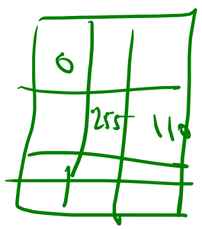
→ Class start at 9:05 pm

→ NLP (Natural language Processing)

↳ Extract information from text and take actions on it

→ Convert text into numerical value and convey all the meaning in sentence

→ Understand the meaning of text



Eg: 1) The food is very good

2) Completely lacking in good taste good service

Sentiment
+ve
-ve

Applications of NLP

→ Translation :

→ Chatbot

→ Generate a summarized review

→ Sentiment analysis

→ Q & A

Syllabus

→ Intro to NLP } Preprocessing,
→ Text Representation } Encoding, BOW, TF-IDF

→ Word Embedding → CBOW, SKipGram

→ Language Modelling

→ RNN, LSTM, GRU

→ NER

→ Transformer, Attention, BERT

Problem

→ Given the corona - tweet,

Classify the sentiment.

Preprocessing technique

→ Removal of special characters

↳ https://___.com

Sentiment

→ Removal of stopword

↳ Commonly used word in any language

↳ a, an, the, and

Corpus : A large & structured collection of text document. Can be book, article, collection of social media post.

Document : Collection of large document is called corpus

Vocabulary

↳ A set of unique words in corpus

[1. The movie is amazing. I watched it today
2. I watched the movie Spiderman]

