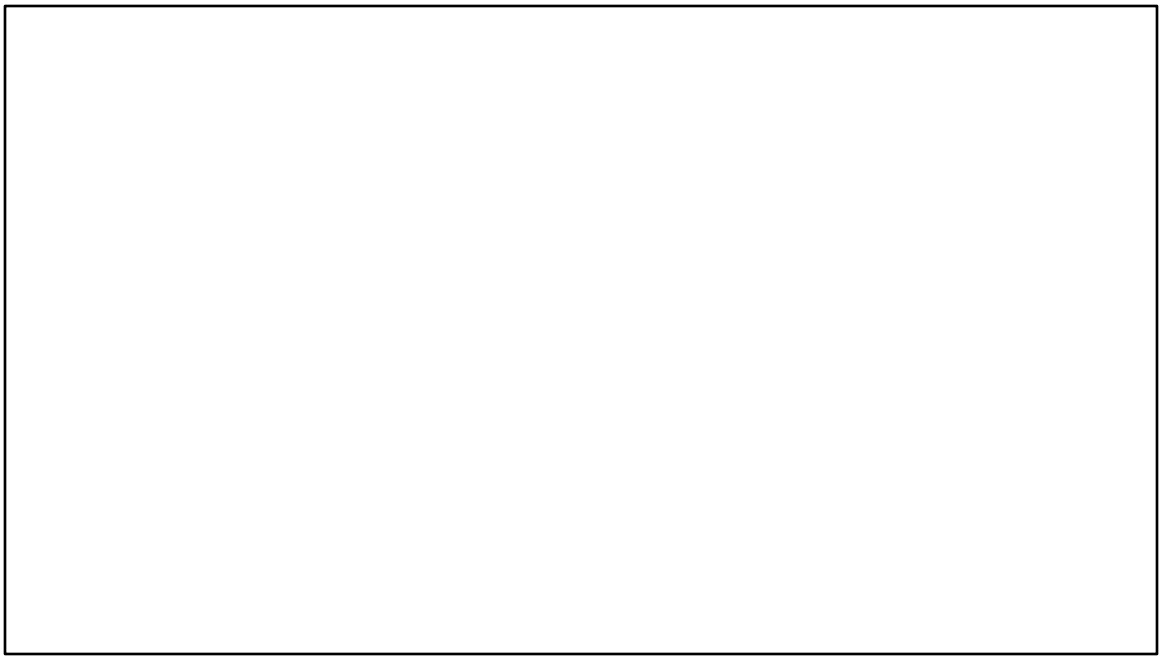
Welcome to

instats

The Session Will Begin Shortly

START



Text Analysis Using Python

Session 7: Fundamentals of Supervised Text Analysis

instats

Outline

1. Unsupervised techniques:
    1. Clustering
    2. Principal Components Analysis (PCA)
    3. Topic Modeling

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.

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

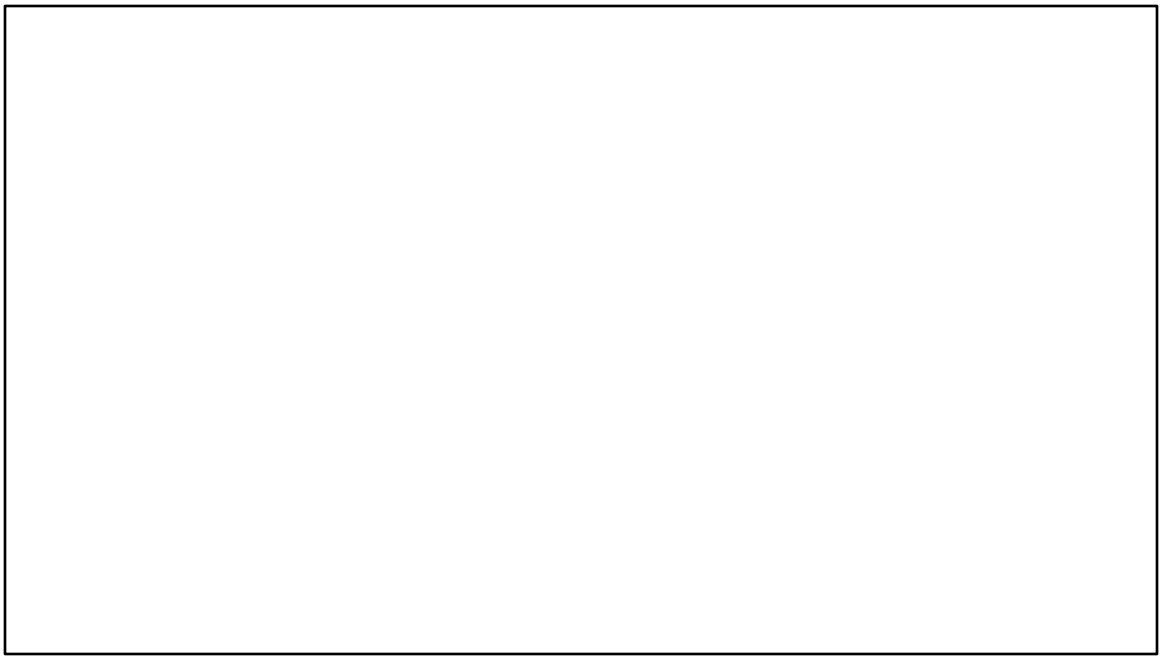
Limitations:

- 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.



STOP