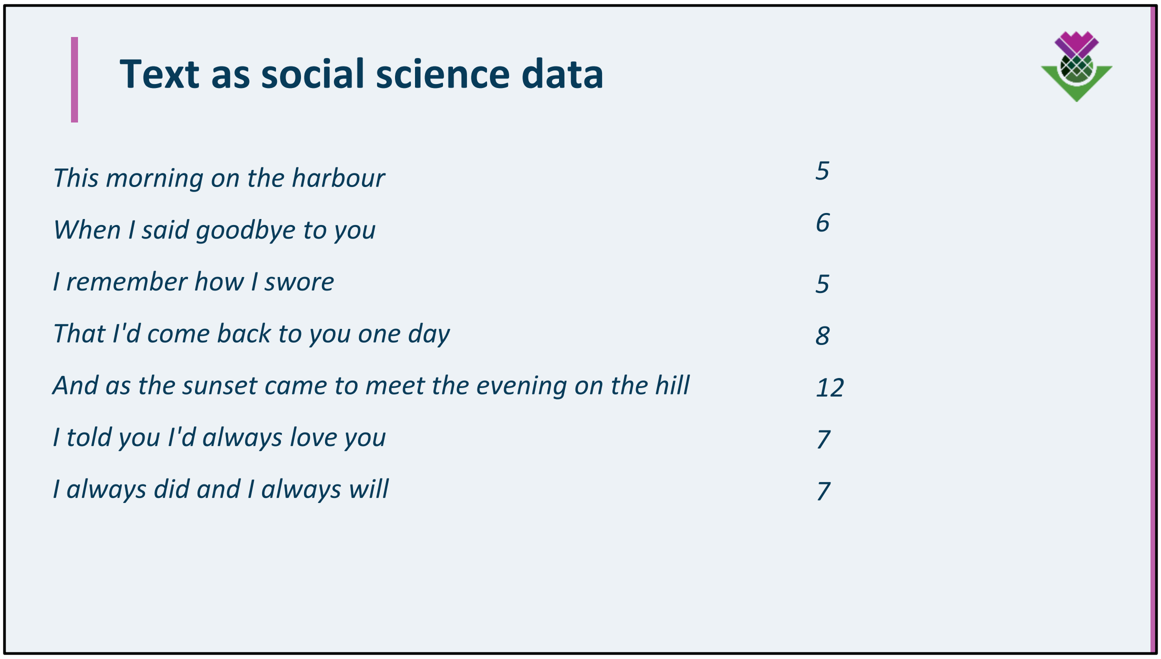


Text as social science data

Representation is all about how we reduce text from a *high-dimensional* to a *low-dimensional* state.

High dimensional = lots of complexity or aspects to the text

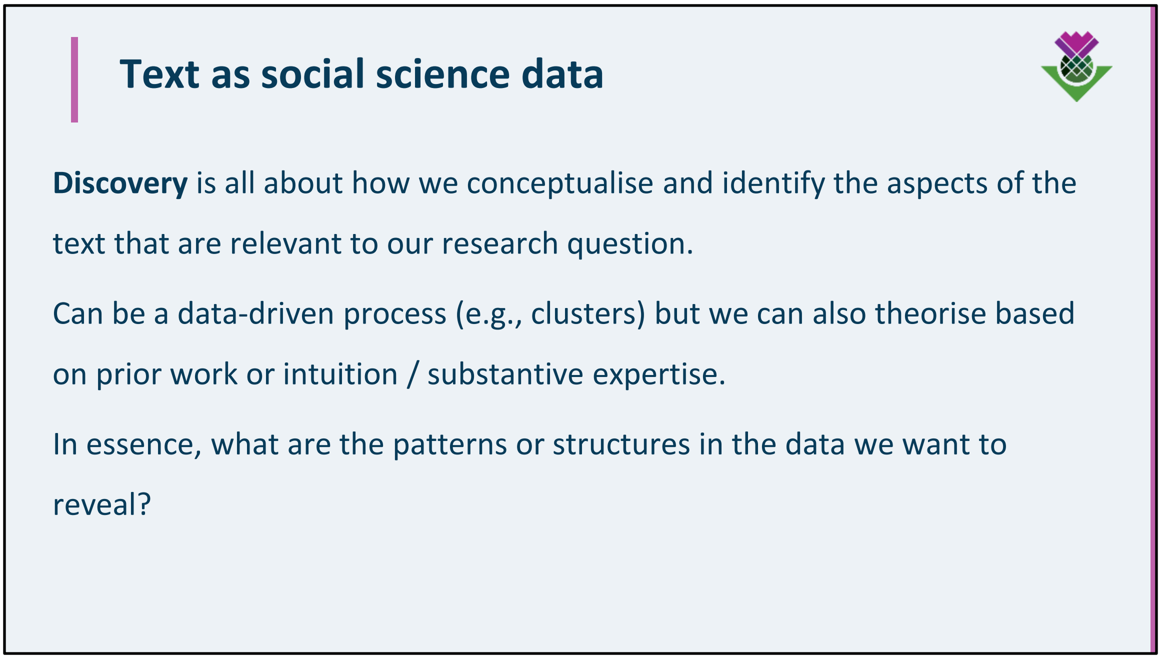
Low dimensional = little complexity or few aspects to the text



Text as social science data

This morning on the harbour	*5*
When I said goodbye to you	*6*
I remember how I swore	*5*
That I'd come back to you one day	*8*
And as the sunset came to meet the evening on the hill	*12*
I told you I'd always love you	*7*
I always did and I always will	*7*

The Body of an American (2001) by The Pogues.



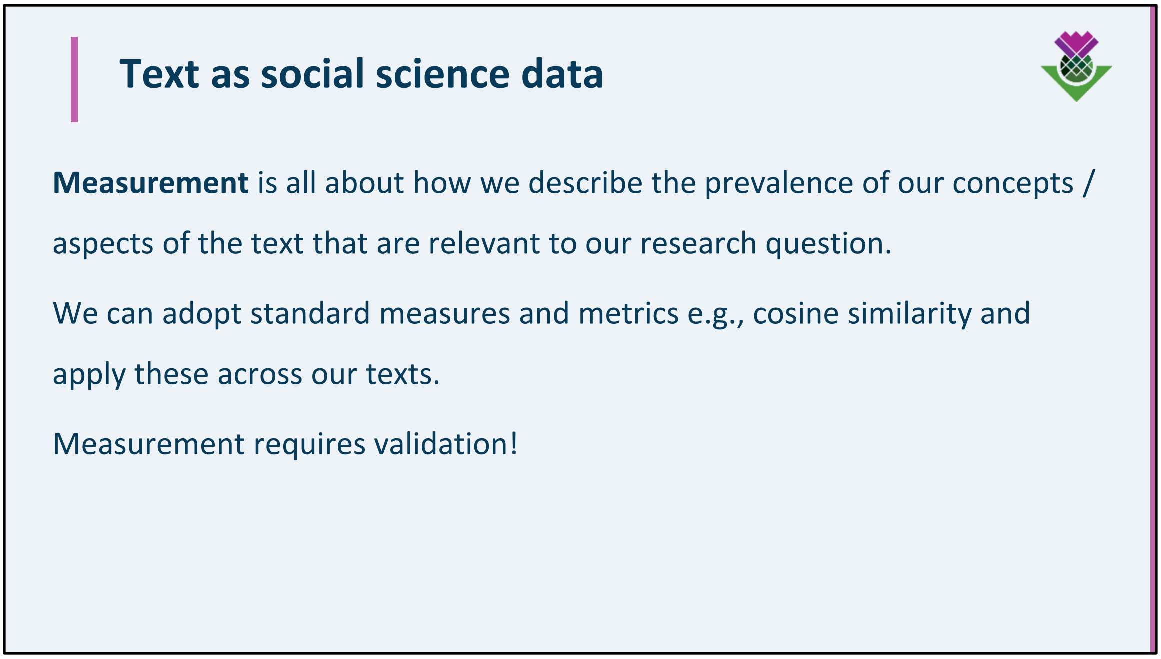
Text as social science data

Discovery is all about how we conceptualise and identify the aspects of the text that are relevant to our research question.

Can be a data-driven process (e.g., clusters) but we can also theorise based on prior work or intuition / substantive expertise.

In essence, what are the patterns or structures in the data we want to reveal?

Returning to our lyrical example, is there a pattern or structure to The Pogues songs? Do they cover similar or diverse themes? Are they structured in a similar way (e.g., number of verses and choruses)?

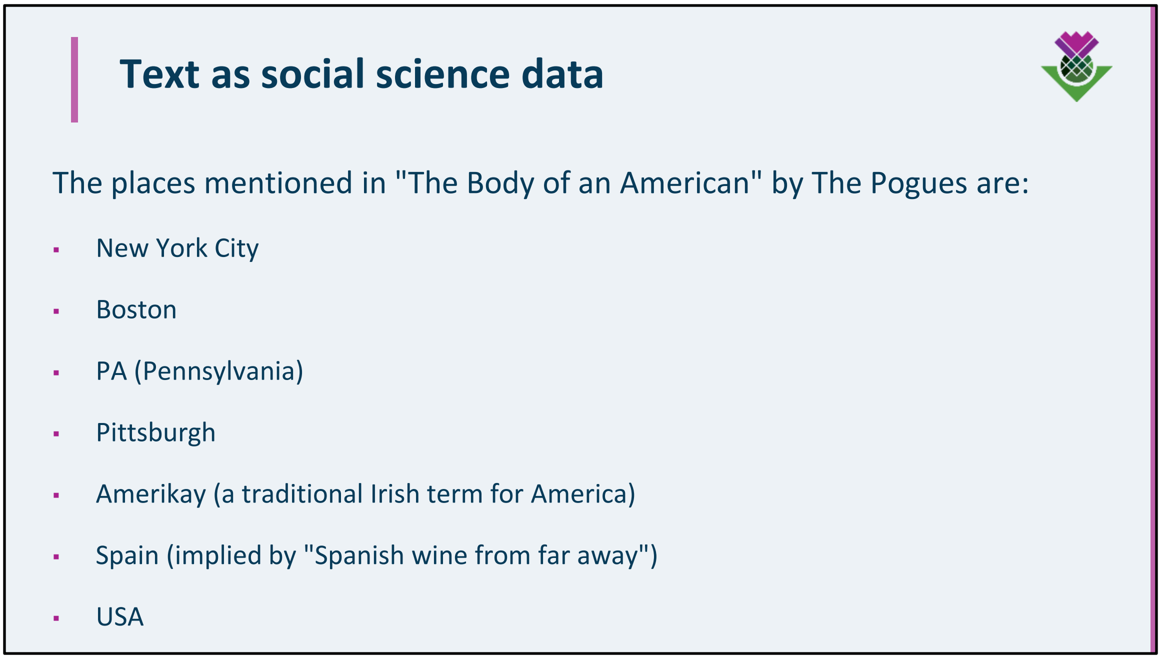


Text as social science data

Measurement is all about how we describe the prevalence of our concepts / aspects of the text that are relevant to our research question.

