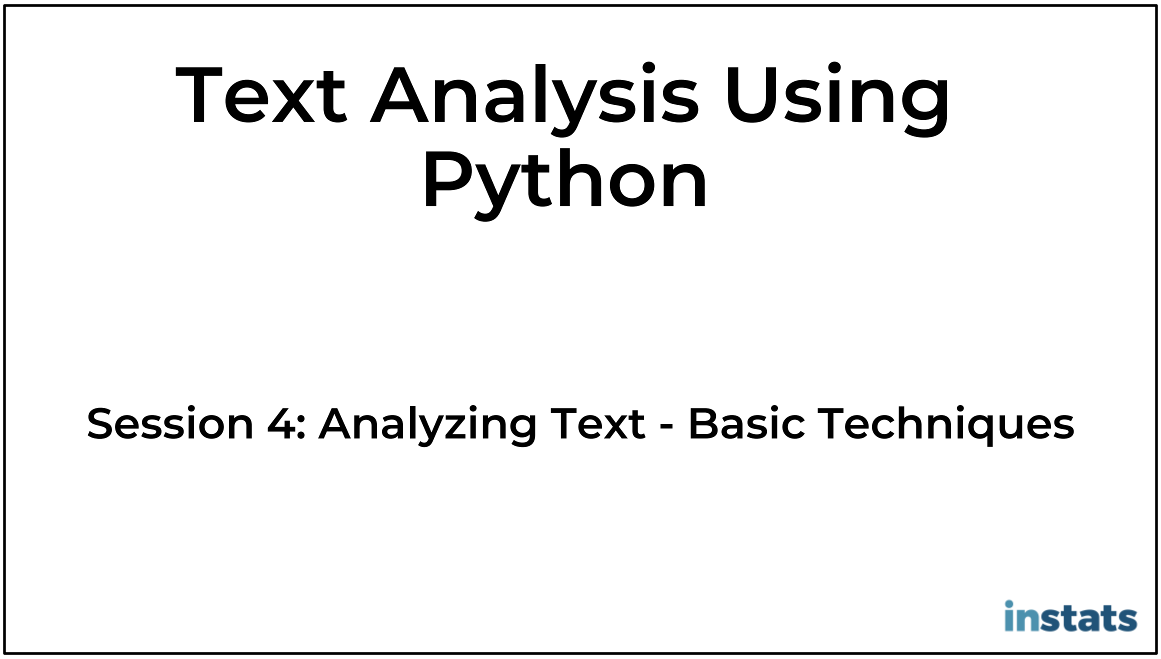
Welcome to instats

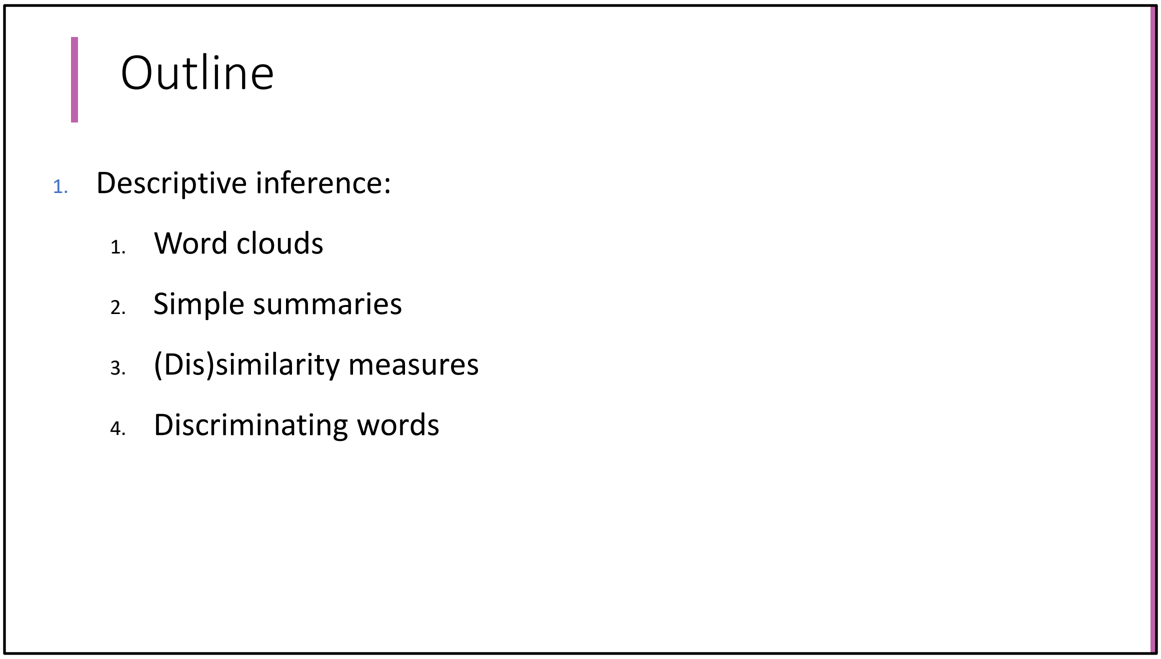
The Session Will Begin Shortly

START



Text Analysis Using Python

