Week2

NLP tasks

NLP tasks

Stemming and Lemmatization helps us to achieve the root forms of inflected (derived) words. Stemming is different to Lemmatization in the approach it uses to produce root forms of words and the word produced.

- **Tokenization**: is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens. The tokens could be words, numbers or punctuation marks. In tokenization, smaller units are created by locating word boundaries. Wait – what are
  word boundaries? These are the ending point of a word and the beginning of the next word. These tokens are considered as a first step for stemming and lemmatization.

Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. **Lemmatization** usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma . If confronted with the token saw, stemming might return just s, whereas **lemmatization** would attempt to return either see or saw depending on whether the use of the token was as a verb or a noun. The two may also differ in that stemming most commonly collapses derivationally related words, whereas lemmatization commonly only collapses the different inflectional forms of a lemma.

- **Lemmatazation** (or lemmatization): The process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form. In computational

linguistics, lemmatisation is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatisation depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighbouring sentences or even an entire document.

- **Stemming**: The process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language.
    - Stem (root) is the part of the word to which you add inflectional (changing/deriving) affixes such as (-ed,-ize, -s,-de,mis). So stemming a word or sentence may result in words that are not actual words. Stems are created by removing the suffixes or prefixes used with a word.
- **Part of Speech (POS) tagging**: Provides insights into the word classes / types of sentences (verb, noun, adverb, preposition, etc.). There is ambiguity in some cases, in which the system picks the most probable ("Visiting aunts can be a nuisance").
- **Parsing sentence structure**: Creates a context-free grammar tree to denote the different parts of the sentence (direct object, subject, etc). There can also be ambiguity.