We can adopt standard measures and metrics e.g., cosine similarity and apply these across our texts.

Measurement requires validation!

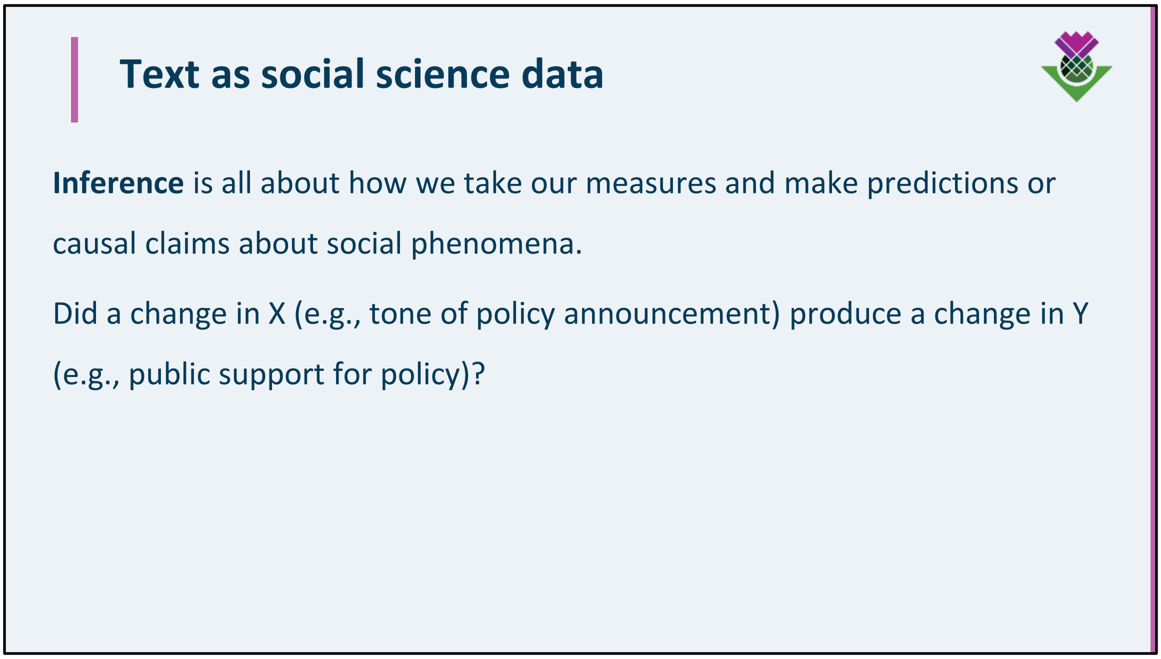


Text as social science data

The places mentioned in "The Body of an American" by The Pogues are:

- New York City
- Boston
- PA (Pennsylvania)
- Pittsburgh
- Amerikay (a traditional Irish term for America)
- Spain (implied by "Spanish wine from far away")
- USA

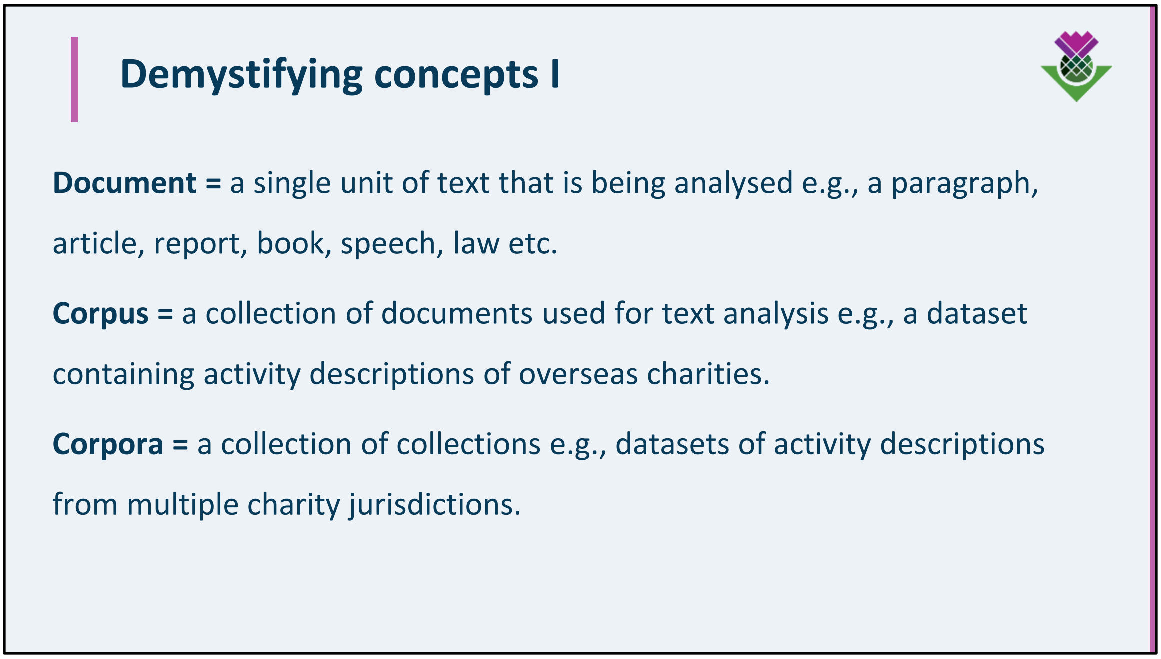
For example, if we look at the full lyrics for the song The Body of an American, there are seven places mentioned.



Text as social science data

Inference is all about how we take our measures and make predictions or causal claims about social phenomena.

Did a change in X (e.g., tone of policy announcement) produce a change in Y (e.g., public support for policy)?



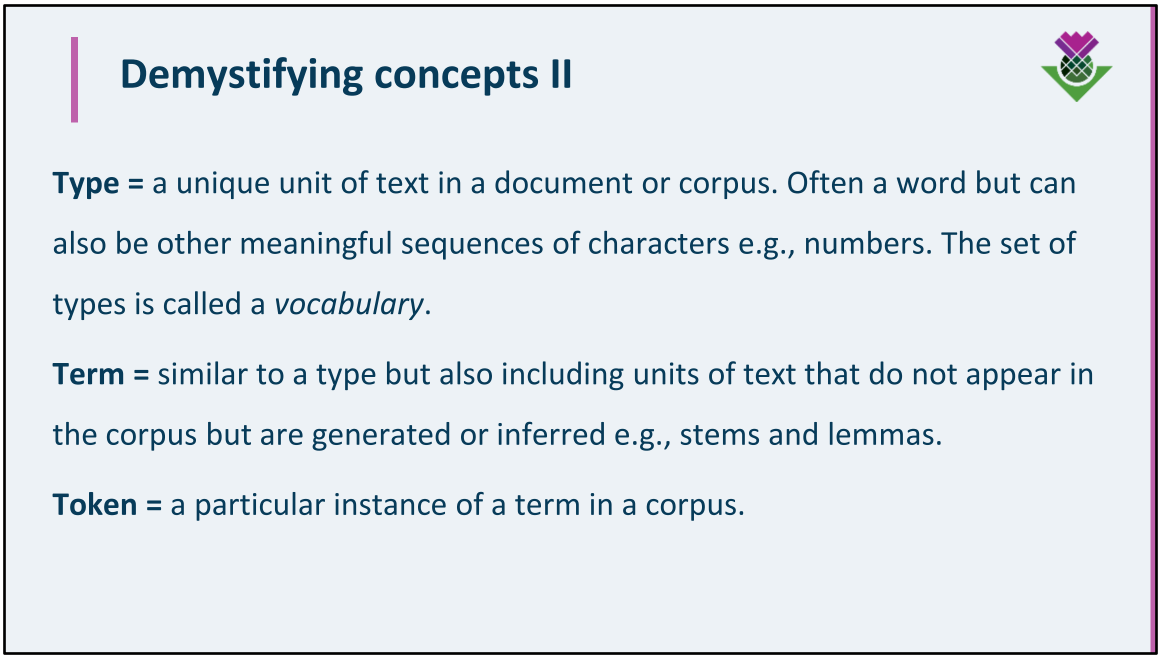
Demystifying concepts I

Document = a single unit of text that is being analysed e.g., a paragraph, article, report, book, speech, law etc.

Corpus = a collection of documents used for text analysis e.g., a dataset containing activity descriptions of overseas charities.

Corpora = a collection of collections e.g., datasets of activity descriptions from multiple charity jurisdictions.

A document is often the unit of analysis in text analysis research.

