NLP

Tuesday, July 08, 2025 1:15 PM

Natural Language Processing :

RNN

Transformer

LLM → ChatGPT

CNN → Images → Computer vision

Google → IR

Text → Numbers

Grammer

NLP : Translator

news Summarization

spam

AutoCorrect

Text : I like the movie ↑ → P

I like the movie but acting was subpar → N

Pipeline Text → num.

1. Get the Entire Data/corpus ✓
2. Tokenization ✓
3. Stopword removal ✓
4. Stemming/Lemmatization ✓
5. Building a vocab ✓
6. Vectorization

nltk

Sklearn

1. Text label
2. Tokenization :

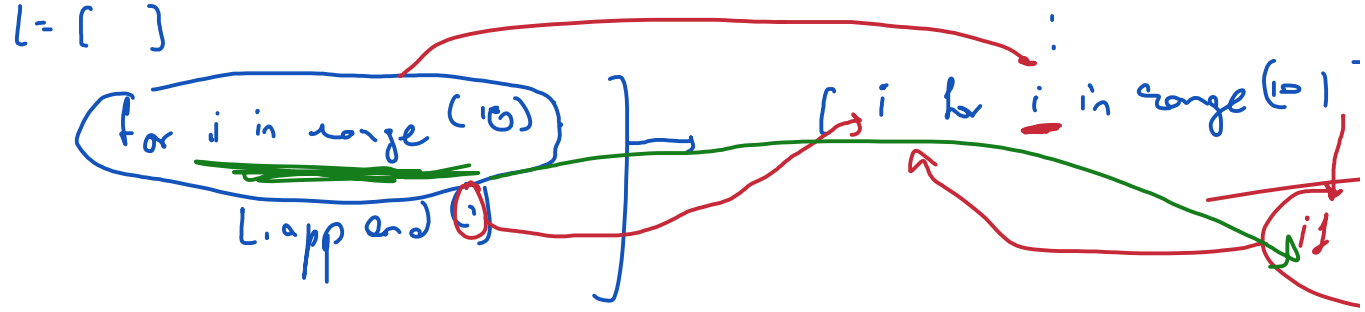
My name is Shubham. I have an interest in teaching ML.