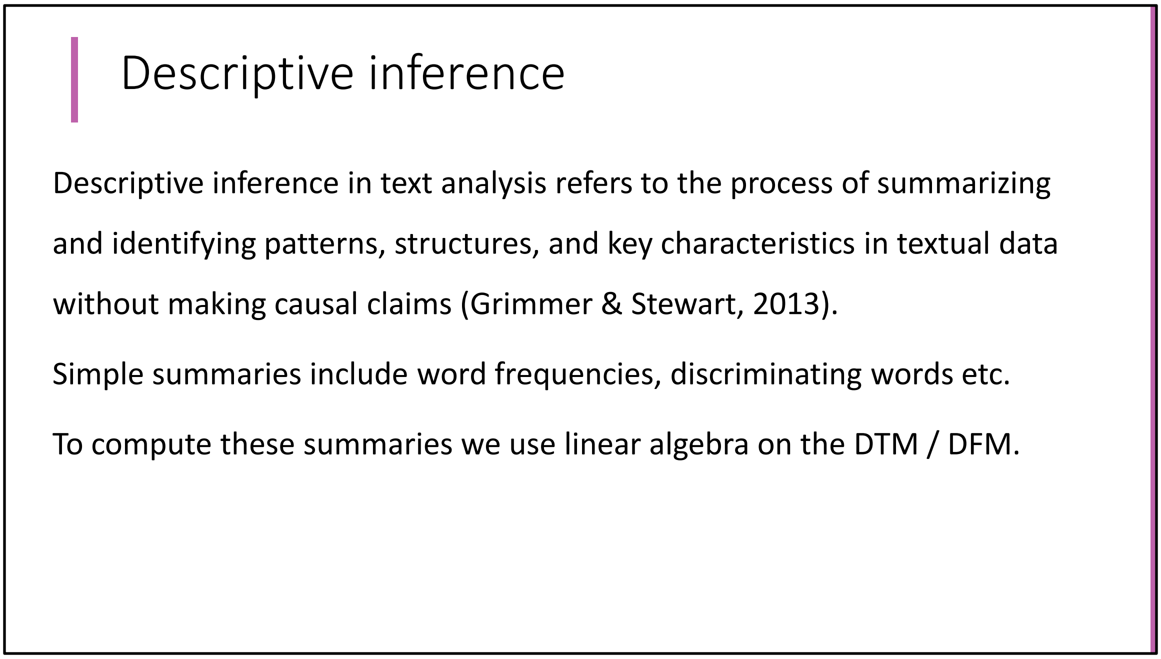
Session 4: Analyzing Text - Basic Techniques



Outline

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

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.

