Stemming & Lemmatization => Reducing Vocab Size

Core idea :words with similar meanings should be represented by a single token, even if they have different surface forms.

Word frequency is particularly important

- **Bag of Words (BoW) Model:** As you mentioned, BoW models represent documents as a collection of words and their frequencies. Reducing the vocabulary size can improve the performance of BoW models by reducing noise and improving generalization.
- **Information Retrieval:** In search engines, lemmatization and stemming can help match search queries with relevant documents, even if the query uses a different form of the word.
- **Text Classification:** Reducing vocabulary size can simplify the feature space and improve the accuracy of text classification models.

When to Use which

- **Stemming:** Use stemming when speed is critical and accuracy is less important. It can be useful for tasks where the exact form of the word is not crucial.
- **Lemmatization:** Use lemmatization when accuracy is important and you need to ensure that the resulting tokens are valid words. It's generally preferred for tasks where the meaning of the words is critical.

Goal of Stemming and Lemmatization

- **Improve model performance:** Simpler models are often more robust and generalize better to unseen data.
- **Reduce computational cost:** Smaller vocabularies require less memory and processing power.
- **Improve interpretability:** Simpler models are easier to understand and interpret.

Stemming

The process of reducing words to their root form (stem) by chopping off suffixes.

Characteristics

- **Simplified Morphological Analysis**
  rule-based version of morphological analysis. It doesn't necessarily involve understanding the word's meaning or its grammatical function.

Heuristic Approach

Stemming often relies on a set of pre-defined rules to remove suffixes. It's a more "**guesswork**" approach compared to lemmatization.

Non-Words

A key characteristic of stemming is that the resulting stem may not be a valid word in the dictionary. This is a trade-off for speed and simplicity.

Stemming:

- **Approach:** A more crude, rule-based process that chops off suffixes from words to reduce them to a common stem.
- **Goal:** To reduce words to a common root, even if the root is not a valid word.
- **Algorithm:** Stemming algorithms typically use a set of rules to remove suffixes like "-ing," "-ed," "-s," etc.
- **Speed:** Generally faster than lemmatization.
- **Accuracy:** Can be less accurate, often resulting in stems that are not actual words.
- **Example:**
- "running" -> "run"
- "easily" -> "easi"
- "studies" -> "studi"

Porter Algorithm

- **A Popular Stemming**

Algorithm: The Porter algorithm is one of the most widely used stemming algorithms. It's a rule-based algorithm that applies a series of transformations to words to reduce them to their stems.

Lemmatization:

- **Approach:** A more sophisticated, dictionary-based process that uses morphological analysis to find the lemma (dictionary form) of a word.
- **Goal:** To reduce words to their base or dictionary form, which is a valid word.
- **Algorithm:** Lemmatization algorithms typically use a lexicon (dictionary) and part-of-speech tagging to determine the correct lemma for a word in a given context.
- **Speed:** Generally slower than stemming.
- **Accuracy:** More accurate than stemming, as it produces valid words.
- **Example:**
- "running" -> "run"
- "easily" -> "easy"
- "studies" -> "study"
- "better" -> "good"

Type to enter text

What's Lemmatization

Lemma : can be understood as the 'base form' or 'dictionary form' of a word

Lemmatization : the process of finding the lemma from different forms of a word. It determines if the number of words can be reduced by finding the root word, even if the words have different forms

How to conduct?

1. first conduct morphological parsing of the word.
morpheme = smallest unit with meaning
morphology = the study of creating words from morphemes.

2 types of morphemes: **stem** and **affix**

1. **Stem:** The core part of the word that carries its primary meaning (cats in cat)
2. **Affix:** The part that gives additional meaning to the word. (-s in cat)

morphologial parsing = process of separating words into stem and affix
Morphological parsing is an important step in lemmatization because it helps identify the stem and affixes, allowing the algorithm to determine the correct lemma for a word. The overall goal is to reduce vocabulary size and improve NLP model performance.

Example

Text: "The striped bats are hanging on their feet for best."

- Stemming:
  - the stripe bat are hang on their feet for best
- Lemmatization:
  - the stripe bat be hang on their foot for good

★Lemmatization provides more meaningful and grammatically correct