1. Sent tokenization
2. word tokenization

1. My name is Shubham
   I have -- - - -- - -
2. My I

name is Shubham have an interest

Stopwords : too — I like this movie



- Stemming

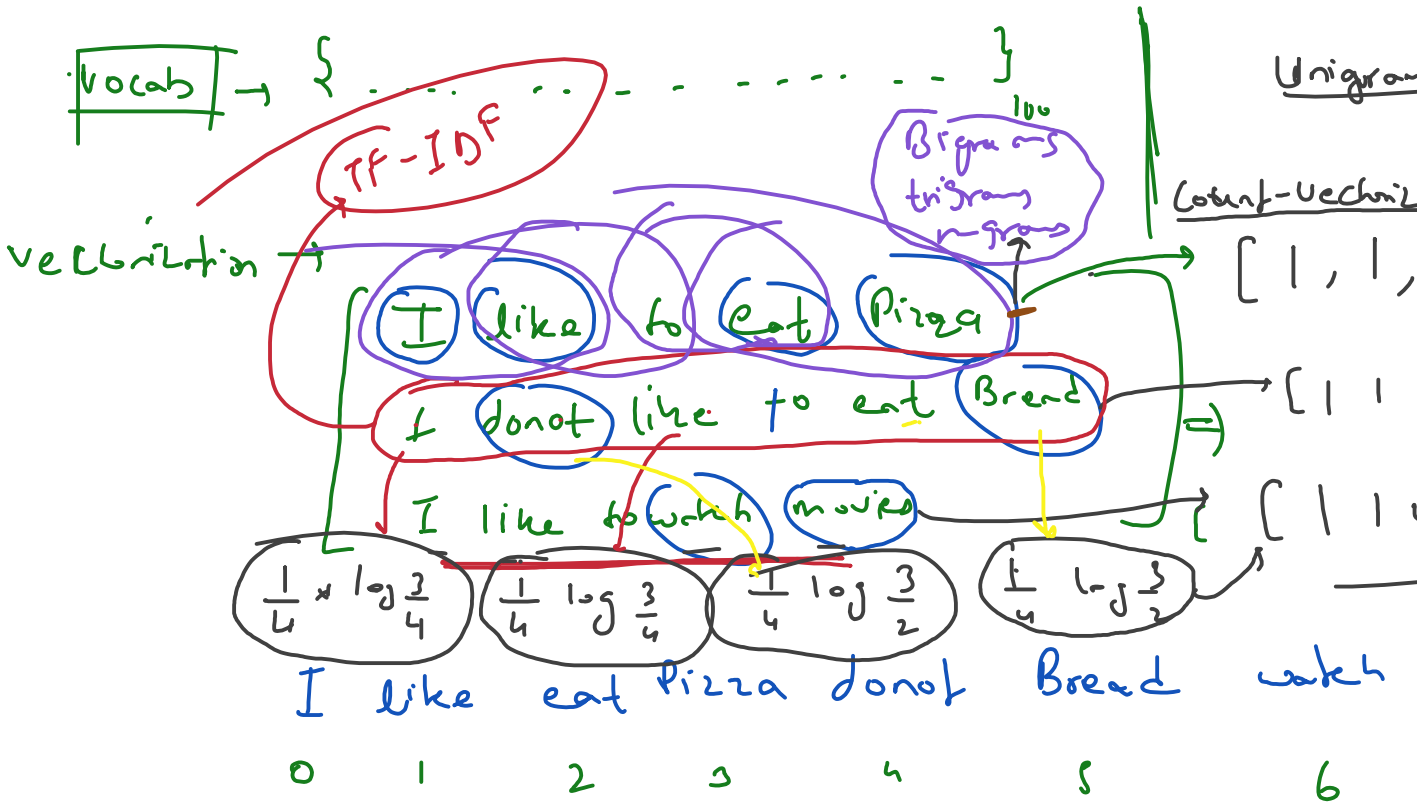
run
running
→ run

give
gave

swim
swimming
→ swim

Lemmati

Carl



I like to eat pizza and I like to watch

[2, 2, 1, 1, 0, 0, 1, 1]

1000 → 800 like ↓ TF-IDF

TF * Term Frequenc

tf(t,d) → No. of times t occurs / Total no. of terms/word

*

IDF(t,c) → log Total number of / 1+ number of Containing t

TF-IDF

New York Bigrams

Word2Vec

RNN, Transformers

Word → Vector

the → [. . . . .] 300

king → [ ] 50

Trying to capture meaning

Male	0.8
Female	0.1
Food	0.2
Kingdom	0.9
City	0.5
Country	0.7

Apple → fruit, Red, taste, mobile, React

1 Dim

2

CBow

Skip gram

Google Word2Vec

Wikipedia

[0 0 1 0 0 0] hidden layer of size d

vocab size

vocab size

The girl is dancing

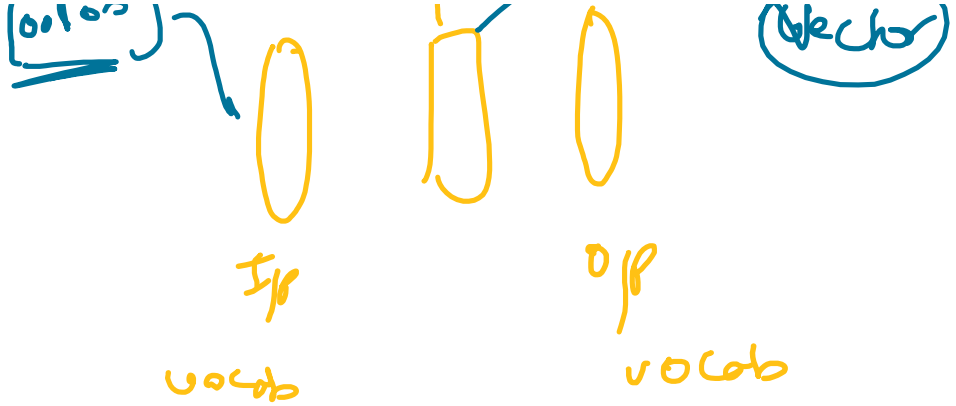
girl

dancing

I/P

[0 0 1 0 0 1 0 0 0 .

dim


Ip
vocab
Op
vocab