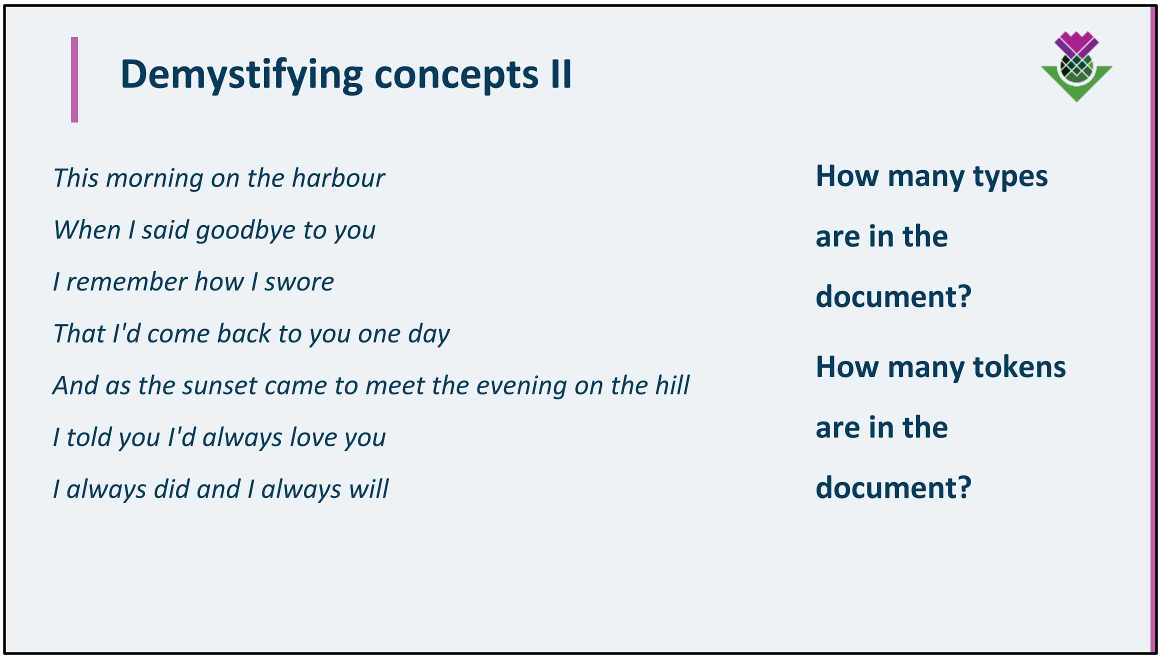


Demystifying concepts II

Type = a unique unit of text in a document or corpus. Often a word but can also be other meaningful sequences of characters e.g., numbers. The set of types is called a *vocabulary*.

Term = similar to a type but also including units of text that do not appear in the corpus but are generated or inferred e.g., stems and lemmas.

Token = a particular instance of a term in a corpus.



Demystifying concepts II

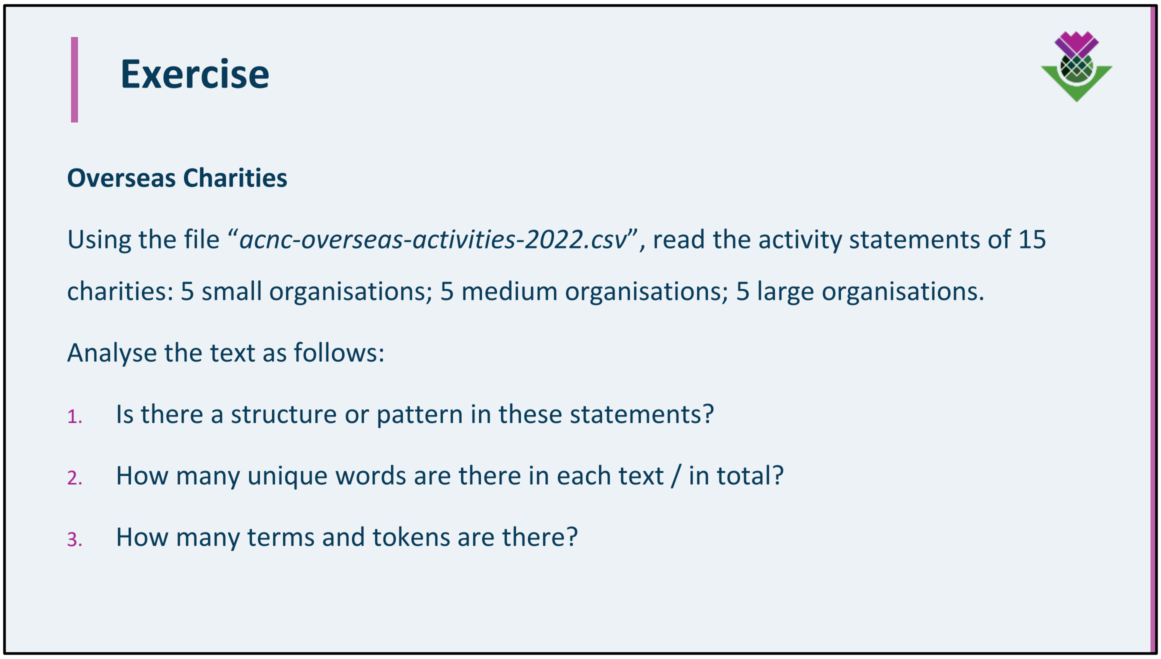
This morning on the harbour
When I said goodbye to you
I remember how I swore
That I'd come back to you one day
And as the sunset came to meet the evening on the hill
I told you I'd always love you
I always did and I always will

How many types are in the document?

How many tokens are in the document?

32 types
50 tokens

However "I" and "I'd" are counted separately. But "And" and "and" are counted together.



Exercise

Overseas Charities

Using the file “*acnc-overseas-activities-2022.csv*”, read the activity statements of 15 charities: 5 small organisations; 5 medium organisations; 5 large organisations.

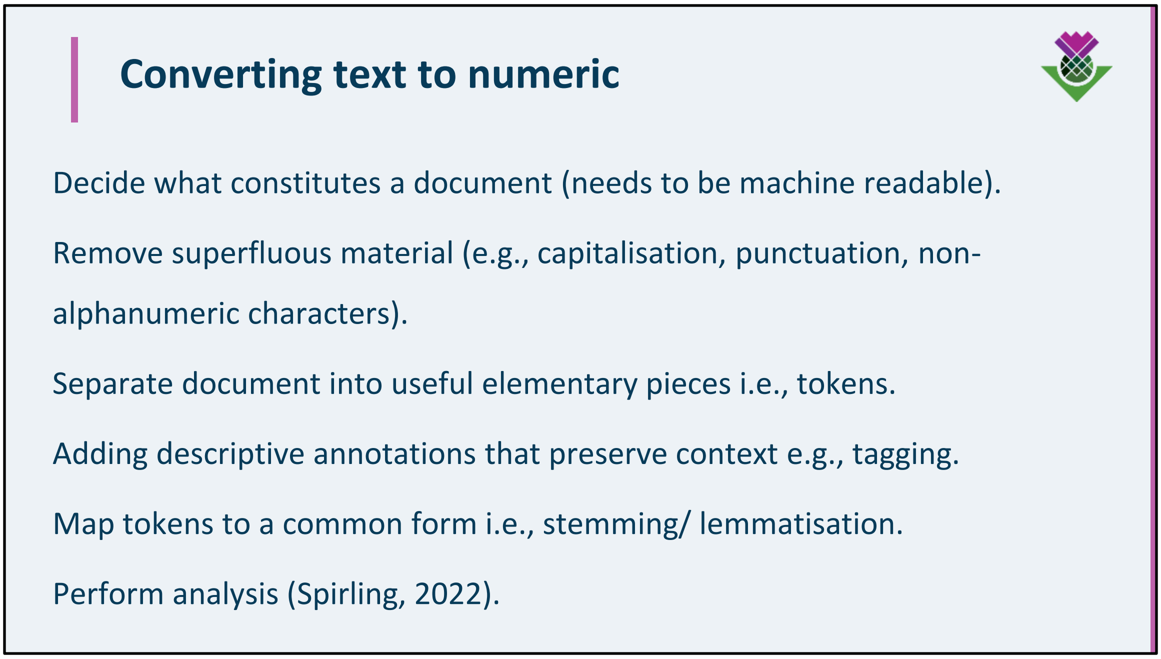
Analyse the text as follows:

1. Is there a structure or pattern in these statements?
2. How many unique words are there in each text / in total?
3. How many terms and tokens are there?

Representation

Outline

1. Converting text to numeric:
    1. Workflow
    2. Pre-processing
2. Simple representations of text:
    1. Bag of words
    2. DFM / DTM
3. Other considerations:
    1. Weighting
    2. N-grams



Converting text to numeric

Decide what constitutes a document (needs to be machine readable).

Remove superfluous material (e.g., capitalisation, punctuation, non-alphanumeric characters).

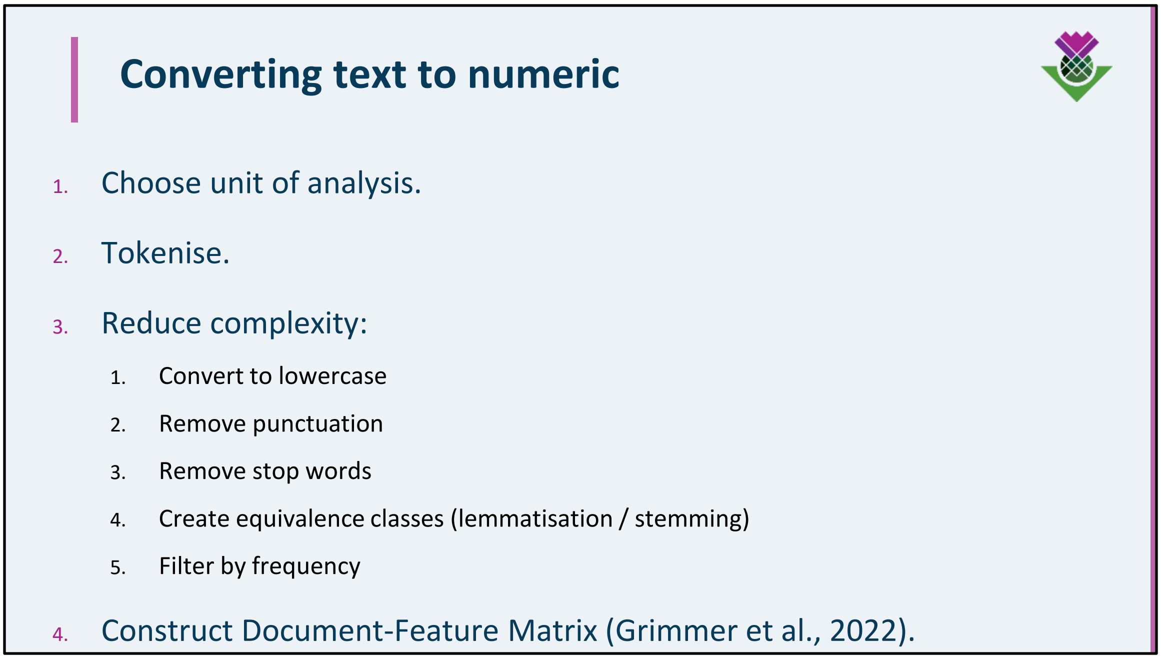
Separate document into useful elementary pieces i.e., tokens.

