Day 1 - Natural Language Processing (ML & Deep Learning)

Agenda

1. Roadmap of Natural Language Processing ✓
2. Why NLP ✓
3. Lot of Examples ✓
4. Tokenization, Stemming, Lemmatization ✓
5. Bag of Words

Prerequisites

1. Python ⇒ community Sessions
2. Stats
3. Machine Learning Algo
4. ANN, Optimizers, Loss functions

Quiz : 5000 Rs

1. 2000Rs INR
2. 1500 INR
3. 1500 INR

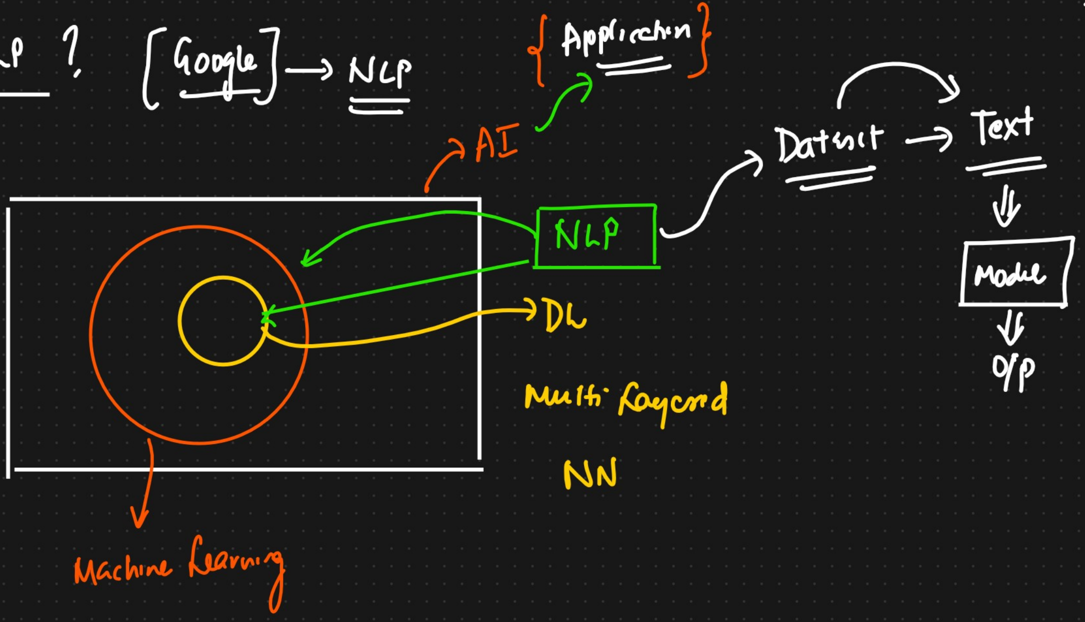
{KrishnaiK06}

SPAM CLASSIFICATION

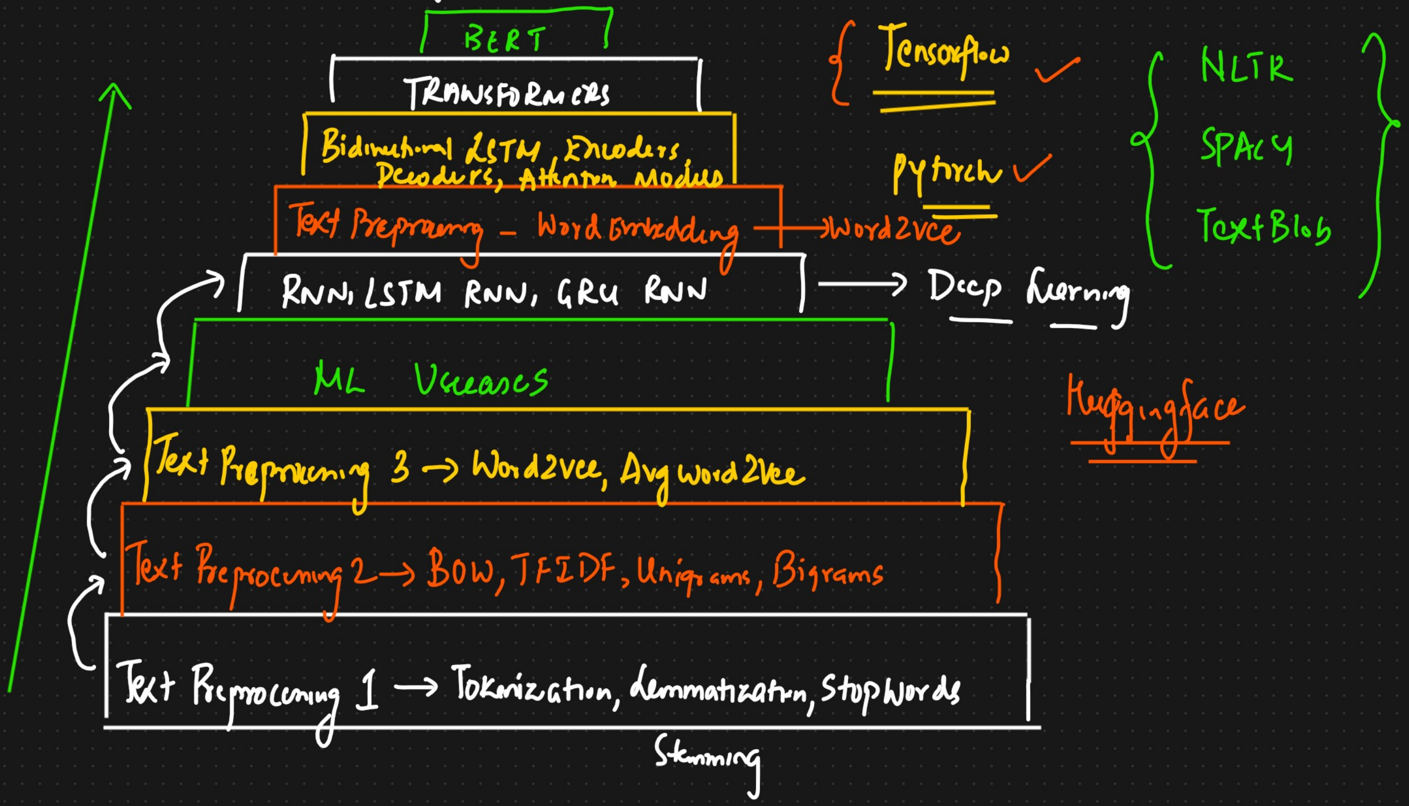
{Machine Translation}

Why NLP ?

[Google] → NLP



Roadmap of NLP



BERT

TRANSFORMERS

Bidirectional LSTM, Encoders, Decoders, Attention Models

Text Preprocessing – Word Embedding → Word2vce

RNN, LSTM RNN, GRU RNN → Deep Learning

ML Usecases

Text Preprocessing 3 → Word2vec, Avg word2vec

Text Preprocessing 2 → BOW, TFIDF, Unigrams, Bigrams

Text Preprocessing 1 → Tokenization, Lemmatization, StopWords, Stemming

{ Tensorflow ✓

Pytorch ✓

Huggingface

{ NLTK

SPACY

TextBlob }

NLP

① Tokenization

ML Usecase

Mail Spam Classifier

I/p features: Email body, Email Subject

Spam/ham

{ ① Tokenization → ② Stemming → ③ Lemmatization
Tokenization ↓ Stopwords

Dataset

