# Applied NLP Session 2 — Professional Summary

—

Youssif Hebesha

# Overwiew

- N-gram analysis (bigrams & trigrams)
- PMI (collocation association)
- Phrase diversity (TTR)

# Bigrams & Trigrams

**What it does:**

- Breaks the book into pairs of words (**bigrams**) and triplets of words (**trigrams**).
- Counts how often each pair or triplet appears.
- Shows which phrases the author uses most often.

**Why it matters:**

It helps you see the author’s writing patterns, like repeated phrases or common expressions.

# PMI Analysis

**What it does:**

- Looks at how strongly two words are connected.
- Not just *what appears most often*, but *which words appear together more than expected*.
- Example: “Mrs. Bennet” appears together a lot → high PMI.

**Why it matters:**

PMI finds meaningful word relationships—words that tend to appear side-by-side in the story.

# Phrase Diversity

**What it does:**

- Measures **how diverse** the author's phrasing is.
- Compares the number of *unique* phrases to the *total* number of phrases.
- Does this for:
  - 1-grams (single words)
  - 2-grams
  - 3-grams
  - etc.

**Why it matters:**

A high diversity score = the author uses many different phrases
A low score = the author repeats phrases more often