Text preprocessing techniques in NLP

1.Convert string to lower or upper case

2.Remove Special Character

3.Stop word removal

4.Stemming

5.Lemmatisation

6.Part-of-Speech Tagging

Convert string to lower or upper case

Convert all the word to lower or upper case. It is done because when we vectorize the data we will have 2 different dimension for same word. For example 'UPPER' and 'upper' will have different dimension. If we convert to lower we will have 1 dimension for every word.

Python code to convert all words to lower case

Remove Special Character

The raw text we have will have lot of noise like punctuations, Special Character & extra White Space. In some cases it does not add meaning to the text/sentence. So those can be removed using popular python library called regex(*re*) or string function

Removes all character other than A-Z,a-z,0–9 using re

They have there own advantage and disadvantage. We can use re to expanding abbreviations, remove white space, remove numbers, replace 1000000 to 1million etc.

Removes [!”#$%&’()*+,-./:;<=>?@[\]^_`{|}~]: using string function

Stop word removal

Text may contain stop words like ‘the’, ‘is’, ‘are’. Stop words can be filtered from the text to be processed. There is no universal list of stop words in nlp research, however the nltk module contains a list of stop words.

Stop word removal depends on the task. For example, if you have task of text classification or sentiment analysis then you should remove stop words since they don’t provide any information to model but if you have task of language translation then stopwords are useful.

we can see in, this, we, are, to are the words which got removed

Stemming

Stemming is 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. Stemming usually trims the word using set of rules for example plays, playing and played is trimmed to play by removing suffix ‘s’, ‘ing’ and ‘ed’.

There are mainly two errors in stemming –over-stemming and under-stemming

Over-stemming occurs when two words are stemmed from the same root that are of different stems. Over-stemming can also be regarded as false-positives.

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. Read more about different types of stemming here.

If u see machine has been stemmed to machin. There is no word called machin in english

Lemmatisation

The difference between stemming and lemmatization is, lemmatization considers the context and converts the word to its meaningful base form, whereas stemming just removes the last few characters, often leading to incorrect meanings and spelling errors.

For example, lemmatization would correctly identify the base form of ‘caring’ to ‘care’, whereas, stemming would cutoff the ‘ing’ part and convert it to car. Read more about different Lemmatisation here.

Notice it didn’t do a good job. Because most of the words remain same. This can be corrected if we provide the correct ‘part-of-speech’ tag (POS tag) as the second argument to lemmatize().

This looks perfect. As we can see the words are brought down to proper meaningful root word.

POS Tagging

POS tagging is a process of labelling each word in a sentence with its appropriate part of speech. We already know that parts of speech include nouns, verb, adverbs, adjectives, pronouns, conjunction and their sub-categories.

Most of the POS tagging falls under Rule Base POS tagging, Stochastic POS tagging and Transformation based tagging.