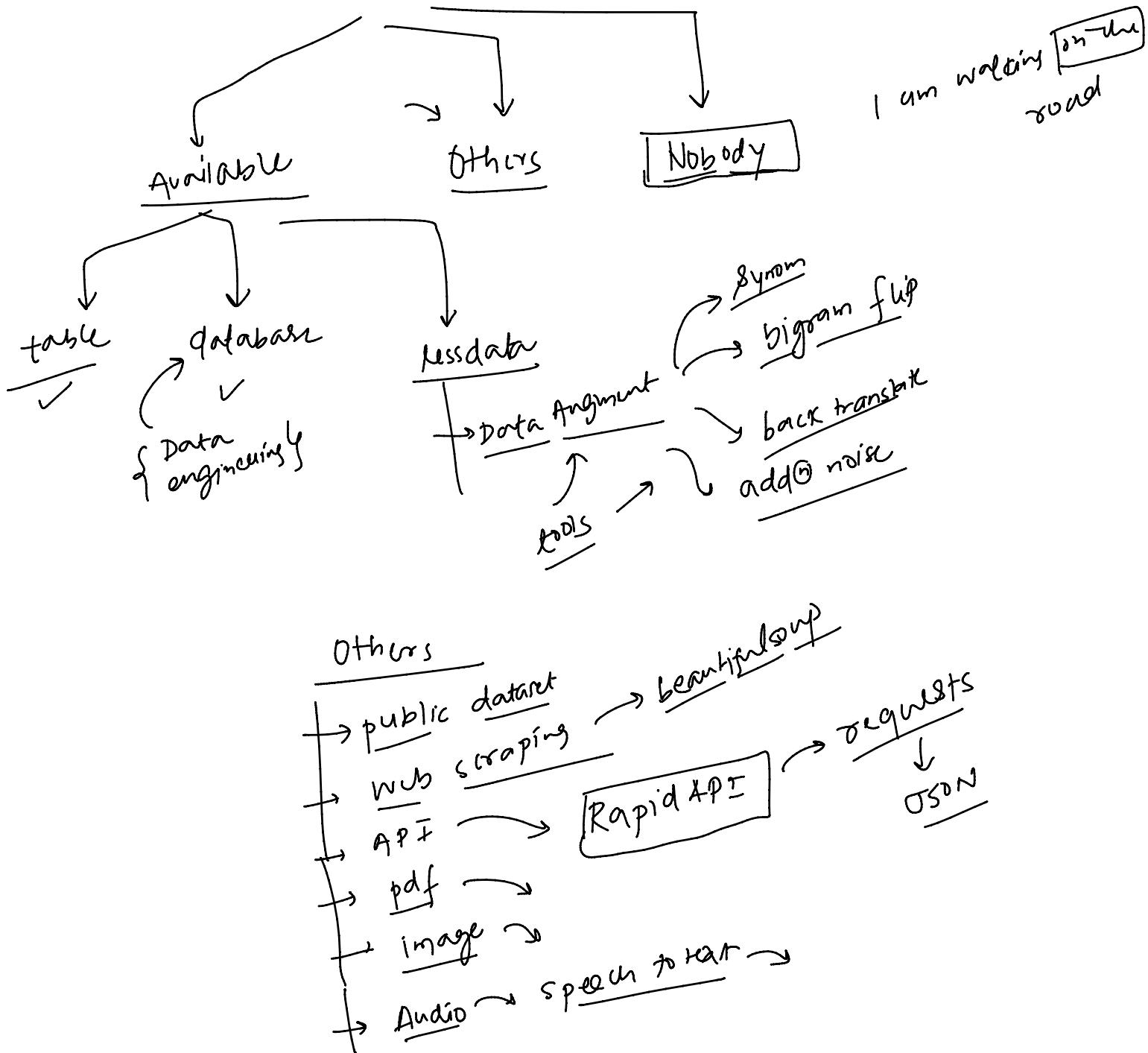


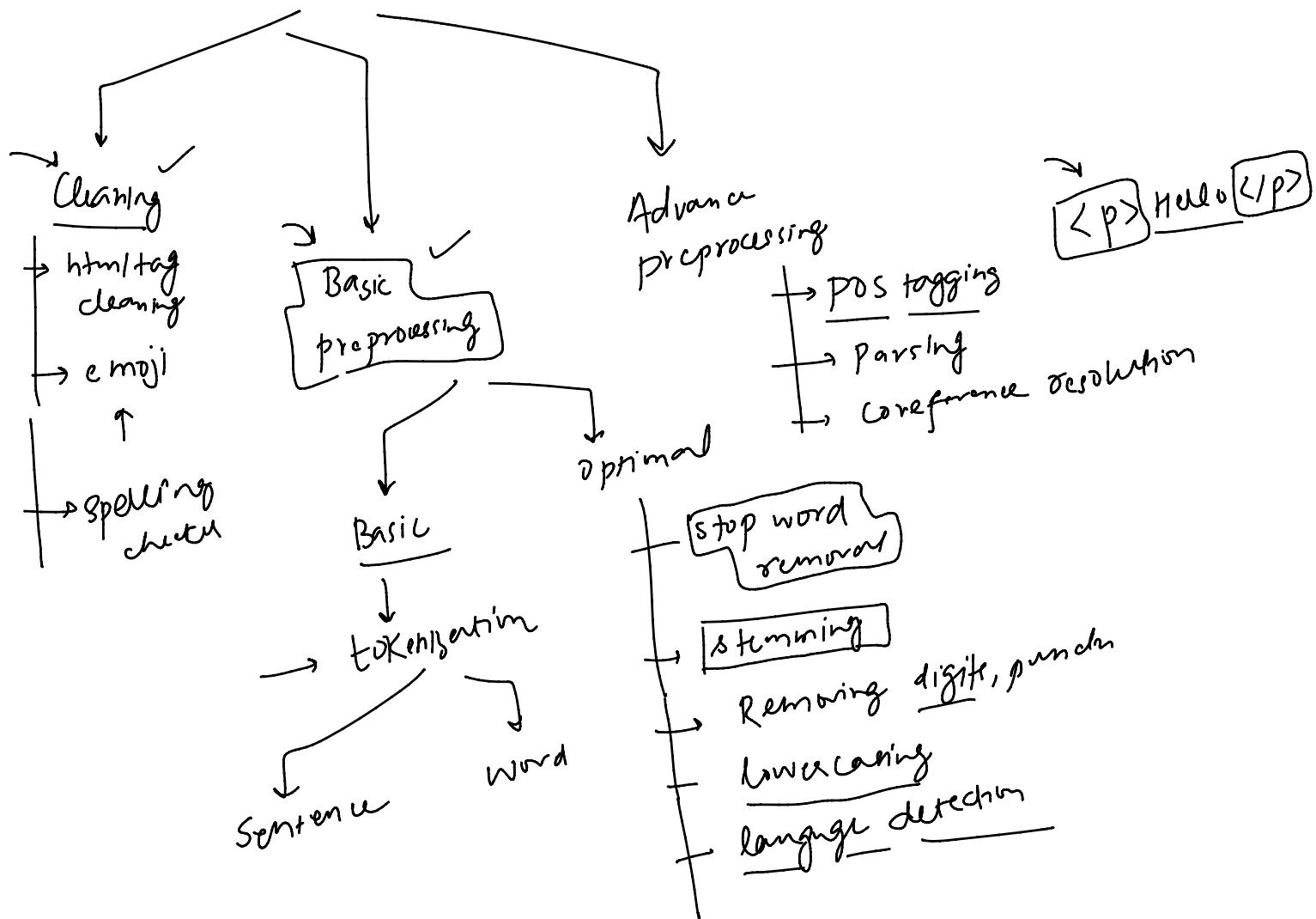
Data Acquisition

Saturday, November 20, 2021 8:02 AM



Text Preparation

Saturday, November 27, 2021 12:47 PM



Input

Chaplin wrote, directed, and composed the music for most of his films.