Adding descriptive annotations that preserve context e.g., tagging.

Map tokens to a common form i.e., stemming/ lemmatisation.

Perform analysis (Spirling, 2022).

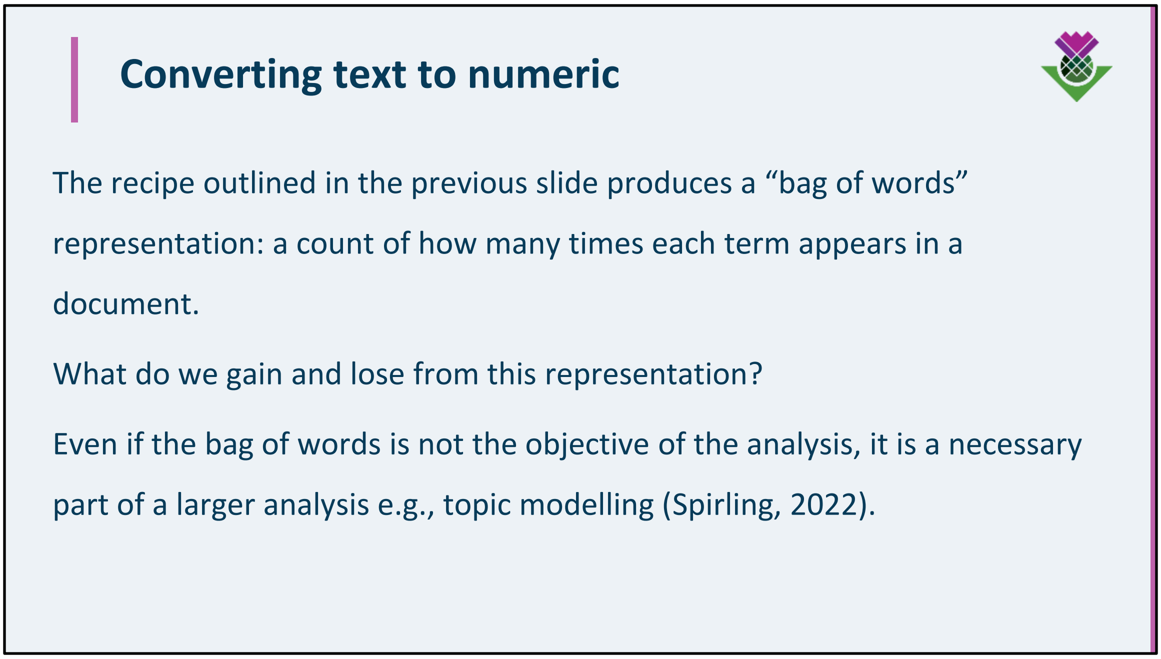
Not all of these steps are necessary for every use of text analysis, and what you do within each step may vary also (e.g., stemming vs lemmatisation). In addition, the order of these steps is your decision!



Converting text to numeric

1. Choose unit of analysis.
2. Tokenise.
3. Reduce complexity:
   1. Convert to lowercase
   2. Remove punctuation
   3. Remove stop words
   4. Create equivalence classes (lemmatisation / stemming)
   5. Filter by frequency
4. Construct Document-Feature Matrix (Grimmer et al., 2022).

This workflow or recipe is a more technical implementation of the general workflow proposed by Spirling (2022). The order also matters, as we will see in the practical. We will look at each of these steps in turn during the practical – for now let’s focus on what this recipe produces.



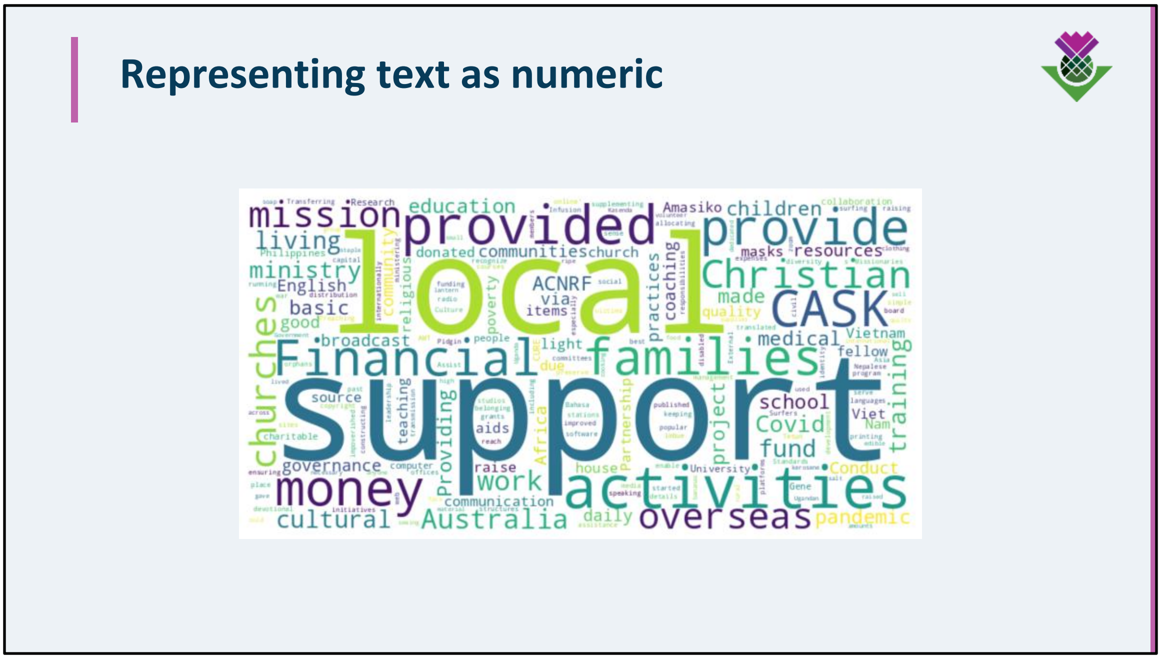
Converting text to numeric

The recipe outlined in the previous slide produces a “bag of words” representation: a count of how many times each term appears in a document.

What do we gain and lose from this representation?

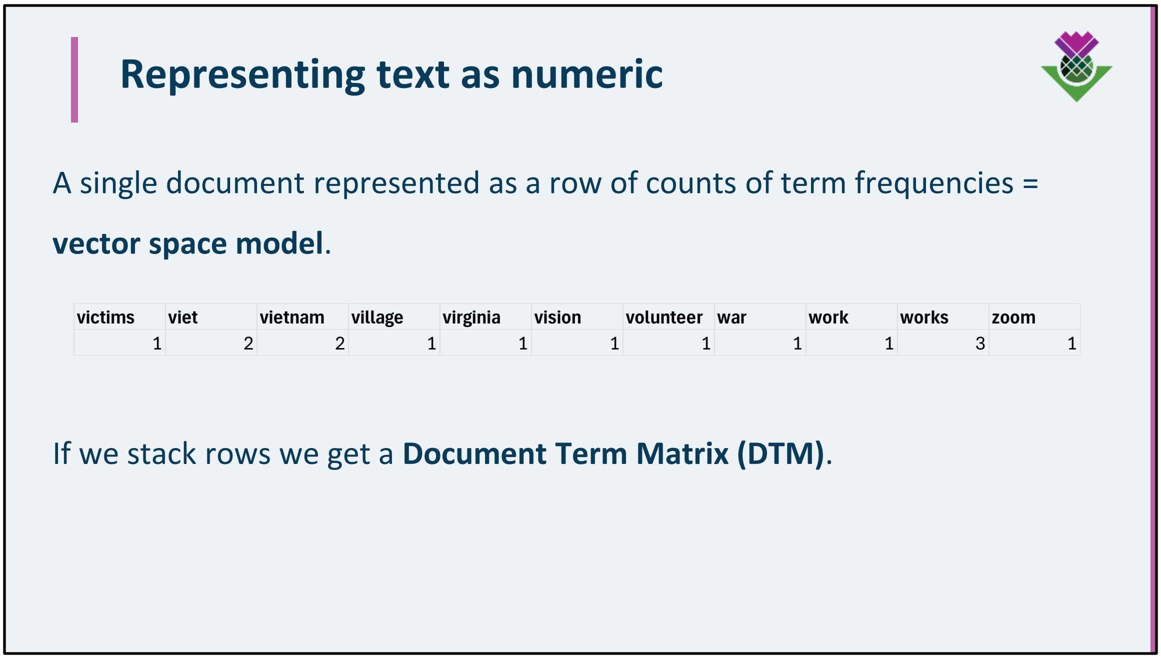
Even if the bag of words is not the objective of the analysis, it is a necessary part of a larger analysis e.g., topic modelling (Spirling, 2022).

Using the bag of words recipe means we are not interested in (or able to analyse) word order.



Representing text as numeric

Here is a word cloud visualisation of the bag of words representation of the activities of a sample of Australian charities operating overseas. What are your thoughts on the interpretation and insights of this approach?



