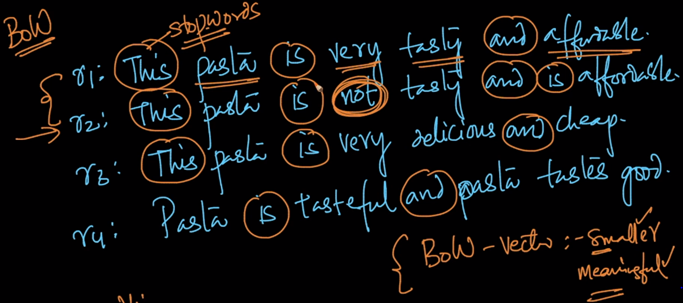
**Text Preprocessing:**

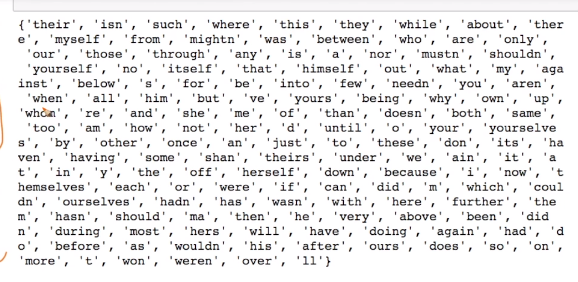
1. **Stopwords:**

As we can seen in below image some words like “this, is, and” will not help much in categorizing statement, and this are known as stop words, so we remove stop words.

Note: It’s not always recommended to remove stop words, because few stopword like ‘not’, which is very useful in current case.



Stopwords present in English language.



1. **Make all words lower case:**

As pasta and Pasta are same but just because of P and p, they are treated as different therefore we convert all words to lower case.

1. **Stemming:**

Stemming is basically replacing each word with their root word, as in our example

Tasty, tastes, tasteful have same meaning, therefore it’s better to replace them with their root word tast.

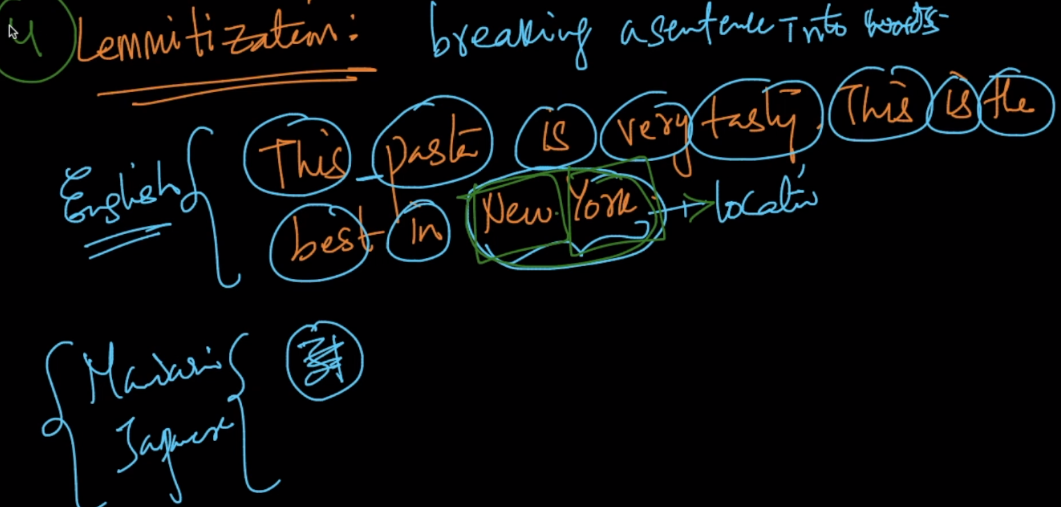


1. **Lemmatization:**

in linguistics it is 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**

**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 neighboring sentences or even an entire document.

Basically it’s the way of breaking a sentence into words, such that preserving the meaning of broken word, example in below image New York is a city and we can’t break them as New and York, that means preserving the semantic meaning.



Example:

from nltk.stem import PorterStemmer, WordNetLemmatizer

stemmer = PorterStemmer()

lemmatiser = WordNetLemmatizer()

print("Stem %s: %s" % ("studying", stemmer.stem("studying")))

print("Lemmatise %s: %s" % ("studying", lemmatiser.lemmatize("studying")))

print("Lemmatise %s: %s" % ("studying", lemmatiser.lemmatize("studying", pos="v")))