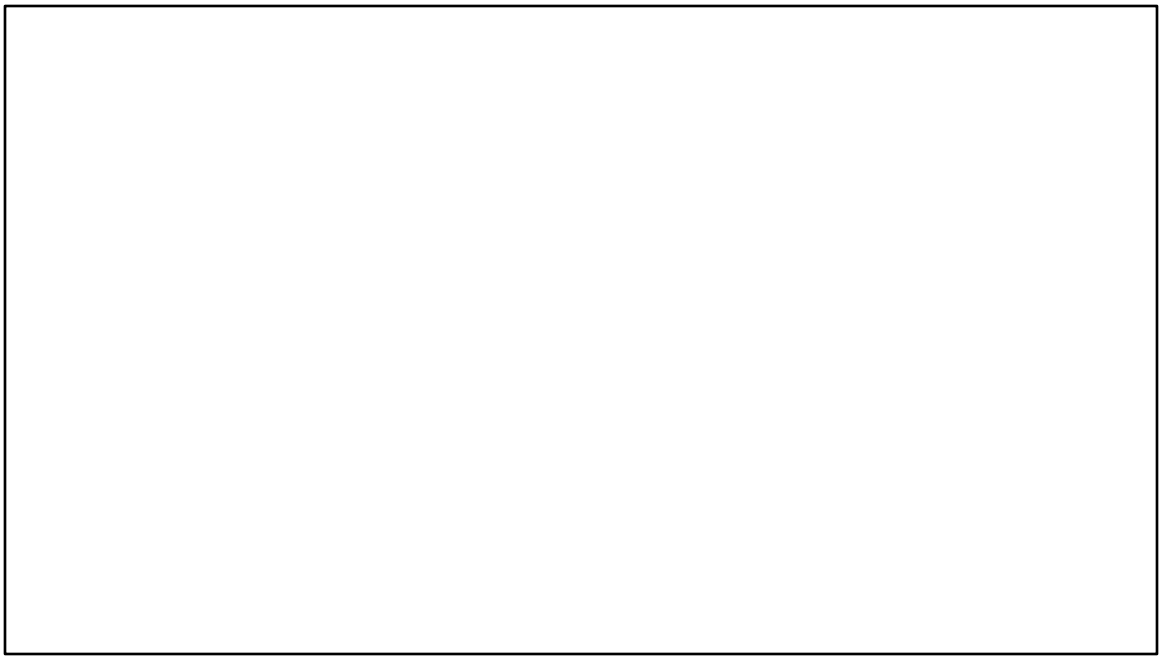
Welcome to

instats

The Session Will Begin Shortly

START



Text Analysis Using Python

Session 1: Fundamentals of Text Analysis

instats

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")

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.

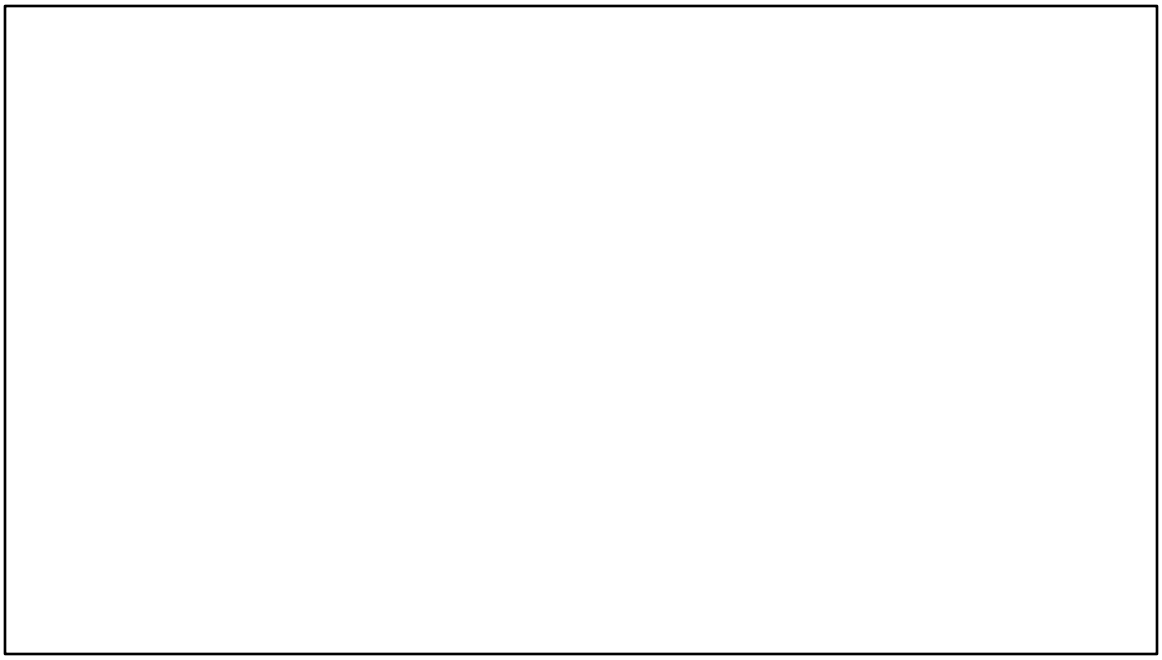
Exercise

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?



STOP