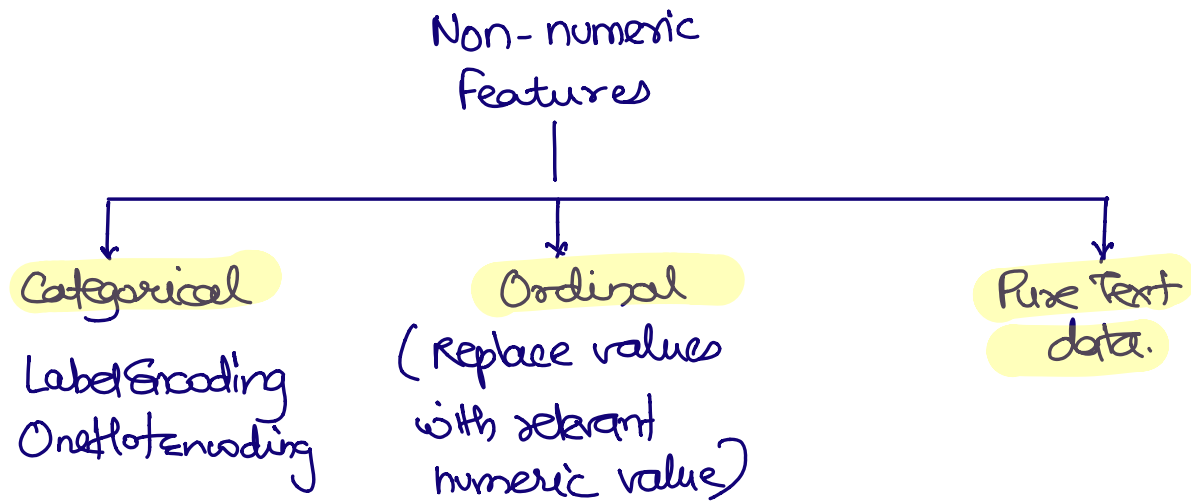
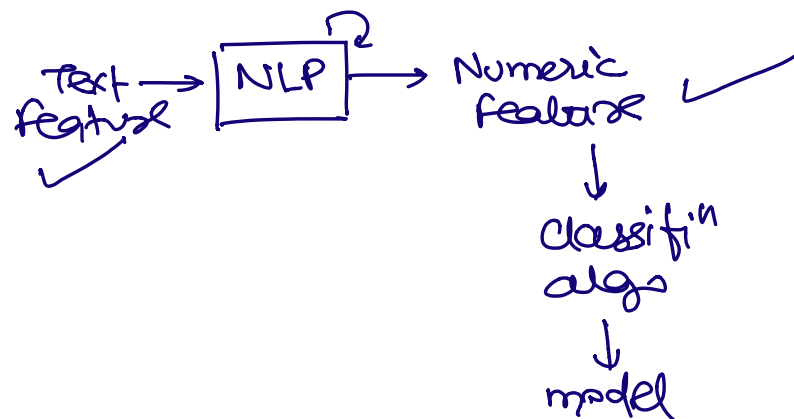


Agenda.

NLP - Natural Language Processing



The scope of NLP is to handle pure text data.



NLP program rules

- ① No punctuations
- ② No stopwords

Python Basics

Nltk package

Text
Feature

Reprocessing

- ① Remove punctuations
- ② Convert sentences into words
- ③ Remove stopwords.

Sms Spam
Classifier

Create BagOfWords

create a vocab. dictionary which
contains text and its count
eg. (Hello, 27223)

convert count into frequency both
(0-1)

TF-IDF

Term Frequency - Inverse Document
Frequency (Hello, 0.8)

Create feature array.

Label

ML algo
Naive
Bayes

model

"win lottery guaranteed" → model → spam

eg.

dataset = ["welcome to simplilearn",
"welcome welcome welcome",
"simplilearn, simplilearn, welcome"]

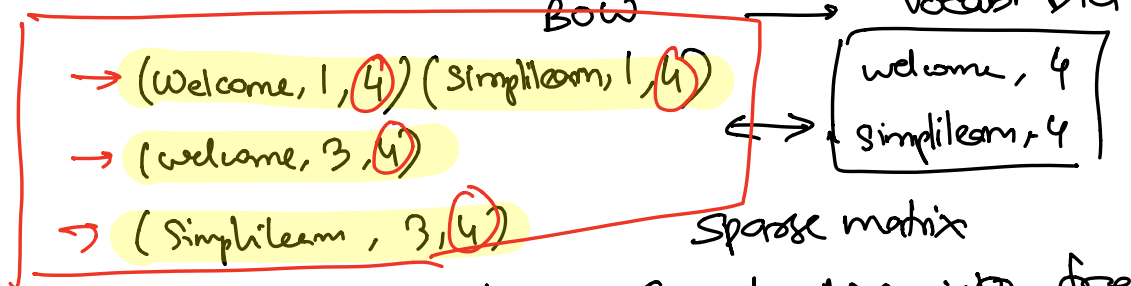
↓

process

(welcome, simplilearn)
(welcome, welcome, welcome)
(simplilearn, simplilearn, simplilearn)

↓

BOW



convert bow into freq using TFIDF

↓

TFIDF-Transformer

fit(bow) →

{ (welcome, 1/4) (simplilearn, 1/4)
(welcome, 3/4)
(simplilearn, 3/4) sparse matrix

←

feature

0.25, 0.25

0.75

0.75

sparse matrix