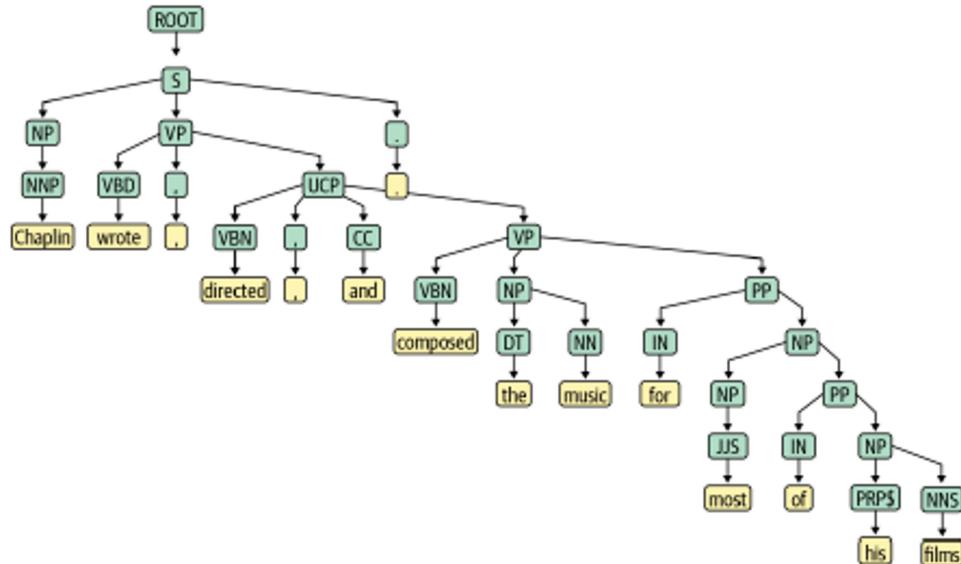
Tokenization with Lemmatization

Chaplin write . direct . and compose . the music for most of he film .
Chaplin wrote, directed, and composed the music for most of his films.

POS Tagging

