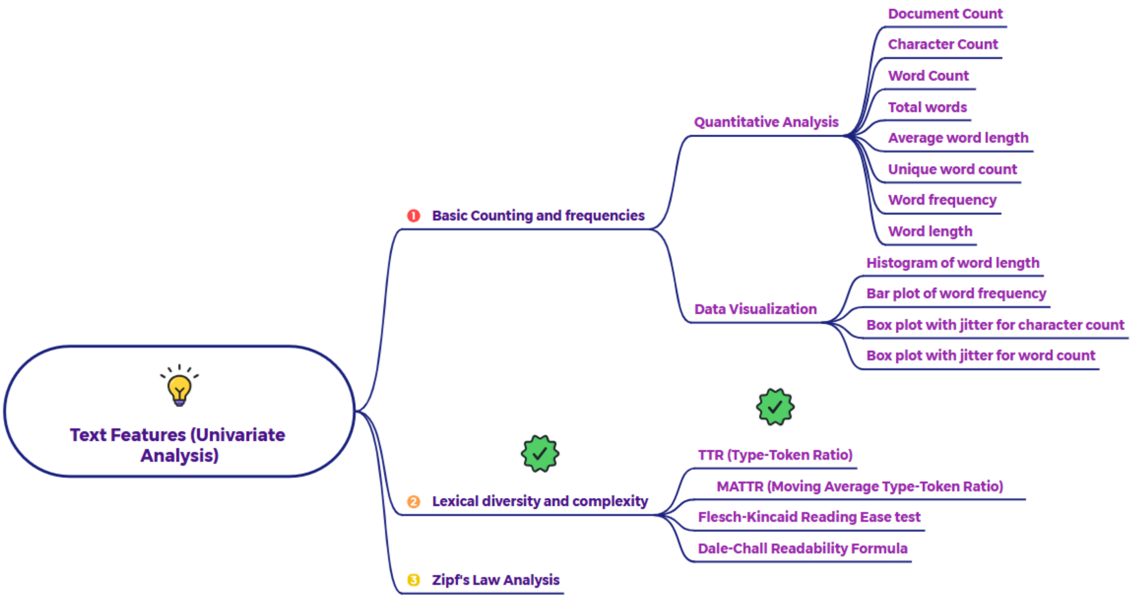
What is TTR (Type token ratio)?


Text Features (Univariate Analysis)
1 Basic Counting and frequencies
Quantitative Analysis
Document Count
Character Count
Word Count
Total words
Average word length
Unique word count
Word frequency
Word length
Data Visualization
Histogram of word length
Bar plot of word frequency
Box plot with jitter for character count
Box plot with jitter for word count
2 Lexical diversity and complexity
TTR (Type-Token Ratio)
MATTR (Moving Average Type-Token Ratio)
Flesch-Kincaid Reading Ease test
Dale-Chall Readability Formula
3 Zipf's Law Analysis

Type-Token Ratio (TTR) is a fundamental and straightforward measure of **lexical diversity** within a text or a collection of texts.

Calculation:

It is calculated by dividing the number of **unique words (types)** in a text by the total number of **words (tokens)** in the same text.

TTR = (Number of Unique Words / Total Number of Words)

How to Interpret Type-Token Ratio (TTR):

The TTR value ranges between 0 and 1 (or can be expressed as a percentage by multiplying by 100).

- **Higher TTR:** A higher TTR value indicates **higher lexical diversity**. This means that a larger proportion of the words in the text are unique, suggesting a richer and more varied vocabulary. The writer is using a wider range of different words and is less repetitive.
- **Lower TTR:** A lower TTR value indicates **lower lexical diversity**. This means that a smaller proportion of the words in the text are unique, suggesting a more limited and potentially repetitive vocabulary. The writer is using the same words more frequently.

Factors Influencing TTR:

It's crucial to understand that TTR is significantly influenced by the **length of the text**:

- **Shorter Texts:** Shorter texts tend to have a higher TTR because every new word encountered is likely to be unique when the total number of words is small. As the text gets longer, the likelihood of encountering previously used words increases, thus lowering the proportion of unique words.
- **Longer Texts:** Longer texts tend to have a lower TTR simply because the total number of words (the denominator) increases while the number of new unique words (the numerator) grows at a decreasing rate as the vocabulary starts to get saturated.

Interpretation Considerations:

- **Direct comparison of TTR values is only meaningful for texts of roughly the same length.** Comparing the TTR of a short story to that of a novel directly can be misleading due to the length effect.
- **TTR provides a basic, surface-level measure of lexical diversity.** It doesn't account for the semantic richness or complexity of the vocabulary used. Two texts might have the same TTR but one could use simpler, more common unique words while the other uses more sophisticated, less frequent unique words.
- **Preprocessing steps can affect TTR.** For example, lemmatization or stemming (reducing words to their base form) will decrease the number of unique words and thus lower the TTR.

In summary, TTR offers a quick way to gauge the variety of vocabulary in a text. A higher TTR generally suggests more diverse word usage. However, always consider the length of the text when interpreting TTR values and be aware of its limitations in capturing the full spectrum of lexical richness and complexity. For more robust measures that address the length effect, techniques like MATTR (Moving Average Type-Token Ratio) are often preferred.