Vocabulary Size:

Stemmed vocabulary size: 26

Lemmatized vocabulary size: 28

Whitespace vocabulary size: 27

Most Frequent Tokens:

Most frequent stemmed tokens: [('t', 6), ('e', 6), ('/', 5), ('r', 4), ('o', 3), ('d', 3), ('n', 2), ('i', 2), ('v', 2), ('s', 2)]

Most frequent lemmatized tokens: [('t', 6), ('e', 6), ('/', 5), ('r', 4), ('o', 3), ('d', 3), ('i', 3), ('a', 3), ('n', 2), ('v', 2)]

Most frequent whitespace tokens: [('t', 6), ('e', 6), ('/', 5), ('r', 4), ('o', 3), ('i', 3), ('a', 3), ('c', 2), ('n', 2), ('d', 2)]

Average Token Length:

Stemmed average token length: 1.0

Lemmatized average token length: 1.0

Whitespace average token length: 1.0

Vocabulary Size:

Stemming – Achieved the smallest vocabulary size (26), suggesting effective removal of word variations.

Lemmatization – Resulted in a slightly larger vocabulary size (28) compared to stemming, potentially preserving more semantic information.

Whitespace – Generated a vocabulary size (27) close to stemming, indicating minimal impact on unique words.

Most Frequent Tokens:

All schemes share the same top three most frequent tokens: "t", "e", and "/".

Lemmatization and Whitespace share an additional frequent token "i", suggesting potentially better retention of common words.

Stemming seems to favor shorter tokens like "r", "o", "d", while Lemmatization and Whitespace preserve longer forms like "a" and "c".

Average Token Length:

All schemes have the same average token length (1.0), demonstrating that there is no substantial difference in token size among these 3 approaches.

Overall Effectiveness:

* Stemming: Effective for reducing vocabulary size and capturing meaning of words.
* Lemmatization: Slightly larger vocabulary size than stemming, providing a balance between vocabulary size and semantic information retention.
* Whitespace: Minimal impact on unique words but retains original word forms without normalization.