NNP VBD . VBD CC VBN DT NN IN JJ\$ IN PRPS NNS .
Chaplin wrote, directed, and composed the music for most of his films.

Parse Tree

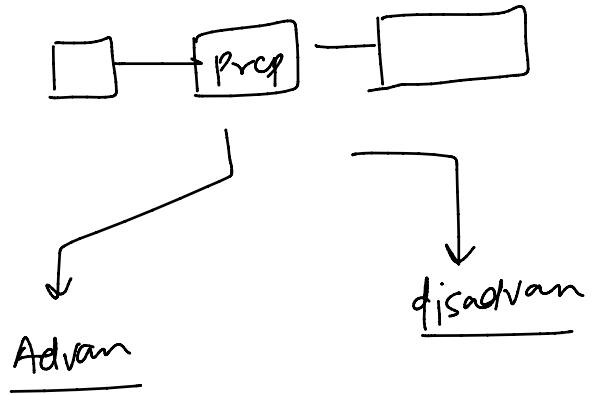
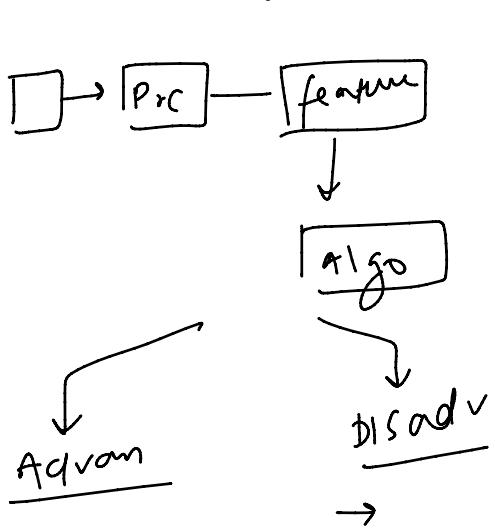
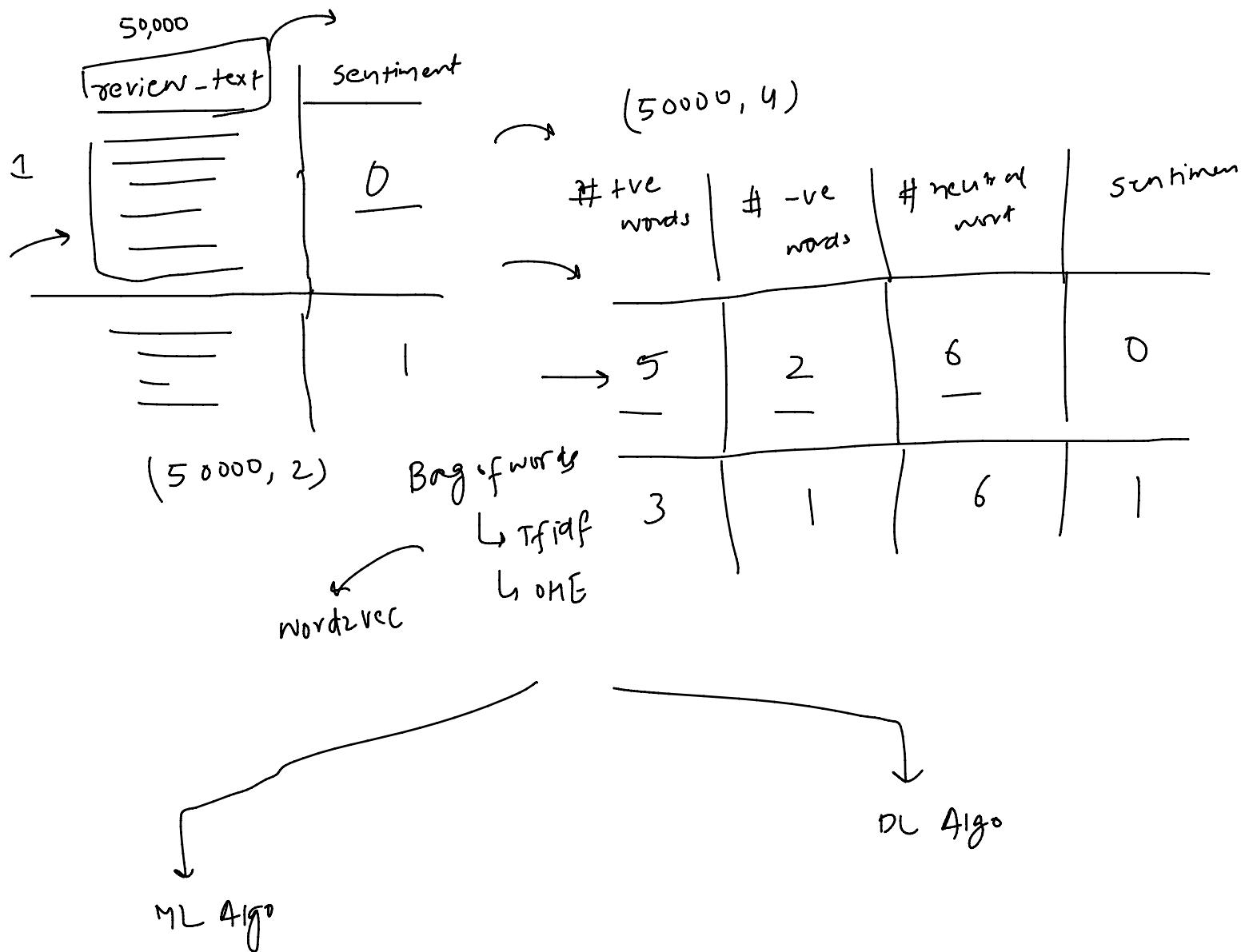


