* Tokenization in NLP

① Topics

① Corpus → Paragraphs

② Docoments → Sentences

③ vocabulary → unique words

④ words → All words present in corpus.

② Tokenization :-

" My name is prathamesh I have intrest in teaching in ML, NLP & DL." I am also a Youtuber".

⇓

Tokens { sentences }

⇓

1) my name is prathamesh & I have intreset in teaching ML, NLP & DL.

2) I am also a Youtuber.

⇓

Tokenisation

③ Another example :- "I like to drink Apple Juice my friends like mango Jucie."

↓ Tokenization.

① ~~Se~~ Corpus : " I like to drink Apple Jcie. & my likes mango Juice."

② Documents :-

① I like to drink Apple Juice.

② my friend like mango Juice.

③ Totel words :- 11 words.

④ vocaboloury := 9 words unique words.

4) Practicle session NLP

→ Diffrence Between NLTK & Spacy.

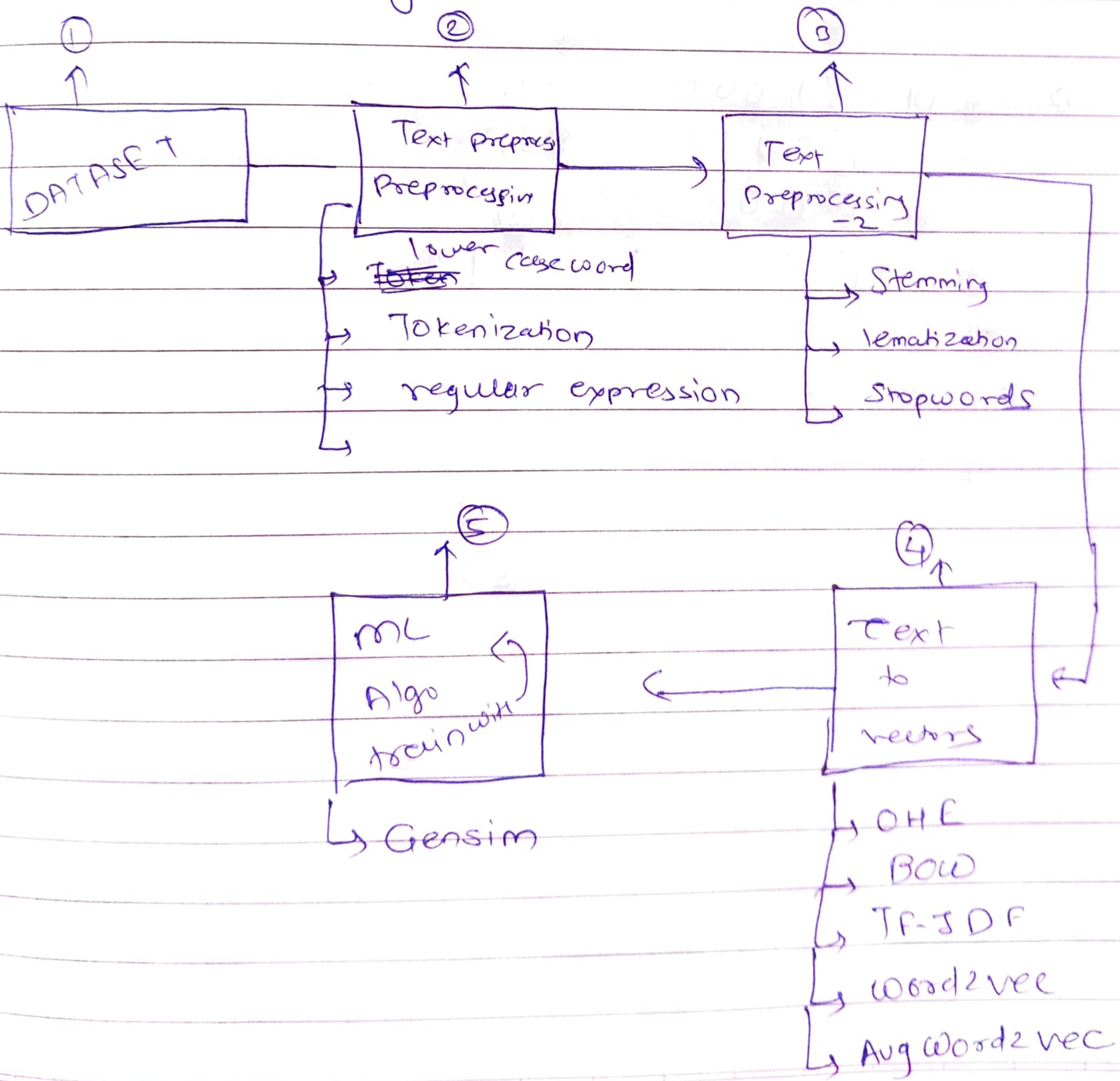
Paramaters	Spacy	NLTK.
Use	High Performane & productio ready application focousing speed & efficiency.	NLTK is more towords research & eduction, offer wider range of algorithum & flexibility.
Performance	High	low
Approch	Object-orient and have rich information ~~about~~ about text	Primary work on strings require manual approch or work.
focous	in production high accuracy & Performance	research & experiments algorithms.

(5) Practice session

→ Tokenization.

10. What we learnt ?

Text preprocessing


1
2
3
DATASET
Text prepros
Preprocessin
Text
Preprocessing
-2
lower ~~Token~~ case word
Tokenization
regular expression
Stemming
lematization
Stopwords
5
4
ML
Algo
train with
Gensim
Text
to
vectors
OHE
BOW
TF-IDF
Word2vec
Avg Word2vec