Representing text as numeric

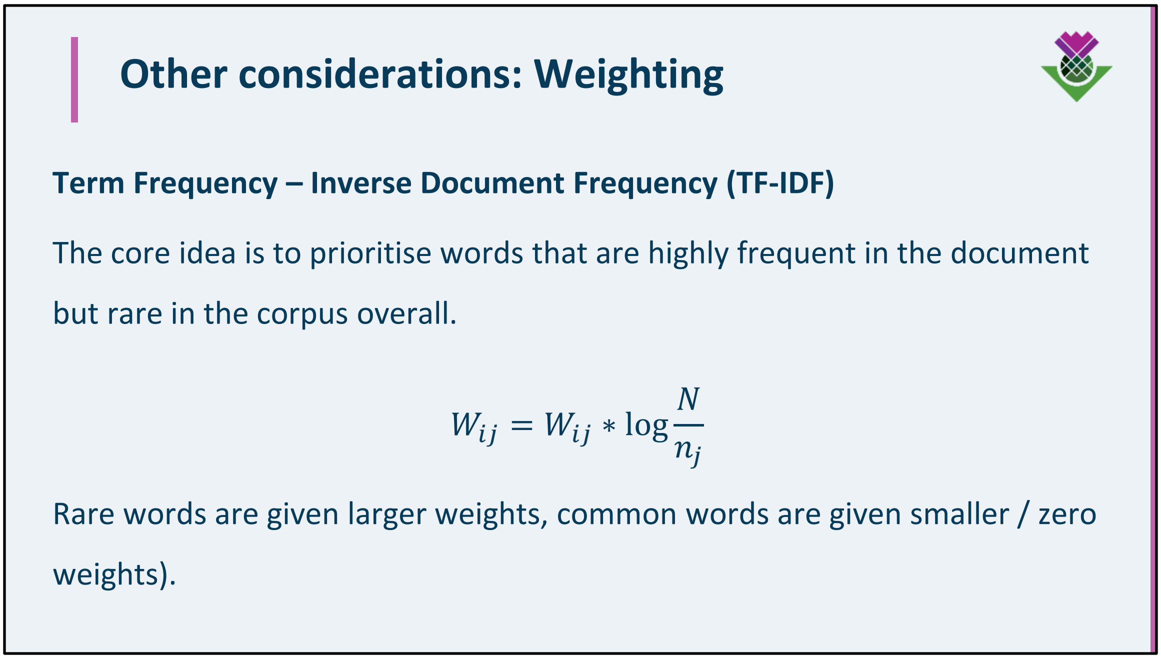
A single document represented as a row of counts of term frequencies = **vector space model**.

victims	viet	vietnam	village	virginia	vision	volunteer	war	work	works	zoom
1	2	2	1	1	1	1	1	1	3	1

If we stack rows we get a **Document Term Matrix (DTM)**.

A vector space model represents documents as a sequence of term counts. This representation is useful as we can not only compare the raw counts but also the sequence.

Word clouds and other visualisations are nice but the actual underpinning data structure / format is what’s called a Document Term Matrix / Document Feature Matrix. The features of the document are the terms found in it and the cells capture counts of those terms in the document.



Other considerations: Weighting

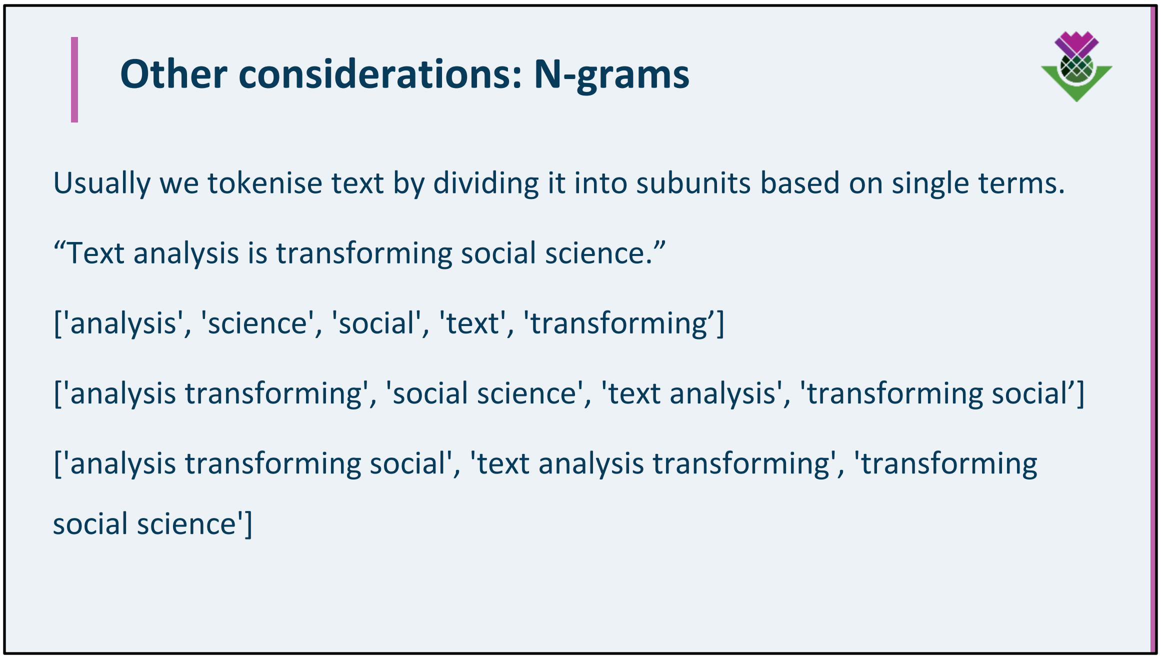
Term Frequency – Inverse Document Frequency (TF-IDF)

The core idea is to prioritise words that are highly frequent in the document but rare in the corpus overall.

$$W_{ij} = W_{ij} * \log\frac{N}{n_j}$$

Rare words are given larger weights, common words are given smaller / zero weights).

You may wonder why we adjust the raw counts of terms in the corpus? The idea is to aid our analysis by a) ensuring documents that use the same terms but in very different frequencies are still considered similar and b) improve our ability to discriminate between documents i.e., easier to identify documents that use particular words that are rare overall – good for information retrieval tasks.



Other considerations: N-grams

Usually we tokenise text by dividing it into subunits based on single terms.

“Text analysis is transforming social science.”

['analysis', 'science', 'social', 'text', 'transforming’]

['analysis transforming', 'social science', 'text analysis', 'transforming social’]

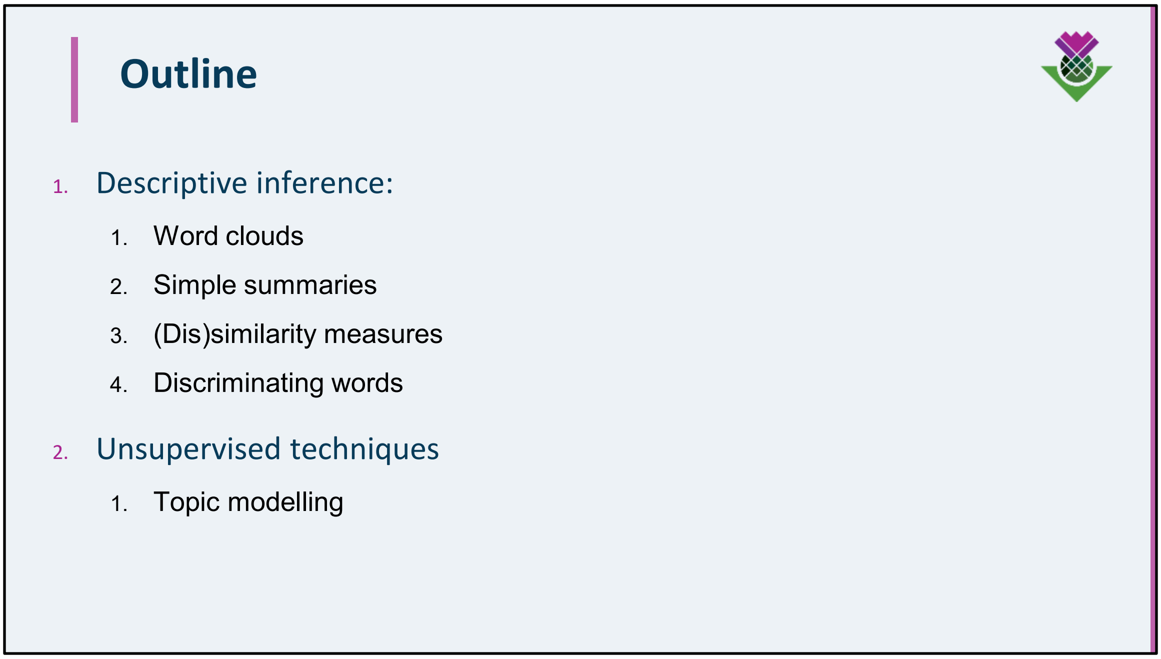
['analysis transforming social', 'text analysis transforming', 'transforming social science']

By tokenising text we create unigrams: each term is treated separately.

Bigrams and trigrams can often be often important depending on your analysis e.g., “social” and “science” will not be linked through a stem or lemma but they are substantively linked. The problem with bigrams and trigrams (and higher) is they can be computationally intensive to process.

When might you use bigrams and trigrams in your analysis? If you want to preserve word order or multiword terms.

