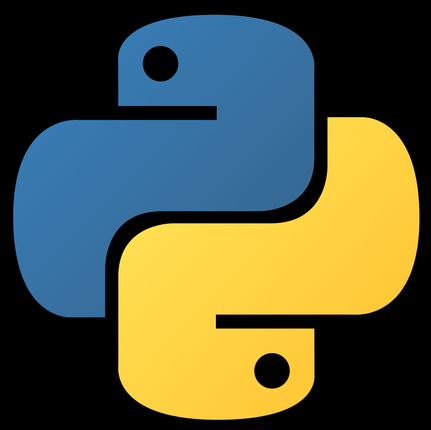
Removing

STOP WORDS

In NLP using NLTK, GENSIM and SPACY



NLTK

CODE

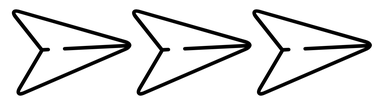
STOP WORDS???

- A stop word is a commonly used word and which have very little meaning (such as "the", "a", "an", "in", etc....)
- Search engines and other enterprise indexing platforms often filter the stop words, both when indexing entries for searching and when retrieving them as the result of a search query.

- Stop words are often removed from the text before training deep learning and machine learning models since stop words occur in abundance, hence providing little to no unique information that can be used for classification or clustering.

STOP WORDS???

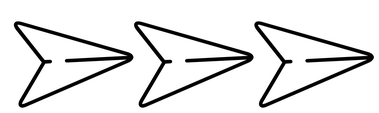
- Sentences with stop words
  - I am part of learn machine learning family
  - I am a data scientist
- Sentences without stop words.
  - part learn machine learning family
  - data scientist



WHY REMOVING STOP WORDS

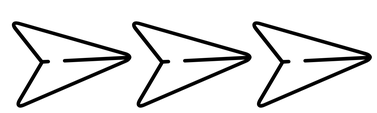
- If we have a task of text classification or sentiment analysis then we should remove stop words as they do not provide any information to our model.
- But if we have the task of language translation, text summarizationthen stopwords are useful, as they have to be translated along with other words.

- It also reduces the dataset size and training time
- It can increase classification accuracy
- Even search engines like Google remove stopwords for fast and relevant retrieval of data from the database



WHEN TO REMOVE??

- No hard rule.
- It depends on the task that you are working on.
- We can remove stopwords while performing the tasks like Text Classification, Spam Filtering, Language Classification, Genre Classification, Caption Generation, Auto-Tag Generation, etc....
- Avoid Stopword Removal for tasks like Machine Translation, Language Modeling, Text Summarization, Question-Answering problems, etc....



REMOVE USING NLTK

- Import NLTK and stop words.

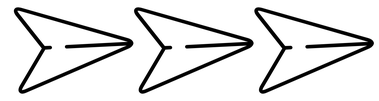

```
>>> import nltk
>>> nltk.download('stopwords')
>>> from nltk.corpus import stopwords
```

- Print the list of available stop words

```
>>> print(stopwords.words('english'))
>>> # it supports 16 different languages
```

- Let's remove stop words from a sentence

```
>>> from nltk.tokenize import word_tokenize
>>> example_sent = "This is a sample sentence, showing
off the stop words filtration."
>>> tokens = word_tokenize(example_sent)
>>> tokens_without_sw = [word for word in tokens if
not word in stopwords.words()]
```



REMOVE USING GENSIM

- Import GENSIM

```
>>> from gensim.parsing.preprocessing import
remove_stopwords
```

- Remove stop words


```
>>> example_sent = "This is a sample sentence, showing
off the stop words filtration."
>>> test_witout_sw = remove_stopwords(example_sent)
```

REMOVE USING SPACY

- Install spacy

```
pip install -U spacy
python -m spacy download en_core_web_sm
```

- Remove stop words


```
>>> import spacy
>>> nlp = spacy.load('en_core_web_sm')
>>> example_sent = "This is a sample sentence, showing
off the stop words filtration."
>>> stopwords = nlp.Defaults.stop_words
>>> tokens = word_tokenize(example_sent)
>>> tokens_without_sw= [word for word in tokens if
not word in stopwords]
```