Coreference Resolution

Mention ----- coref ----- Mention
Chaplin wrote, directed, and composed the music for most of his films.

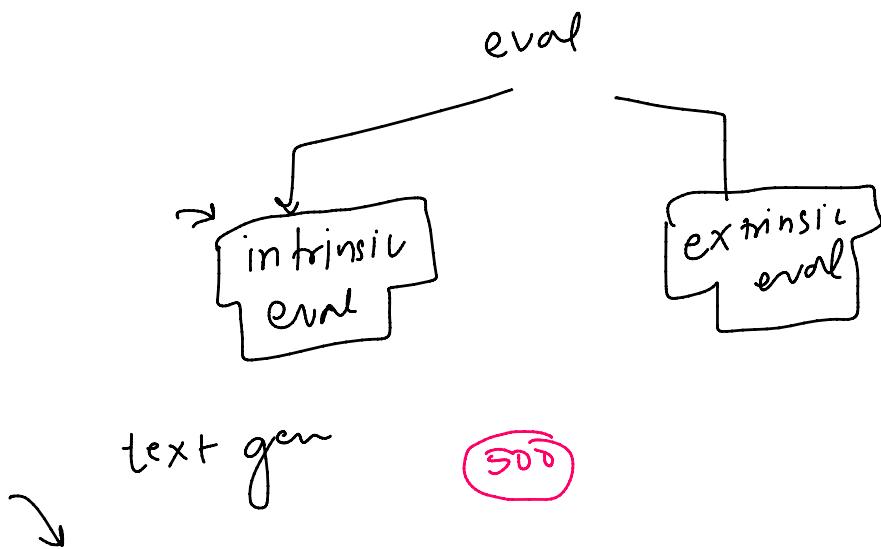
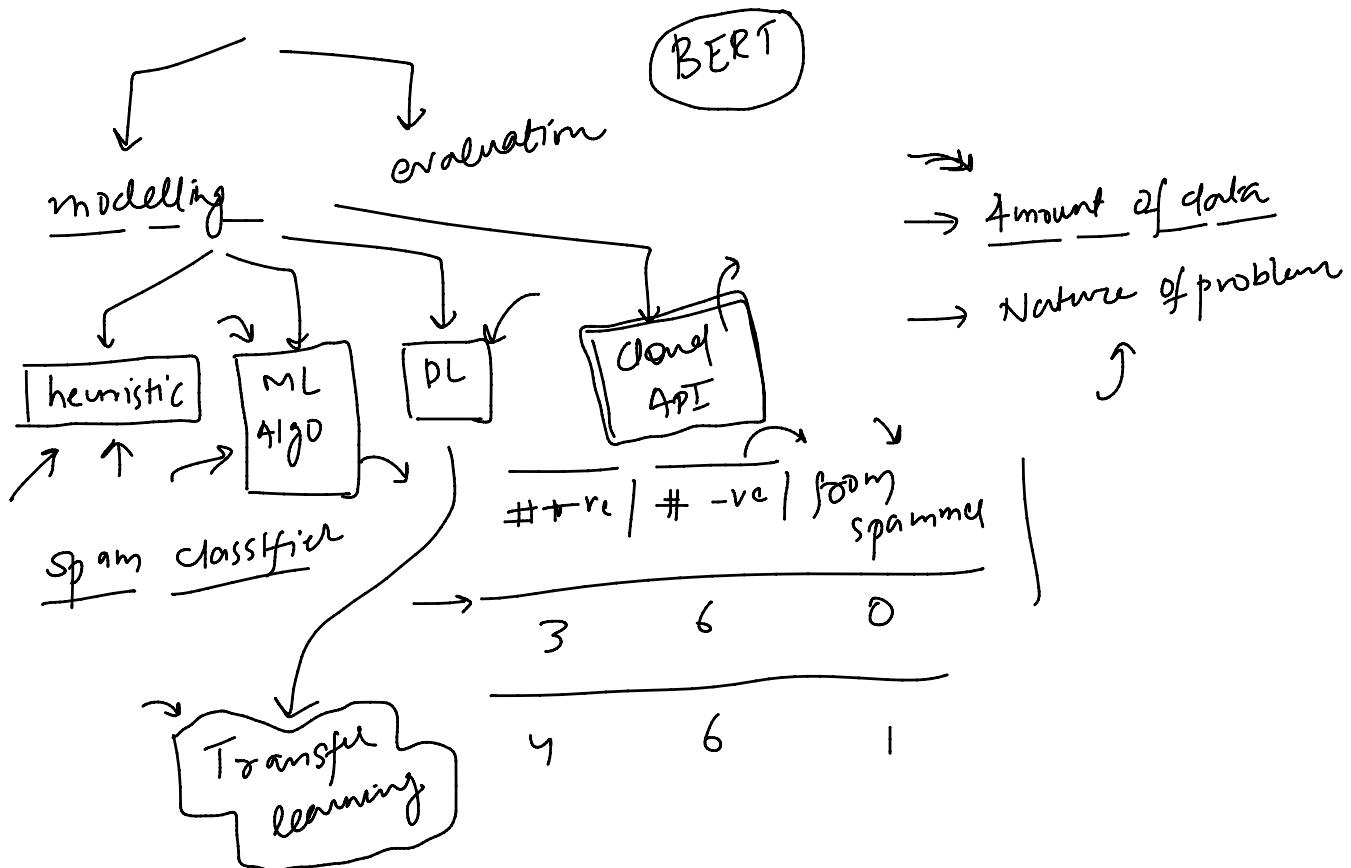
Feature Engineering