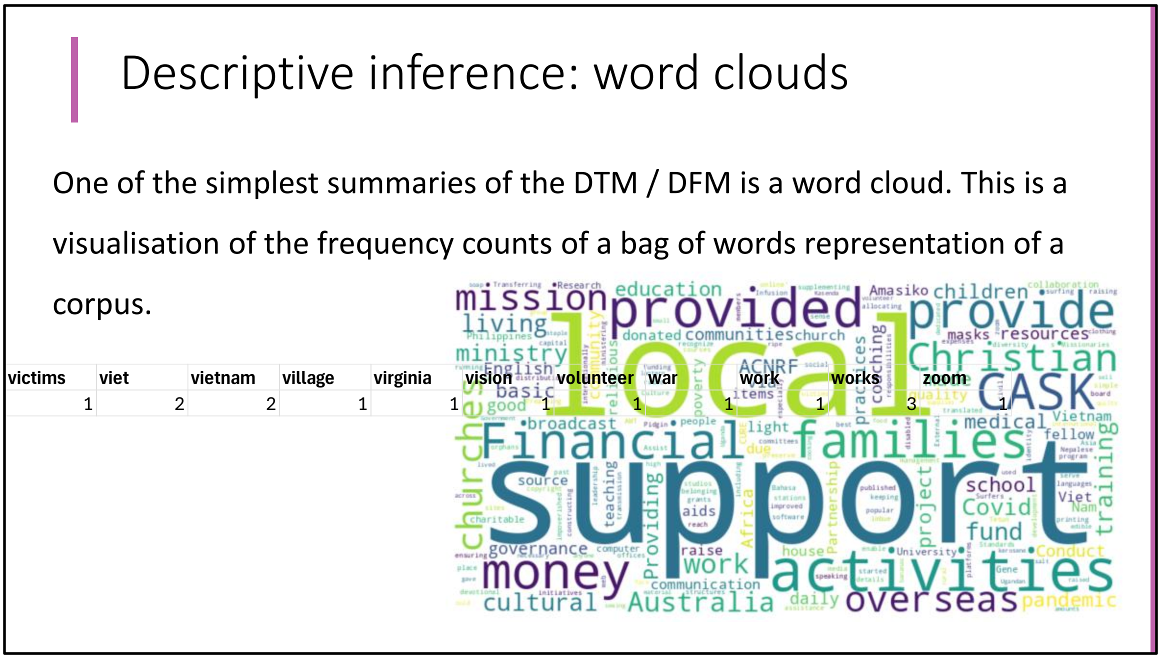


Descriptive inference

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.