	f1 Email body	f2 Email Subject	O/P
1)	You won 1000000 $$	Billionaire	Spam
2)	Hey KRISH, How ARE YOU	Hello	HAM
③	Credit Cards worth	Winner	Spam

You | won | 1000000 | $$

Text Preprocessing

① Tokenization :{ Sentence into words} Sentence → Converts

[Hey buddy I want ~~to~~ go ~~to~~ your ~~of~~ house] → not ←

⇓

stopwords → yes

[.]

③ **Stemming** { Not have any meaning }

Processing of reducing words to their Base word Stem

⇓

Root word

or

Base form

historical
history } histori ⇑ Not have any meaning

finally
final
finalized } ⇒ fina

⇑ Meaning is gone

going
goes
gone } ⇒ go

⇑ { Meaningful word }

Advantages

① Stemming is really fast

Disadvantage

① It is removing the meaning of the word

④ **Lemmatization**

history
historical } history

finally
final
finalized } final

Advantages

① Meaningful world

Disadvantage

① It is slow.

Usecase

Stemming

① Spam classification

② Review classification

Lemmatization

① Text Summarization

② Language Translation

③ Chatbot

Step 1

Text Preprocessing

① Tokenization ② Stopwords ③ Stemming ④ Lemmatization

Step 2 : Words → Vectors

① Bag of Words ② TF-IDF ③ Word2vec

⇓

Term Frequency – Inverse Document Frequency.