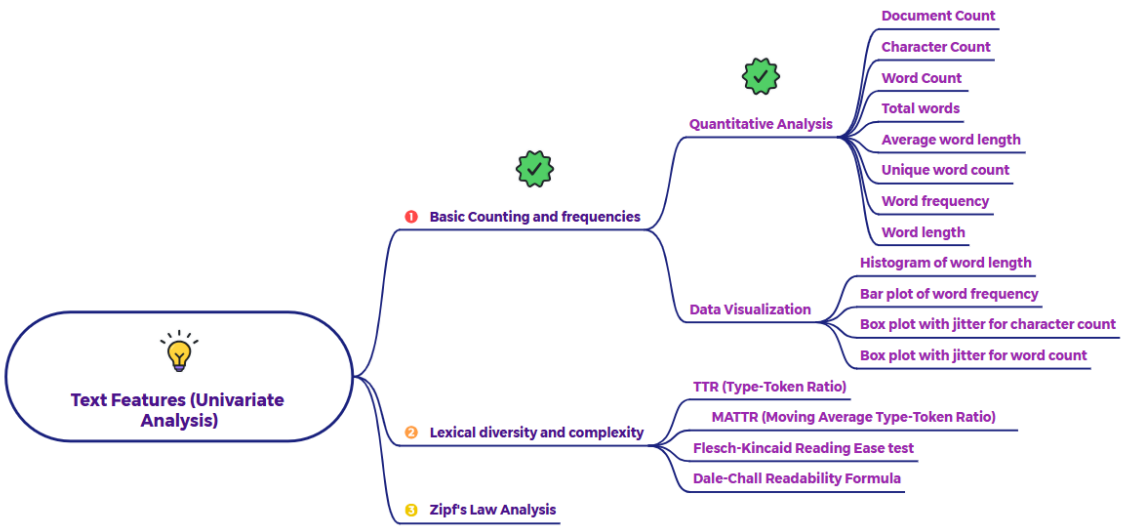
Univariate analysis of text features – Quantitative Analysis



1. Document Count:

- **Explanation:** This refers to the total number of individual text units (documents) in your dataset. A "document" here could be an email, a tweet, a paragraph, a full article, a book, or any other discrete piece of text you are analyzing as a single unit.
- **Univariate Analysis Use:** Knowing the document count provides the overall size of your text corpus. It's a fundamental piece of information for context when interpreting other quantitative measures. For example, a high word count might mean something different in a corpus of 10 documents versus a corpus of 10,000 documents.

2. Character Count:

- **Explanation:** This is the total number of characters in a document (or across all documents in the corpus). This includes letters, numbers, punctuation marks, spaces, and any other symbols.
- **Univariate Analysis Use:**

- **Document Length:** Character count is a basic measure of the length of a document. You can analyze the distribution of character counts across your documents to understand their typical length and identify any outliers (very short or very long documents).
- **Preprocessing Insights:** Extremely high or low character counts might indicate issues with data collection or the need for specific preprocessing steps.

3. Word Count:

- **Explanation:** This is the total number of words in a document (or across all documents). Typically, words are separated by spaces or punctuation. Tokenization rules will influence how words are counted (e.g., whether hyphenated words are counted as one or two).
- **Univariate Analysis Use:**
  - **Document Length (Word-Based):** Similar to character count, word count is a primary way to measure document length. Analyzing the distribution of word counts helps understand the typical length of documents in your corpus.
  - **Comparison:** You can compare the average word count across different subsets of your data (e.g., different authors, different topics).

4. Total Words:

- **Explanation:** In the context of analyzing a collection of documents (a corpus), "Total Words" usually refers to the sum of the word counts of all the documents in the corpus.
- **Univariate Analysis Use:** This provides the overall size of the vocabulary used in your entire text collection. It's a fundamental statistic for understanding the scale of the language data you are working with.

5. Average Word Length:

- **Explanation:** This is calculated by dividing the total number of characters in a document (or corpus) by the total number of words in that document (or corpus). It provides a measure of the average length of the words used.

- **Univariate Analysis Use:**
    - **Text Complexity:** Average word length can be a simple indicator of text complexity. Longer words are sometimes associated with more complex or formal writing. Analyzing the average word length across documents or authors can reveal stylistic differences.
    - **Language Characteristics:** Different languages tend to have different average word lengths.

6. Unique Word Count:

- **Explanation:** This is the number of distinct or unique words present in a document (or across all documents). Variations in word forms (e.g., "run," "running," "ran") might be counted as unique depending on whether stemming or lemmatization has been applied.
- **Univariate Analysis Use:**
    - **Vocabulary Richness:** Unique word count is a basic measure of the vocabulary richness or lexical diversity of a text. A higher unique word count in a document of a given length suggests a more varied vocabulary.
    - **Author Style:** Different authors might have different tendencies in terms of vocabulary richness.

7. Word Frequency:

- **Explanation:** This refers to how often each unique word appears in a document or a corpus. It can be expressed as a raw count (number of times a word appears) or as a relative frequency (the proportion of times a word appears out of the total number of words).
- **Univariate Analysis Use:**
    - **Identifying Important Words:** Analyzing the frequency of individual words can help identify the most common or important terms in a text or a collection of texts. Stop words (common words like "the," "a," "is") are often removed before this analysis.
    - **Topic Identification:** High-frequency content words can provide clues about the main topics discussed in the text.

- **Distribution of Words:** You can analyze the overall distribution of word frequencies (e.g., using a histogram) to understand how common and rare words are in your data.

8. Word Lengths:

- **Explanation:** This refers to the lengths of individual words in a document or corpus (measured by the number of characters in each word).
- **Univariate Analysis Use:**
  - **Distribution of Word Lengths:** You can analyze the distribution of word lengths (e.g., using a histogram) to understand the prevalence of short, medium, and long words in the text. This can provide further insights into text complexity and style.
  - **Comparison:** Comparing the distribution of word lengths across different authors or genres can reveal stylistic differences.

These quantitative measures provide a foundational understanding of the textual data and are often the starting point for more sophisticated text analysis techniques. By examining the distributions and central tendencies of these features, you can gain initial insights into the characteristics of your text corpus.