Tuesday, November 30, 2021 10:44 AM

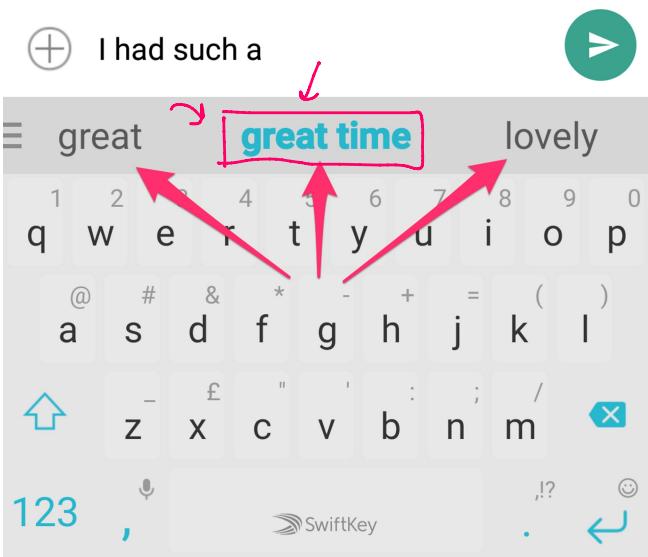


Modelling

Tuesday, November 30, 2021 11:14 AM



⊕ I had such a



perplexity

NP

extrinsic

1000

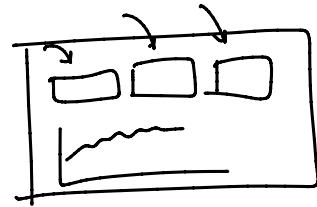
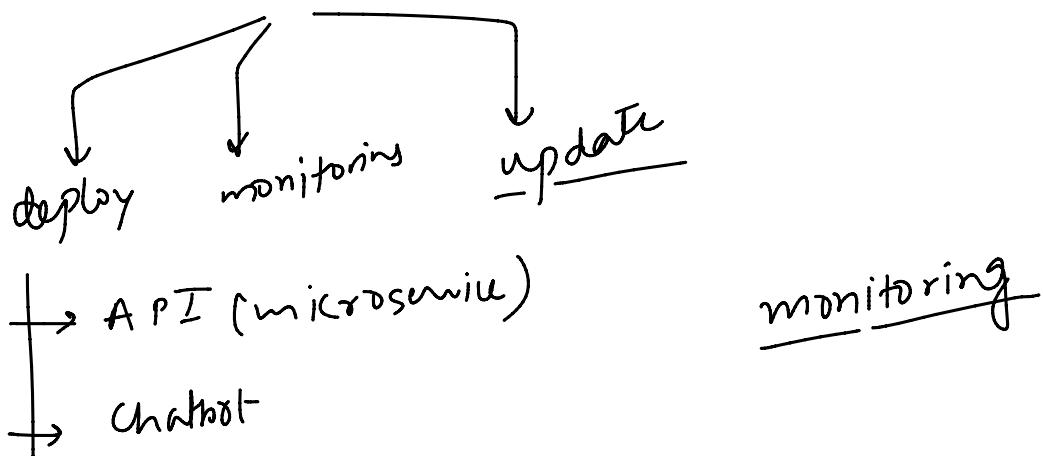


