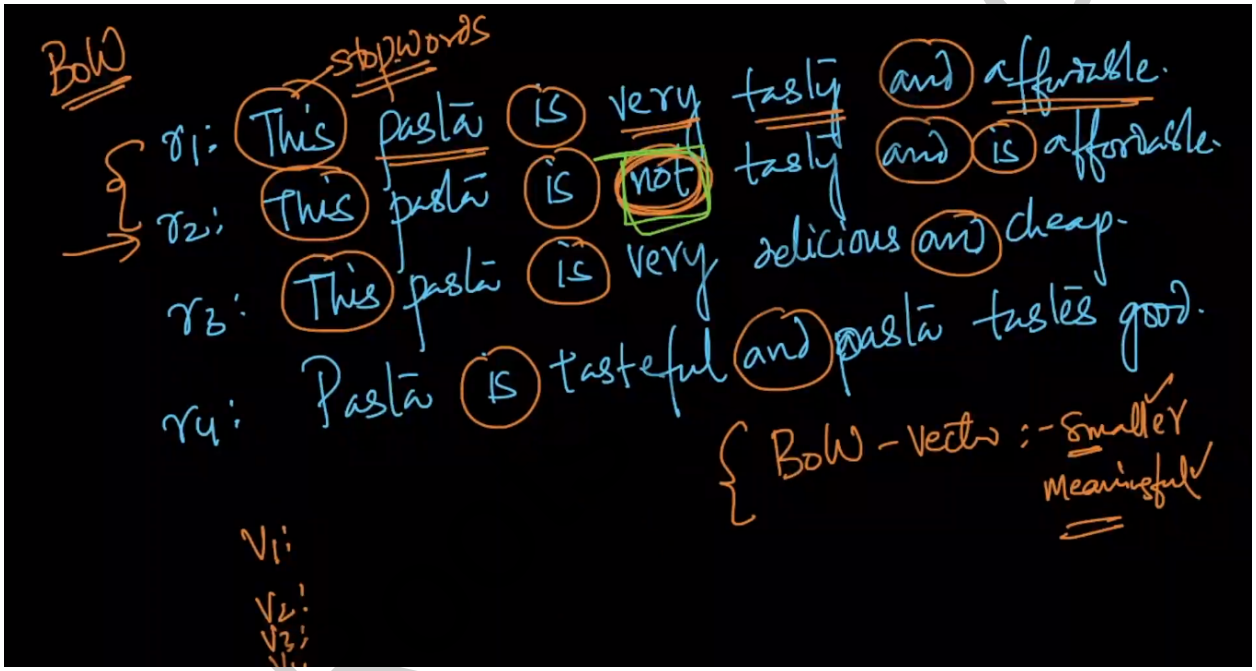
28.5 Text Preprocessing: Stemming, Stop-word Removal, Tokenization, Lemmatization

In a document, there will be certain words which make more sense about the data and there will be certain words which do not make much sense about the data. These words that do not make much sense about the data are present in the document just for sentence completion. Such words are called **Stopwords**.

These stopwords can be removed in order to make the bag of words vector smaller and more meaningful. Removal of stopwords is one of the text preprocessing steps and it has to be performed only when needed.



Steps in Text Preprocessing

1) Stopword Removal
2) Conversion of all words into lowercase
3) Stemming

In Stemming, the words like ‘taste’, ‘tasteful’, ‘tastes’ are reduced to the root form ‘tast’. There are many stemming algorithms in Natural Language Processing (NLP). Two of the majorly used Stemming algorithms are PorterStemmer and SnowballStemmer. SnowballStemmer is much more powerful than PorterStemmer.

4) Lemmatization

Lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatization 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.

Note:

Let us look at the reviews 'r1' and 'r3' given below.

$\mathbf{r_1}$: this pasta is very tasty and affordable

$\mathbf{r_3}$: this pasta is delicious and cheap

These two reviews give the same meaning, but in BOW vectorization, the words 'tasty' and 'delicious' are treated as two different features. Similarly the words 'affordable' and 'cheap' are treated as two different features. This is the main disadvantage with the Bag of Words approach, as it doesn't preserve the semantic meaning.

The semantic meanings of the words are taken into consideration in the **Word2Vec vectorization techniques**. So finally using **Text Preprocessing + Bag of Words**, we are converting the text into a 'd' dimensional vector that could not guarantee the semantic meanings of the words. Bag of words doesn't take the semantic meanings into consideration.

References:

Refer to the below blogs to learn about the differences between Stemming and Lemmatization.

https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/

https://towardsdatascience.com/stemming-vs-lemmatization-2daddabcb221

Tokenization

Tokenization is the process of splitting a given string into a sequence of sub-strings. There are two types of Tokenization. They are **Word Tokenizer** and **Sentence Tokenizer**.

The Word Tokenizer splits the given sentence into a sequence of words, on the basis of space whereas the Sentence Tokenizer splits the given sentence into a sequence of words/sentences on the basis of dot(.).

Example: "Hello Mr.Rajeev, How are you doing today?

Word Tokenizer Output:

["Hello", "Mr.", Rajeev", "How", "are", "you", "doing", "today"]

Sentence Tokenizer Output:

["Hello Mr", "Rajeev How are you doing today?"]