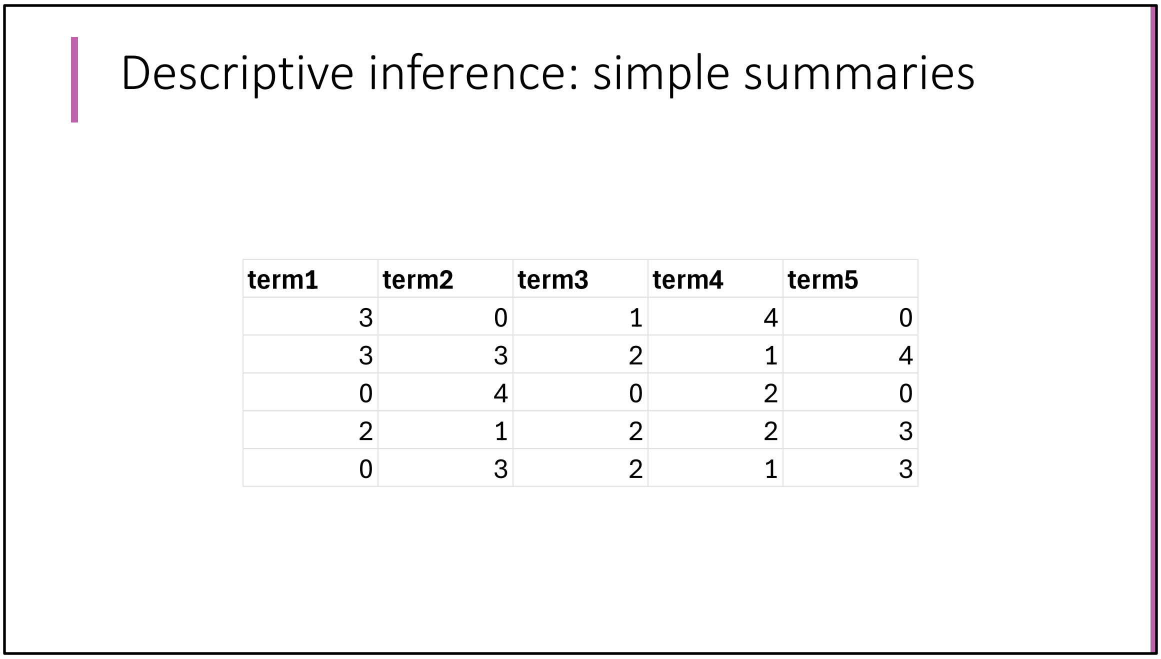


Descriptive inference: word clouds

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?



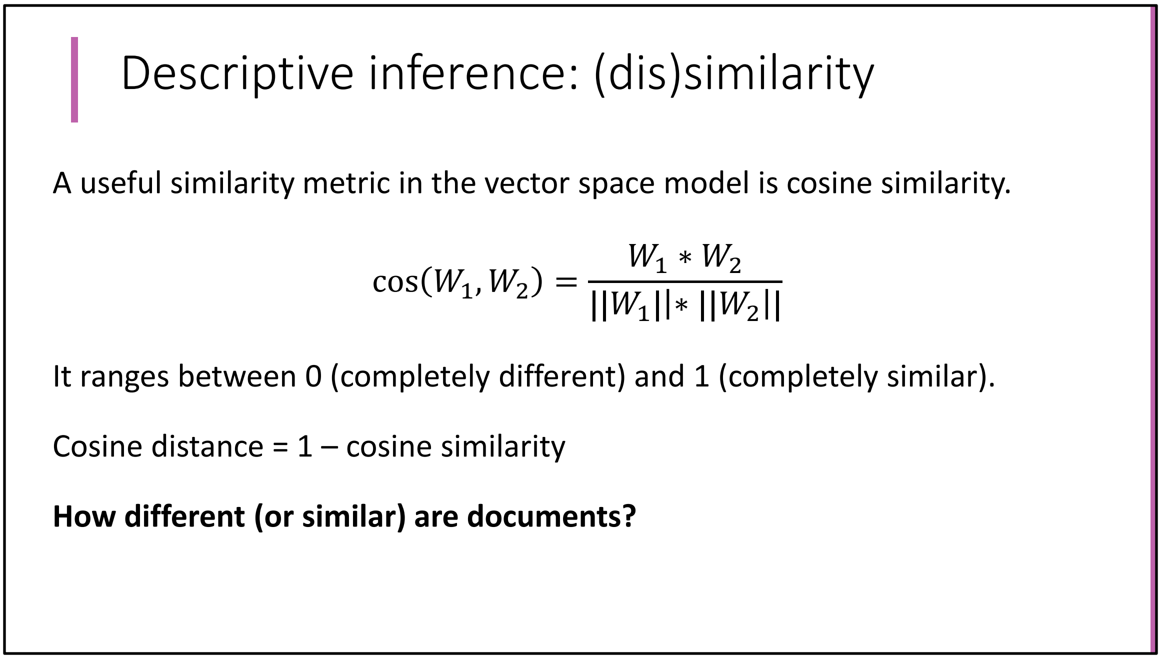
Descriptive inference: simple summaries

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).