0

Deployment

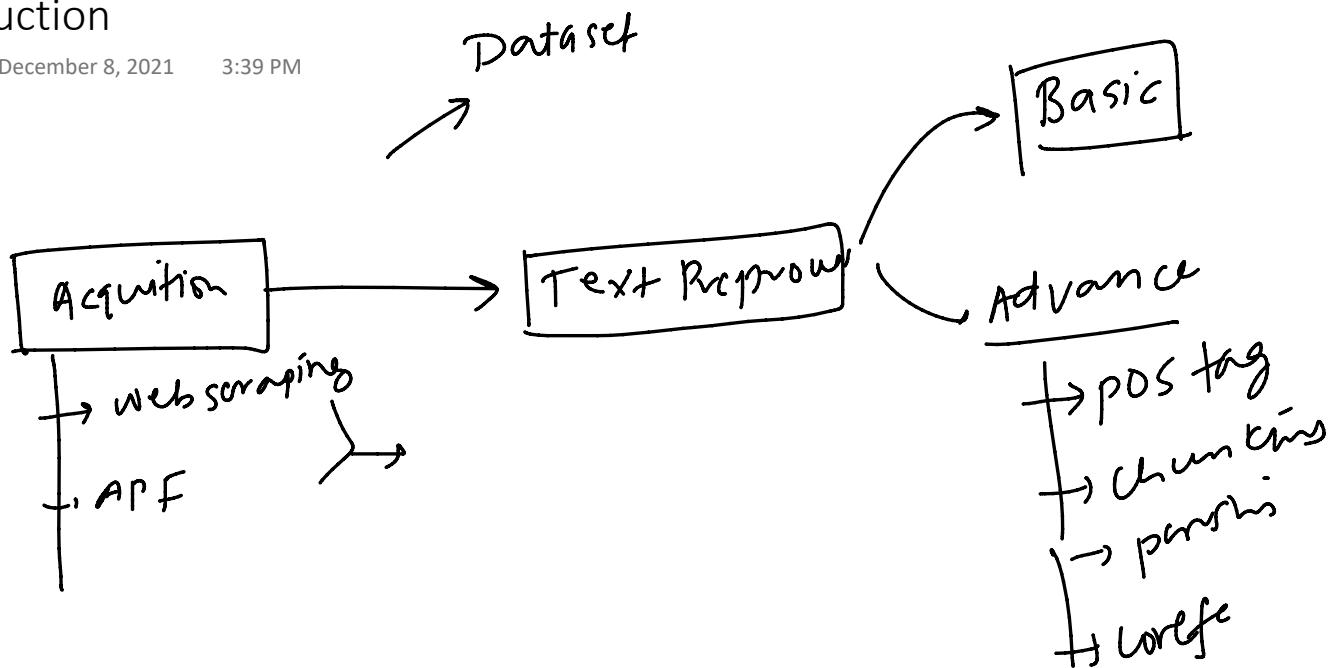
Tuesday, November 30, 2021

12:50 PM



Introduction

Wednesday, December 8, 2021 3:39 PM



Lowercasing

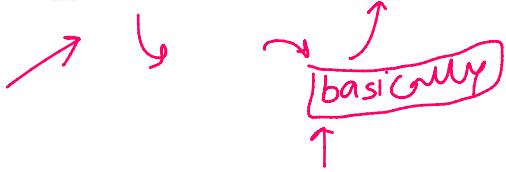
Wednesday, December 8, 2021 3:40 PM

"Basically there's a family where a little boy (Jake) thinks there's a zombie in his closet & his parents are fighting all the time.

This movie is slower than a soap opera... and suddenly, Jake decides to become Rambo and kill the zombie.

OK, first of all when you're going to make a film you must Decide if its a thriller or a drama! As a drama the movie is watchable. Parents are divorcing & arguing like in real life. And then we have Jake with his closet which totally ruins all the film! I expected to see a BOOGEYMAN similar movie, and instead i watched a drama with some meaningless thriller spots.

3 out of 10 just for the well playing parents & descent dialogs. As for the shots with Jake: just ignore them."



Remove HTML Tags

Wednesday, December 8, 2021 3:53 PM

"basically there's a family where a little boy (jake) thinks there's a zombie in his closet & his parents are fighting all the time.

this movie is slower than a soap opera... and suddenly, jake decides to become rambo and kill the zombie.

ok, first of all when you're going to make a film you must decide if its a thriller or a drama! as a drama the movie is watchable. parents are divorcing & arguing like in real life. and then we have jake with his closet which totally ruins all the film! i expected to see a boogeyman similar movie, and instead i watched a drama with some meaningless thriller spots.

3 out of 10 just for the well playing parents & descent dialogs. as for the shots with jake: just ignore them."

