CHAPTER 1

Stemming and Lemmatization

Stemming and lemmatization are two fundamental techniques in natural language processing that are used to prepare text data. They help in reducing inflectional forms of a word to a common base form.

1.0.1 Stemming

Stemming is the process of reducing a word to its word stem (its basic form). For example, a stemming algorithm would reduce the words "jumping", "jumped", and "jumps" to the stem "jump".

It's important to note that stemming may not always lead to actual words. For example, the stem of the word "running" could be "runn" depending on the stemming algorithm used.

Stemming is generally simpler and faster than lemmatization, but it is also less precise.

1.0.2 Lemmatization

Lemmatization, on the other hand, reduces words to their base or root form, which is linguistically correct. For example, "running" and "runs" are both changed to "run".

Lemmatization uses a more complex approach to achieve this. It considers the morphological analysis of the words and requires detailed dictionaries, which the algorithm can look through to link the form back to its lemma.

1.0.3 Applications

To summarize, both stemming and lemmatization help in text normalization and preprocessing, but while stemming can be faster and simpler, lemmatization is more accurate, as it uses more informed analysis to create groups of words with similar meanings based on the context. These are both used to improve natrual language processing (NLP) in software such as search engines, machine learning models, algorithmic searches, and even baseline programs for autofill and text prediction.

This is a draft chapter from the Kontinua Project. Please see our website (`https://kontinua. org/`) for more details.

Answers to Exercises

INDEX