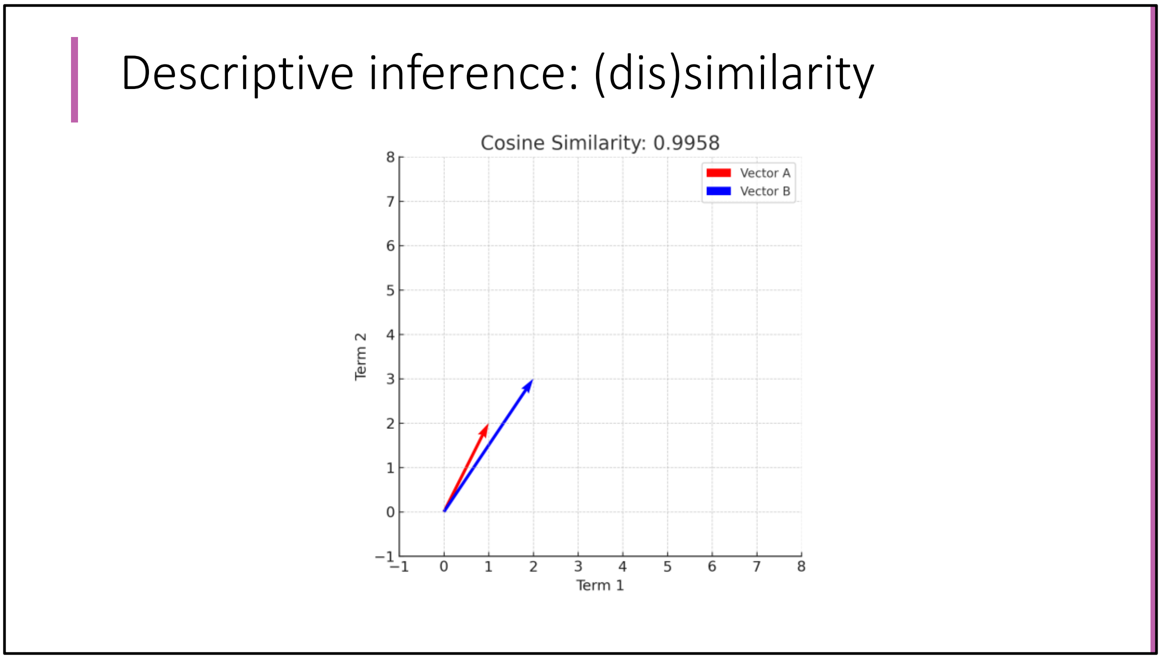
Cosine distance = 1 – cosine similarity

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.

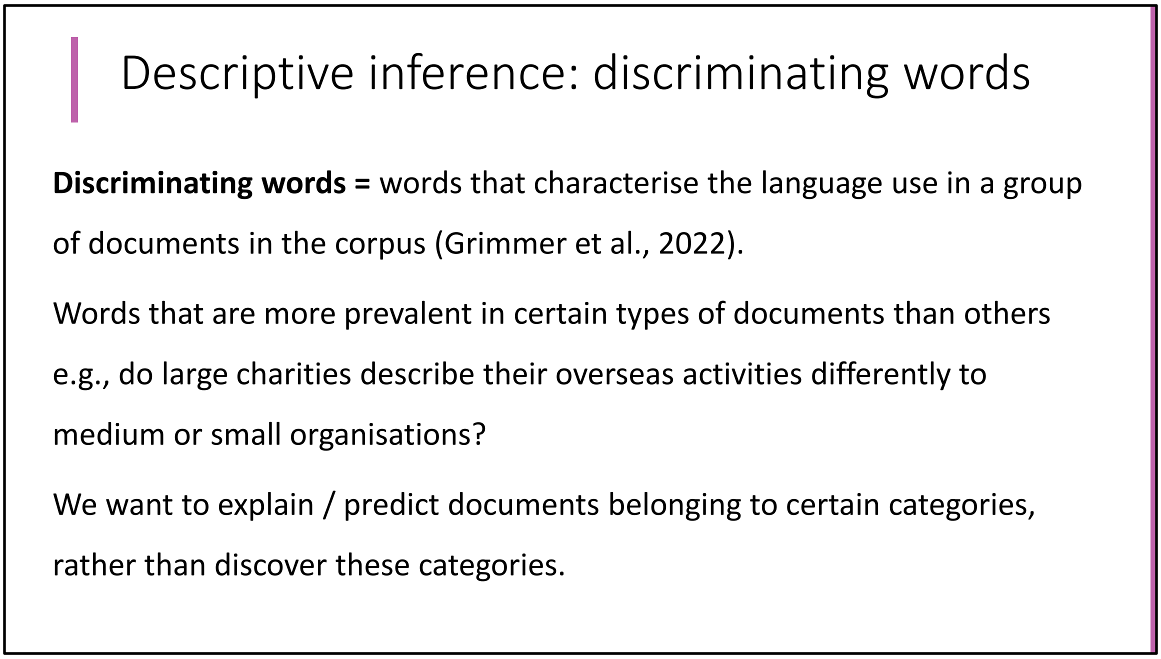
There are lots of difference measures / metrics of similarity or distance:

- 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.

