STEMMING

INTRODUCTION

- Stemming is the process which is reducing a word to its word stem. In stemming affixes are removed from the word.
- Stemming programs are commonly referred to as stemming algorithms or stemmers.
- For example, the stem of the words ***eating, eats, eaten*** is ***eat***.
- Search engines use stemming for indexing the words. That's why rather than storing all forms of a word, a search engine can store only the stems. In this way, stemming reduces the size of the index and increases retrieval accuracy.
- Stemming is an important part of the pipelining process in Natural language processing (NLP).

Algorithm

There are several types of stemming algorithms which differ in respect to performance and accuracy.

Simple stemming algorithm-

In this algorithm we have to create a words table where every inflected form of words are there, and we have to search the words. The advantages of this approach are that it is simple, fast, and easily handles exceptions. The disadvantages are that all inflected forms must be present in listed table: new or unfamiliar words are not handled, even if they are perfectly regular (e.g. cats ~ cat), and the table may be large.

The production technique-

In this technique the words table created automatically. For example, if the word is "eat", then the inverted algorithm might automatically generate the forms "eating", "eats", "eatery", and "eaten".

Suffix-stripping algorithms-

Suffix stripping algorithms do not depend on a word table that consists of inflected forms and root form relations of words. There are some rules are work to provide path to the algorithm.

For examples-

- if the word ends in 'ed', remove the 'ed'
- if the word ends in 'ing', remove the 'ing'
- if the word ends in 'ly', remove the 'ly'

Suffix stripping algorithms are sometimes unprocessed , and given the poor performance when dealing with exceptional relations like 'ran' and 'run'.

Suffix stripping algorithms may differ in results for a variety of reasons. One such reason is whether the output word must be a real word in the given language. Some approaches do not require the word to actually exist in the language lexicon (the set of all words in the language). Alternatively, some suffix stripping approaches maintain a database (a large list) of all known morphological word roots that exist as real words. These approaches check the list for the existence of the term prior to making a decision. Typically, if the term does not exist, alternate action is taken. This alternate action may involve several other criteria.

Errors in Stemming:

There are mainly two errors in stemming –

- over-stemming
- under-stemming

over-stemming:

Over-stemming is when two words with different stems are stemmed to the same root. This is also known as a false positive.

- Universal
- University
- universe

All the above 3 words are stemmed to univers which is wrong behavior. Though these three words are etymologically related, their modern meanings are in widely different domains, so treating them as synonyms in NLP/NLU will likely reduce the relevance of the search results.

Under-Stemming

Under-stemming occurs when two words are stemmed from the same root that are not of different stems. Under-stemming can be interpreted as false-negatives.