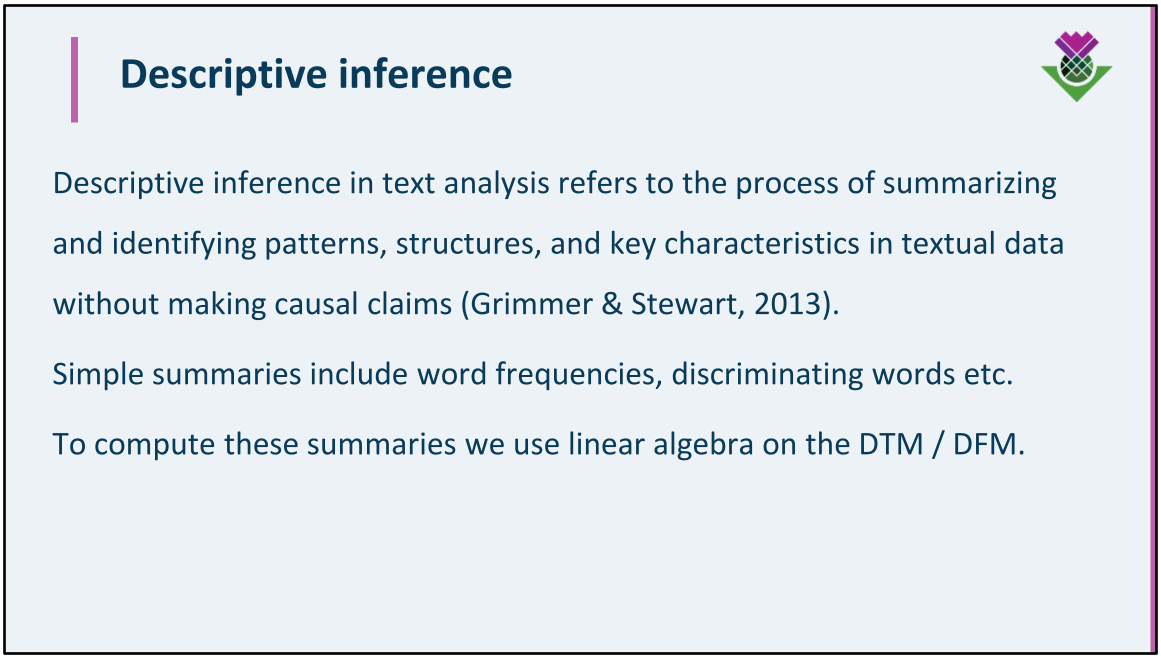
Analysis



Outline

1. Descriptive inference:
   1. Word clouds
   2. Simple summaries
   3. (Dis)similarity measures
   4. Discriminating words
2. Unsupervised techniques
   1. Topic modelling

There are so many other techniques and methods we could use. Text analysis is a huge topic. We will add to these materials over time but please suggest ideas for other approaches you would like to learn.

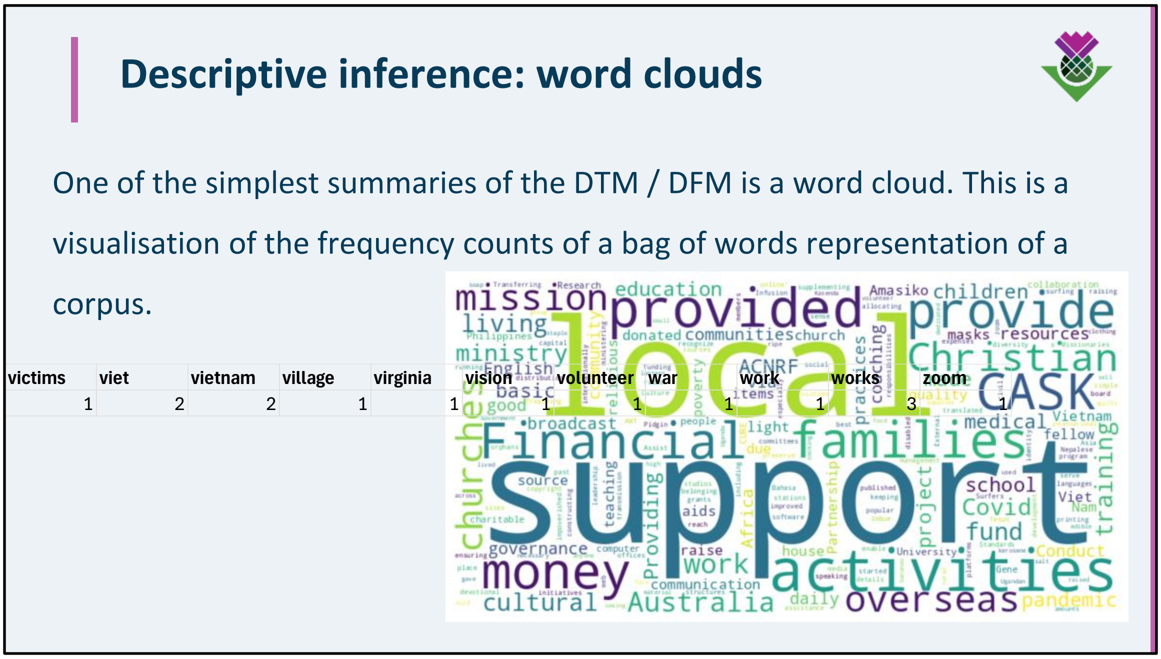


Descriptive inference

Descriptive inference in text analysis refers to the process of summarizing and identifying patterns, structures, and key characteristics in textual data without making causal claims (Grimmer & Stewart, 2013).

Simple summaries include word frequencies, discriminating words etc.

To compute these summaries we use linear algebra on the DTM / DFM.

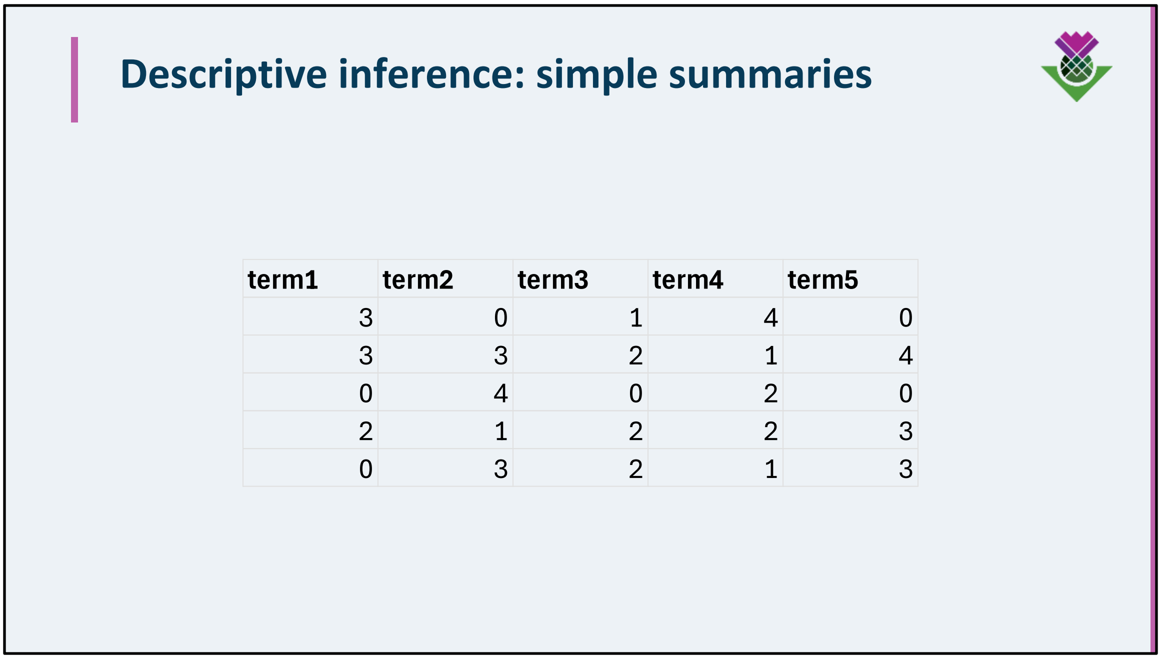


Descriptive inference: word clouds

One of the simplest summaries of the DTM / DFM is a word cloud. This is a visualisation of the frequency counts of a bag of words representation of a corpus.

victims	viet	vietnam	village	virginia	vision	volunteer	war	work	works	zoom
1	2	2	1	1	1	1	1	1	3	1

What are the advantages and disadvantages of a word cloud as a means of text analysis?



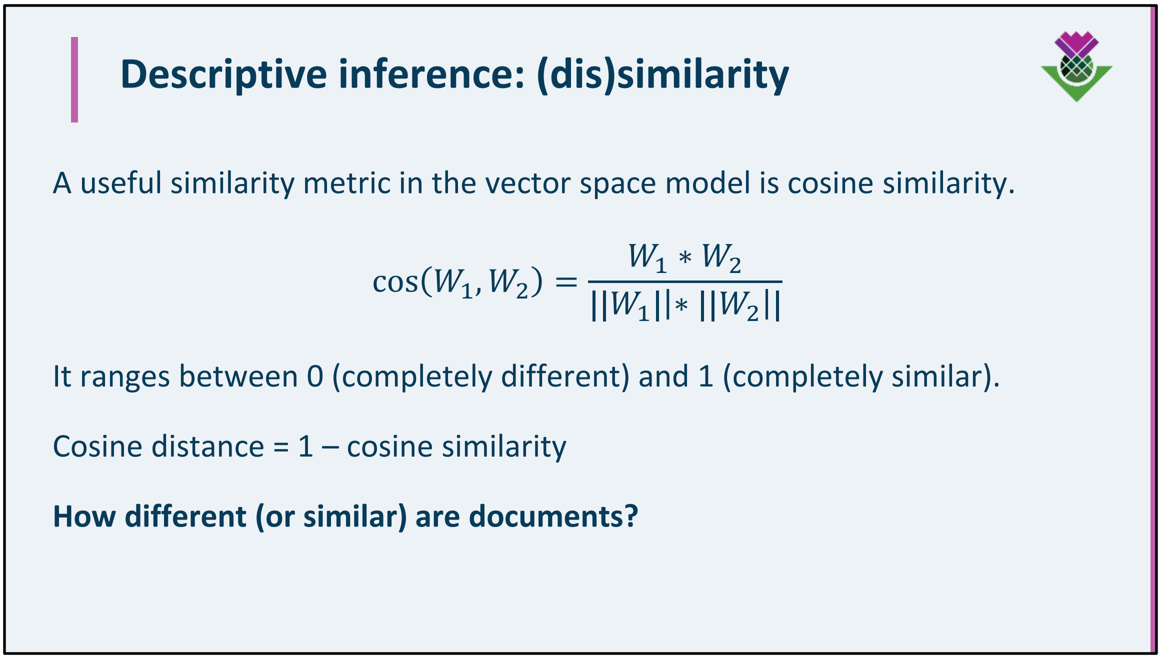
Descriptive inference: simple summaries

term1	term2	term3	term4	term5
3	0	1	4	0
3	3	2	1	4
0	4	0	2	0
2	1	2	2	3
0	3	2	1	3

Let’s say we have a simple DTM / DFM with five documents and five terms. Answer the following questions:

- How many times is term 3 mentioned in the corpus?
- How many terms and tokens are in document 5?
- Compute the row and column totals and write down their interpretations.