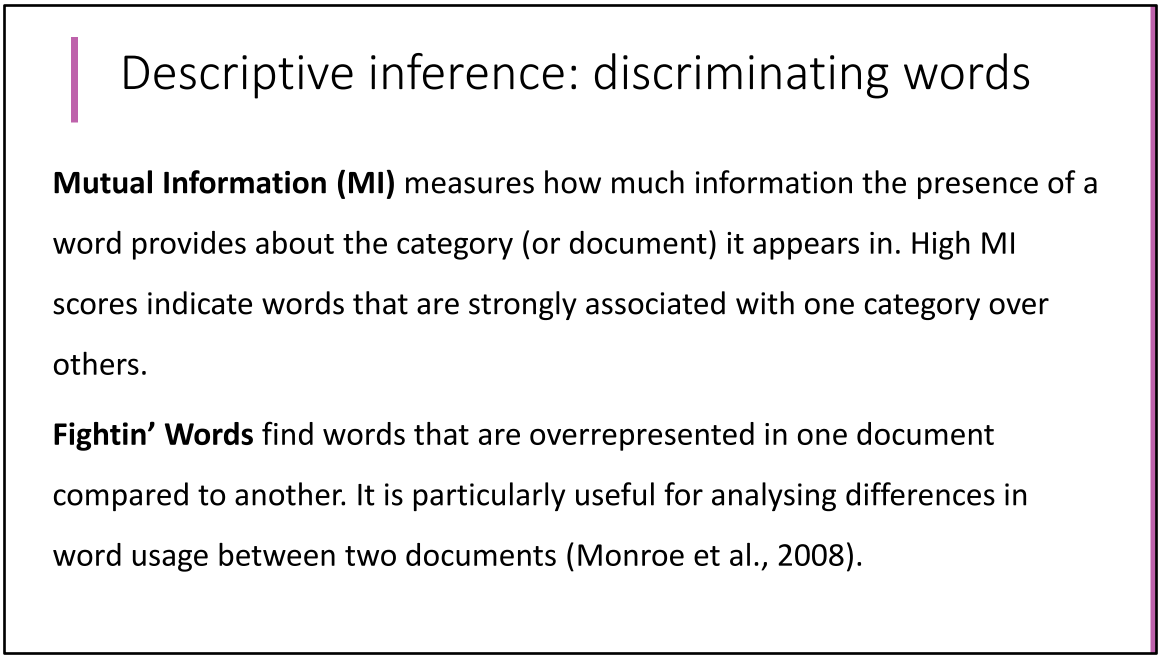


Descriptive inference: discriminating words

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.