Vocabulary = {the, movie, is, amazing, I, watched, it, Spiderman}

Tokenization

↳ Breaking text into smaller units called token

↳ Word → I, the, movie, is, amazing
I, watched, it, today

↳ Character

↳ Subword

→ Break until 10:10pm

Stemming & lemmatization

→ Helps in converting words into their base form

Eg: have, having, had → have

run, running, ran → run

Stemming

histori
historical

→ May not be understood by human

→ Works by cutting begin or end of word

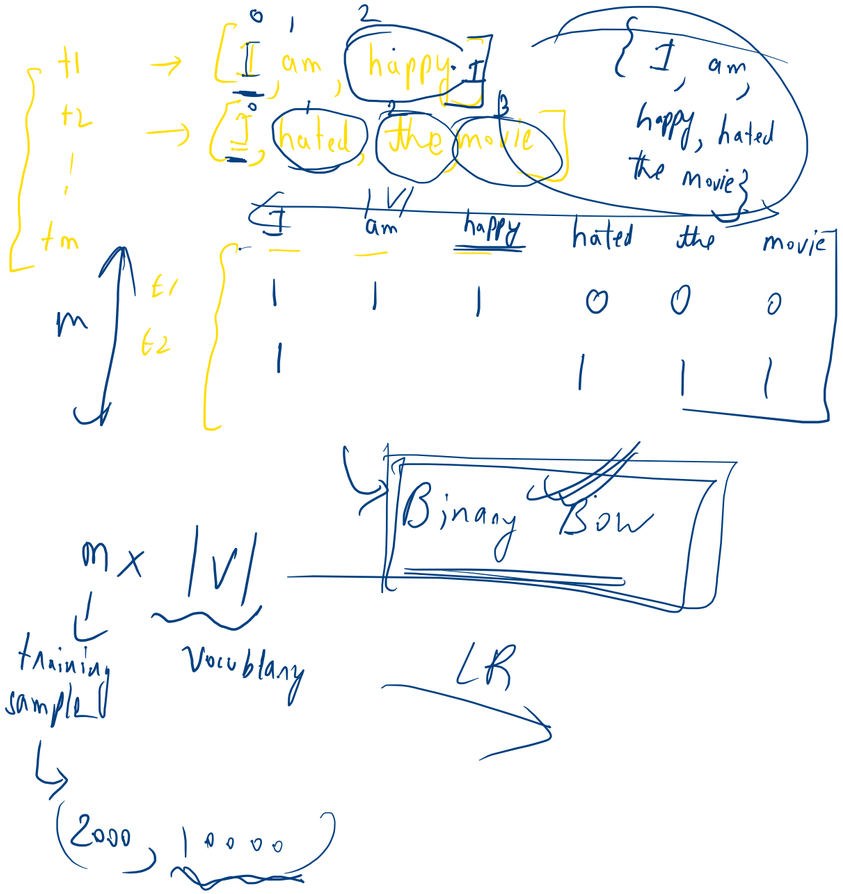
Lemmatization

history
historical → history

→ Understood by human

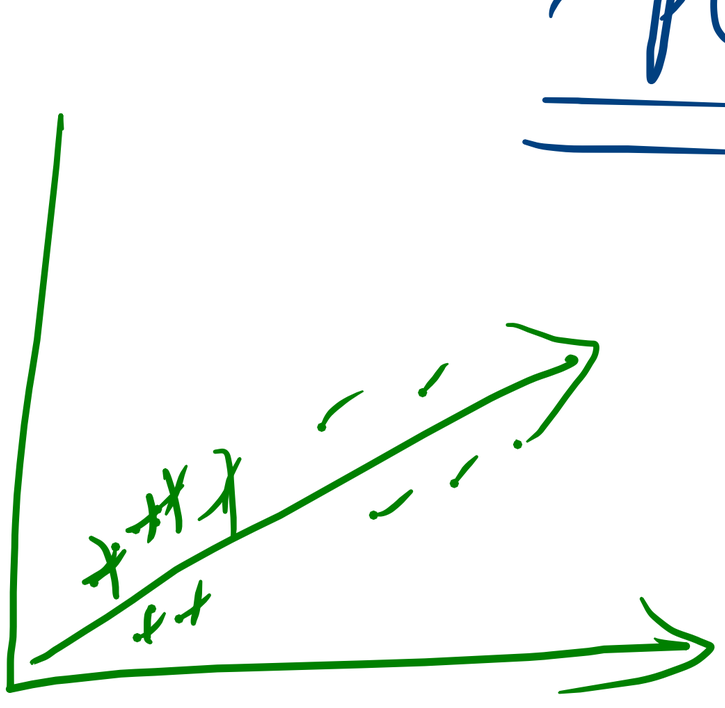
→ Considers the context and convert into more meaningfull

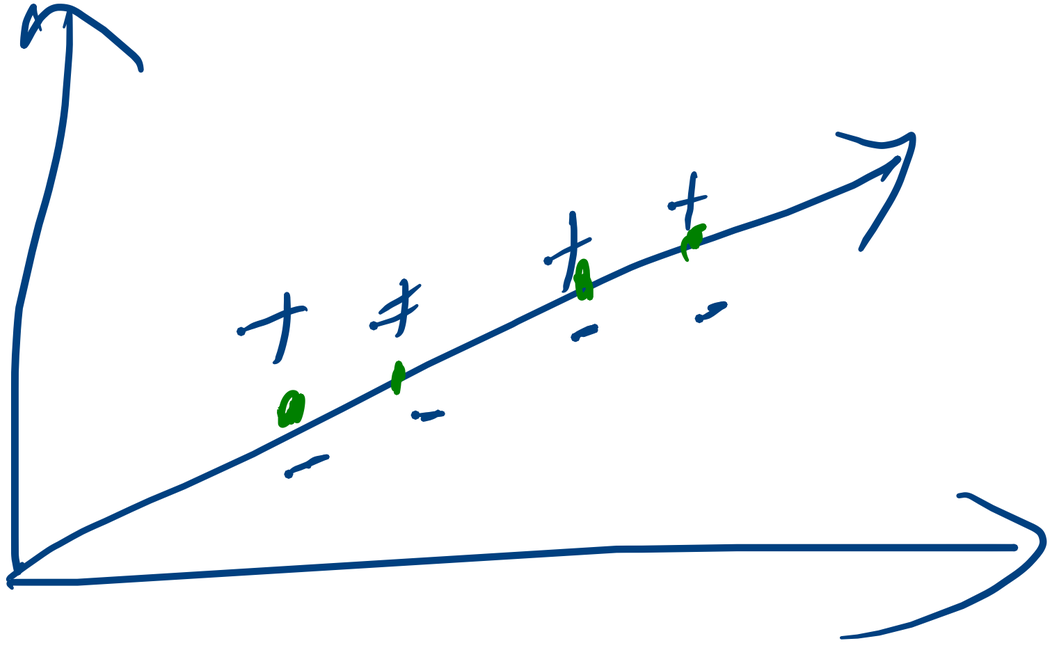
1. Preprocess
2. Tokenize
3. Stem & lemm

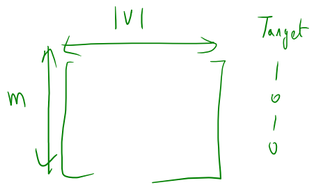

t1
t2
tm
I am happy. I
I hated the movie
{I, am, happy, hated the movie}
I am happy hated the movie
m
E1
E2
1 1 1 0 0 0
1 1 1 1
Binary BOW
m x |V|
training samples
Vocublary
LR
(2000, 10000)
Very sparse
LR will have (|V|+1) parameters
|V|
m
BOW
LR
Logistic

→ How can we reduce dimension?

↳ PCA







freqs: dictionary mapping from (word, class) to frequency

$$X_m = [1, \sum_w freqs(w,1), \sum_w freqs(w,0)]$$

Features of tweet m | Bias | Sum Pos. Frequencies | Sum Neg. Frequencies

2 features

LR

1. I am happy because I am learning NLP — Target 1
2. I am happy — 1
3. I am sad, I am not learning NLP — 0
4. I am sad — 0

Vocab = { I, am, happy, because, learning, NLP, sad, not }

(I, 1) = 3 (am, 1) = 3
(I, 0) = 3 (am, 0) = 3

Vocab	1	0
I	3	3
am	3	3
happy	2	0
because	1	0
learn	1	1
NLP	1	1
sad	0	2
not	0	1

Surprise

I am sad, I am not learning NLP

→ (1, 2)

$$X_m = \{ \sum_w freq(w,1) \quad \sum_w freq(w,0) \}$$

$$\sum_w freq(w,1) = freq(I,1) + freq(am,1) + freq(sad,1) + freq(not,1) + freq(learn,1) + freq(NLP,1)$$

$$= 3 + 3 + 0 + 0 + 1 + 1 = 8$$

$\sum_{\omega} freq(\omega, 0)$

11

$m \times |V|$

$\sum freq(w,1)$

$m \times 2$

$\sum freq(w,0)$

LR

m