As you can see we can use some simple algebra to calculate some quantities or summaries of interest.



Descriptive inference: (dis)similarity

A useful similarity metric in the vector space model is cosine similarity.

$$\cos(W_1, W_2) = \frac{W_1 * W_2}{||W_1||* ||W_2||}$$

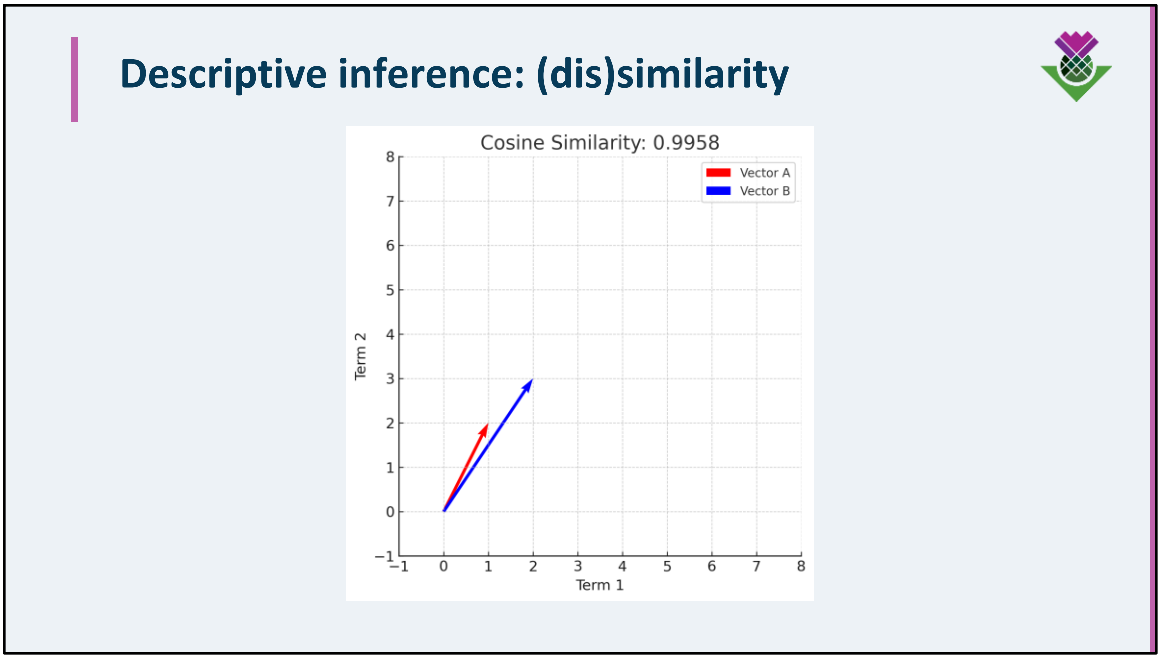
It ranges between 0 (completely different) and 1 (completely similar).

Cosine distance = 1 – cosine similarity

How different (or similar) are documents?

Where W1 is the vector of document 1 and ||W1|| is the normalised length of vector W1. It is normalised because we want to compare documents (and therefore vectors) of different lengths.

There are lots of difference measures / metrics of similarity or distance:
- Euclidean
- Manhattan
- Jaccard



Descriptive inference: (dis)similarity

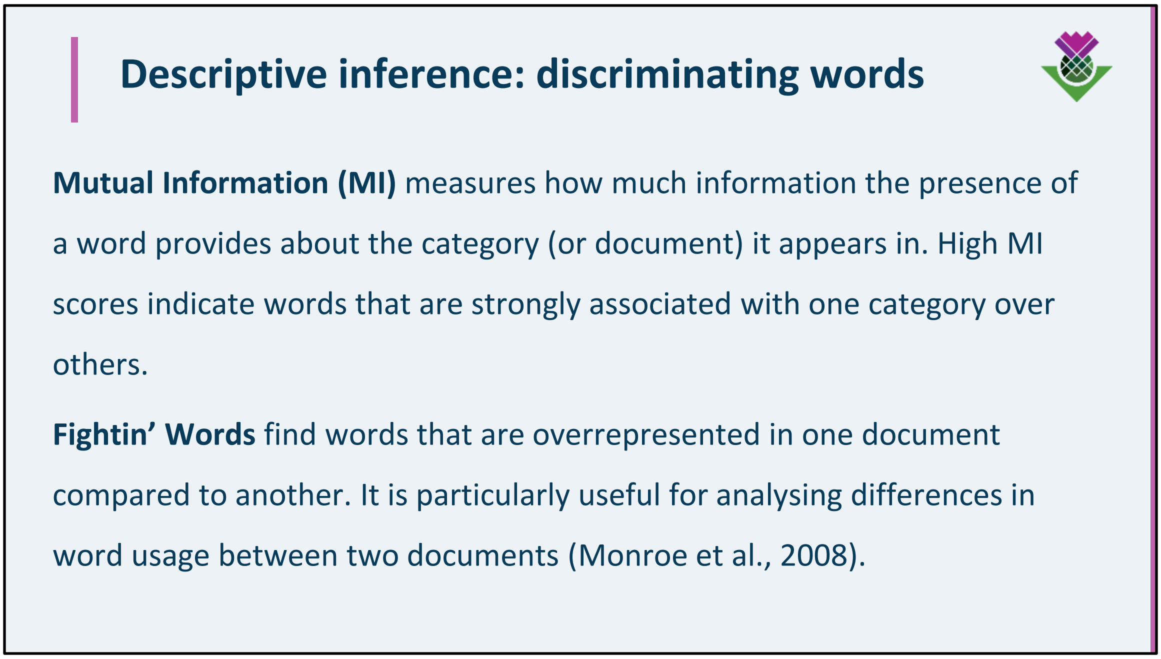
If vectors are sequences of terms, we can compute the angle between these sequences as a measure of how similar they are.

Descriptive inference: discriminating words

Discriminating words = words that characterise the language use in a group of documents in the corpus (Grimmer et al., 2022).

Words that are more prevalent in certain types of documents than others e.g., do large charities describe their overseas activities differently to medium or small organisations?

We want to explain / predict documents belonging to certain categories, rather than discover these categories.



Descriptive inference: discriminating words

Mutual Information (MI) measures how much information the presence of a word provides about the category (or document) it appears in. High MI scores indicate words that are strongly associated with one category over others.

Fightin' Words find words that are overrepresented in one document compared to another. It is particularly useful for analysing differences in word usage between two documents (Monroe et al., 2008).

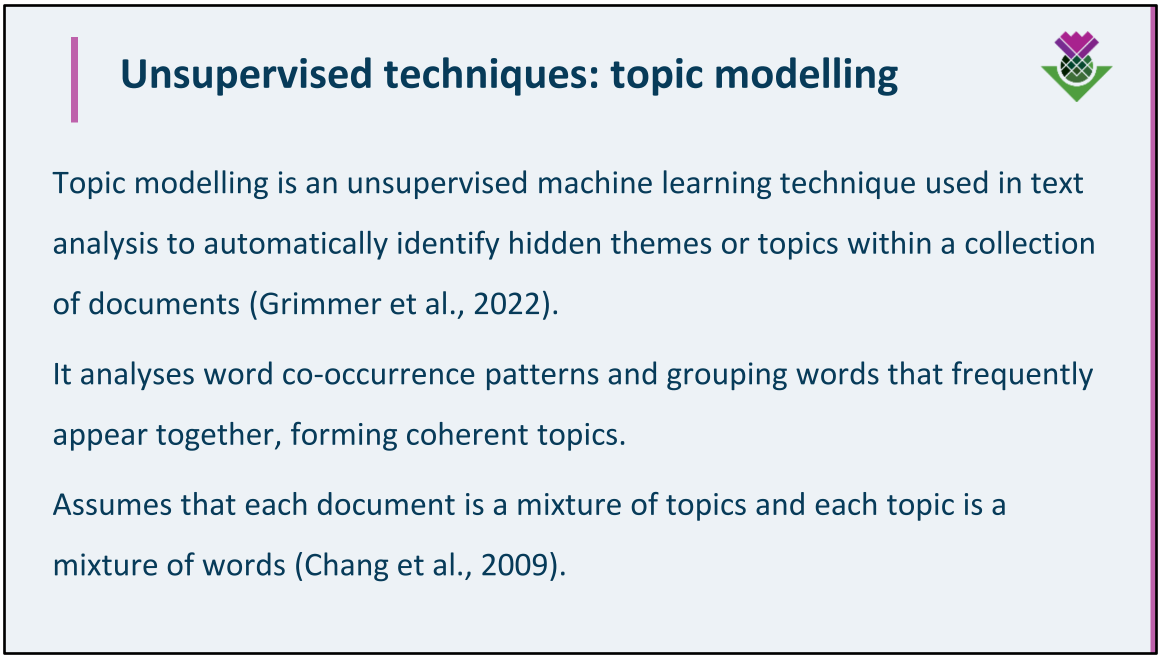
What is a limitation of MI? A downside of MI is it only considers the presence or absence of words, not occurrence (and therefore not the probability of the word occurring).

Fightin' Words = Feature Weighting using Log-Odds Ratio with Informative Dirichlet Priors

Descriptive inference: discriminating words

Considerations when using discriminating words approaches:

- Rare words: represent genuine differences in language / discourse or just random chance?
- There is a difference in words that distinguish between categories and those that are indicative of categories.