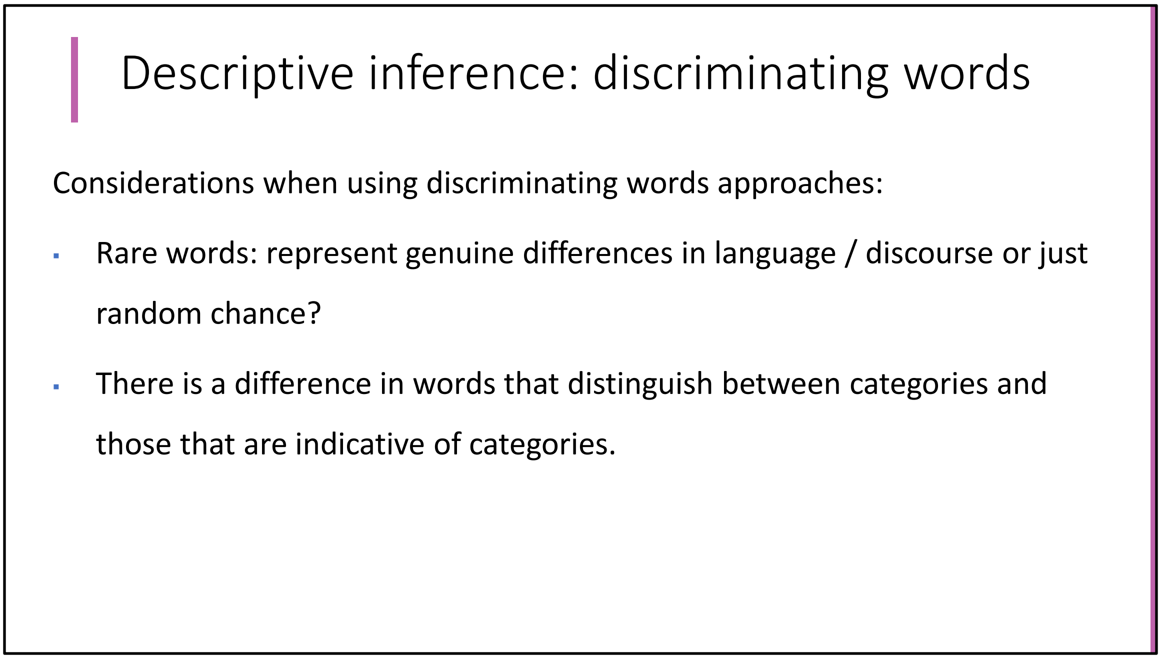
Descriptive inference: discriminating words

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).

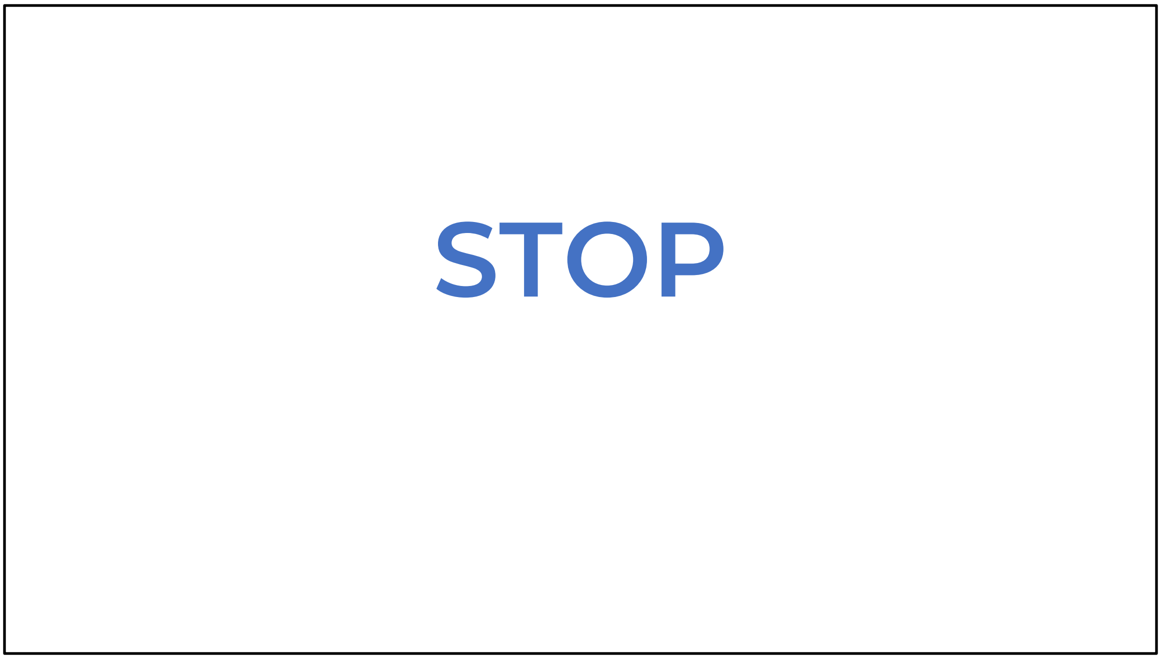
Fightin' Words **=** Feature Weighting using Log-Odds Ratio with Informative Dirichlet Priors



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.



STOP