Unsupervised techniques: topic modelling

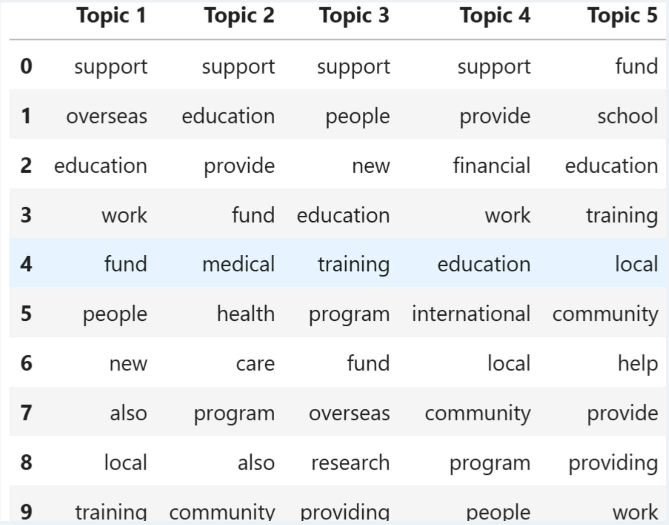
Topic modelling is an unsupervised machine learning technique used in text analysis to automatically identify hidden themes or topics within a collection of documents (Grimmer et al., 2022).

It analyses word co-occurrence patterns and grouping words that frequently appear together, forming coherent topics.

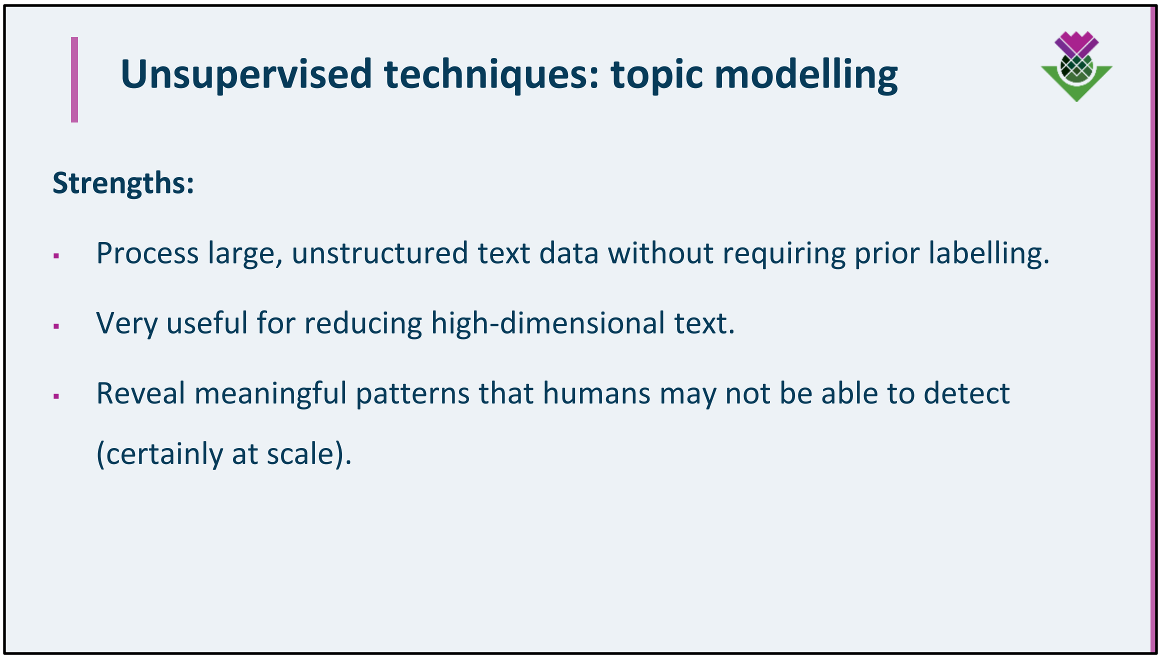
Assumes that each document is a mixture of topics and each topic is a mixture of words (Chang et al., 2009).

Therefore documents can belong to multiple topics (or classes) – in contrast to cluster analysis which assigns documents to a single class or group. You often hear topic modelling referred to as a "mixed membership" model.

Unsupervised techniques: topic modelling



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
0	support	support	support	support	fund
1	overseas	education	people	provide	school
2	education	provide	new	financial	education
3	work	fund	education	work	training
4	fund	medical	training	education	local
5	people	health	program	international	community
6	new	care	fund	local	help
7	also	program	overseas	community	provide
8	local	also	research	program	providing
9	training	community	providing	people	work

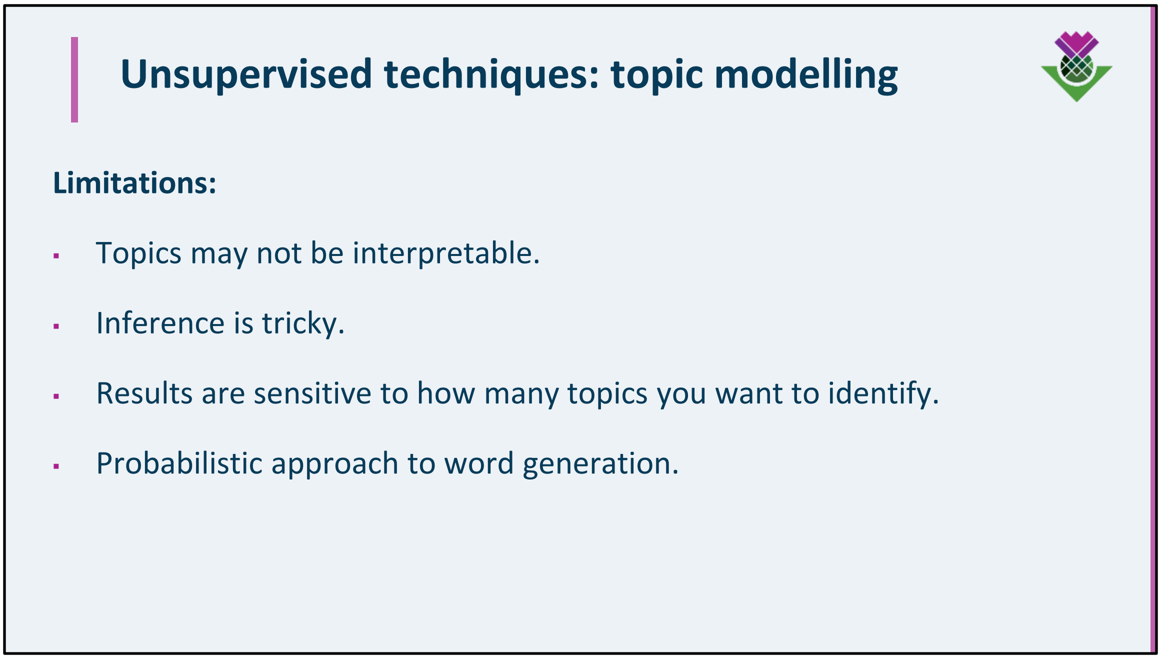


Unsupervised techniques: topic modelling

Strengths:

- Process large, unstructured text data without requiring prior labelling.
- Very useful for reducing high-dimensional text.
- Reveal meaningful patterns that humans may not be able to detect (certainly at scale).

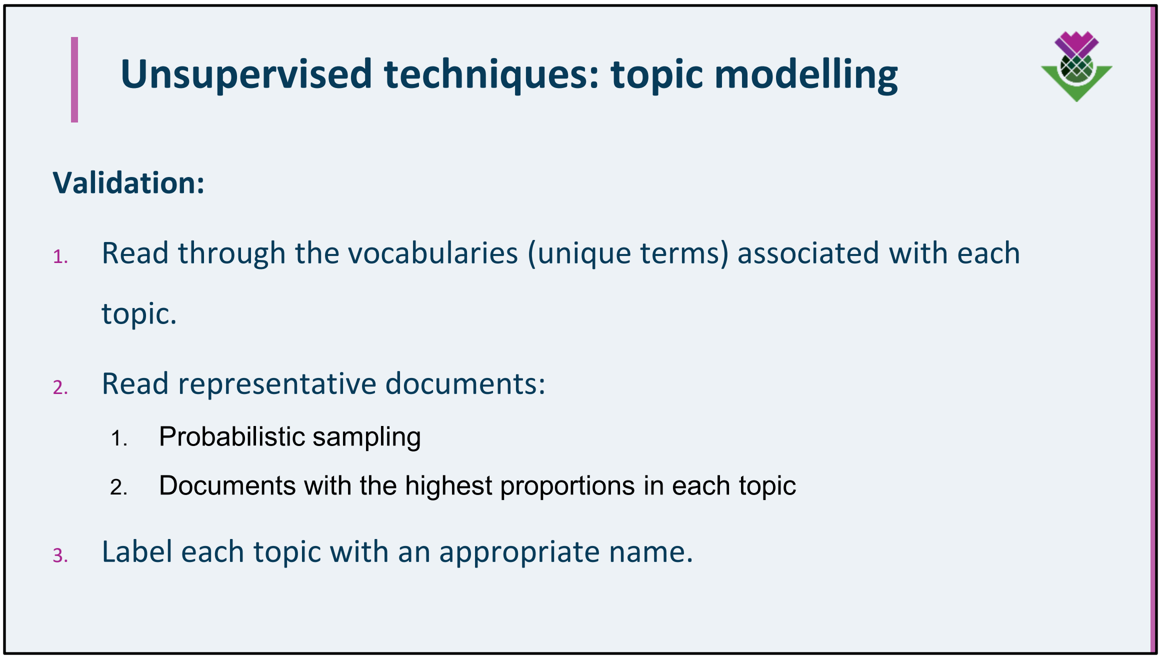
We can characterise a document well even if we have never seen another like it before (it’s an inductive, data-driven approach).



Unsupervised techniques: topic modelling

Limitations:

- Topics may not be interpretable.
- Inference is tricky.
- Results are sensitive to how many topics you want to identify.
- Probabilistic approach to word generation.



Unsupervised techniques: topic modelling

Validation:

1. Read through the vocabularies (unique terms) associated with each topic.
2. Read representative documents:
    1. Probabilistic sampling
    2. Documents with the highest proportions in each topic
3. Label each topic with an appropriate name.

Conclusion:

https://github.com/SGSSSonline/text-analysis-summer-school-2025

Contact me: diarmuid.mcdonnell@uws.ac.uk