  regex

Remove URLs

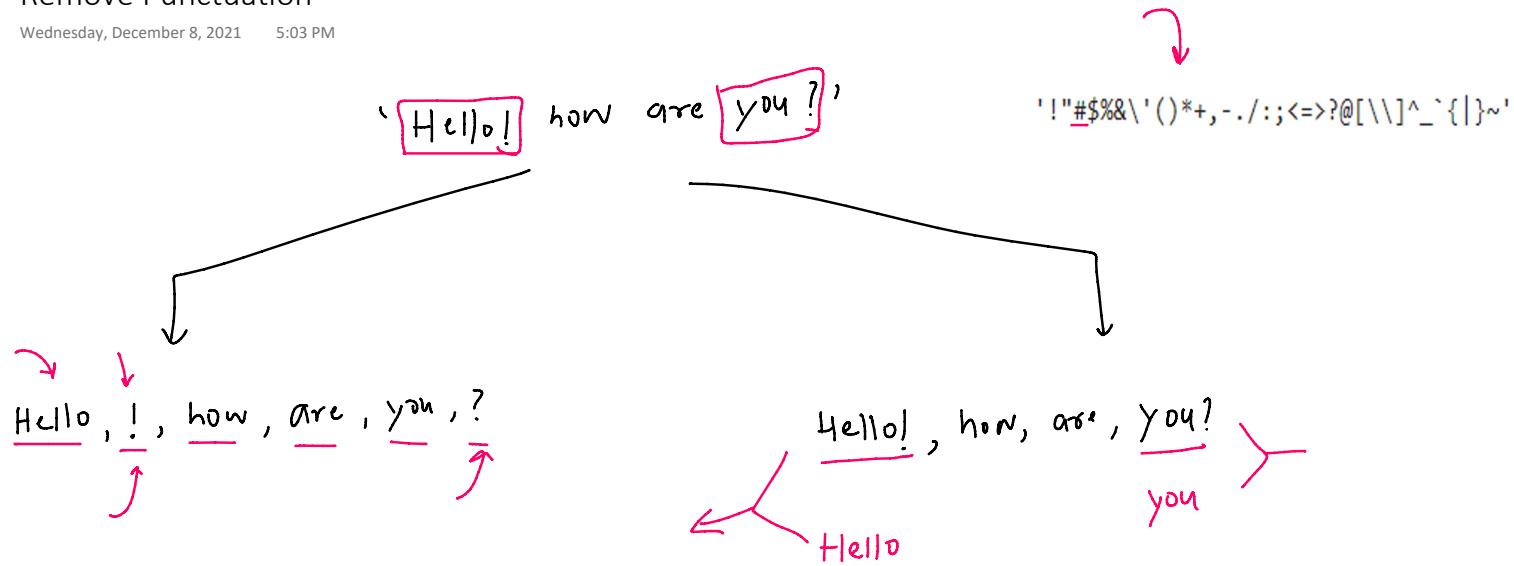
Wednesday, December 8, 2021 4:23 PM

'For notebook click <https://www.kaggle.com/campusx/notebook8223fc1abb> to search check www.google.com'



Remove Punctuation

Wednesday, December 8, 2021 5:03 PM



Chat word treatment

Thursday, December 9, 2021 1:54 AM

- { ✓ rofl →
 - ✓ lmao ↗
 - ✓ imho
-
- { ✓ fyi
 - ✓ ASAP ↗
 - ✓ gn → good night

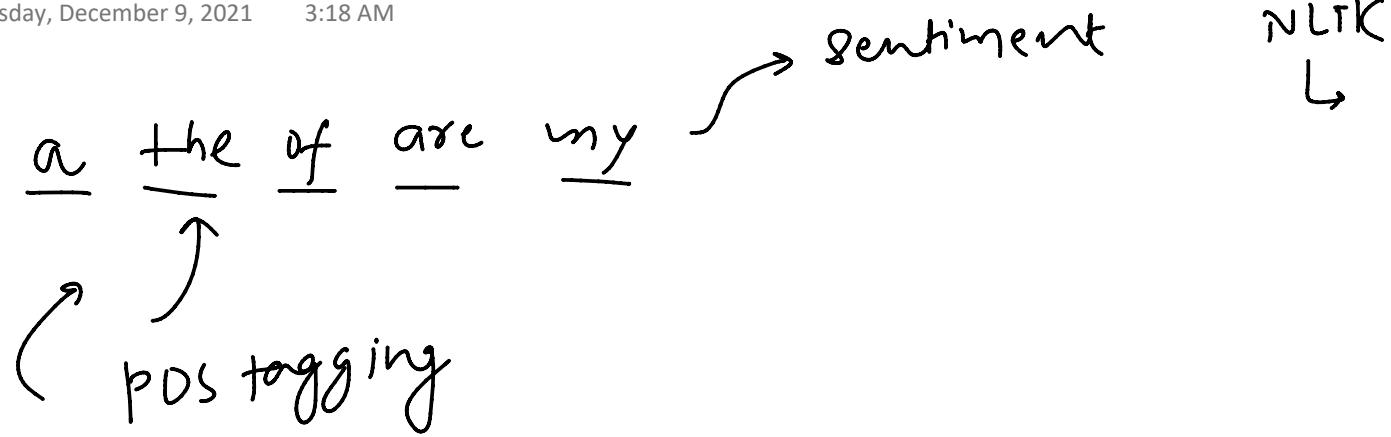
Spelling Correction

Thursday, December 9, 2021 2:19 AM

Please read the note book, and also like the notebook

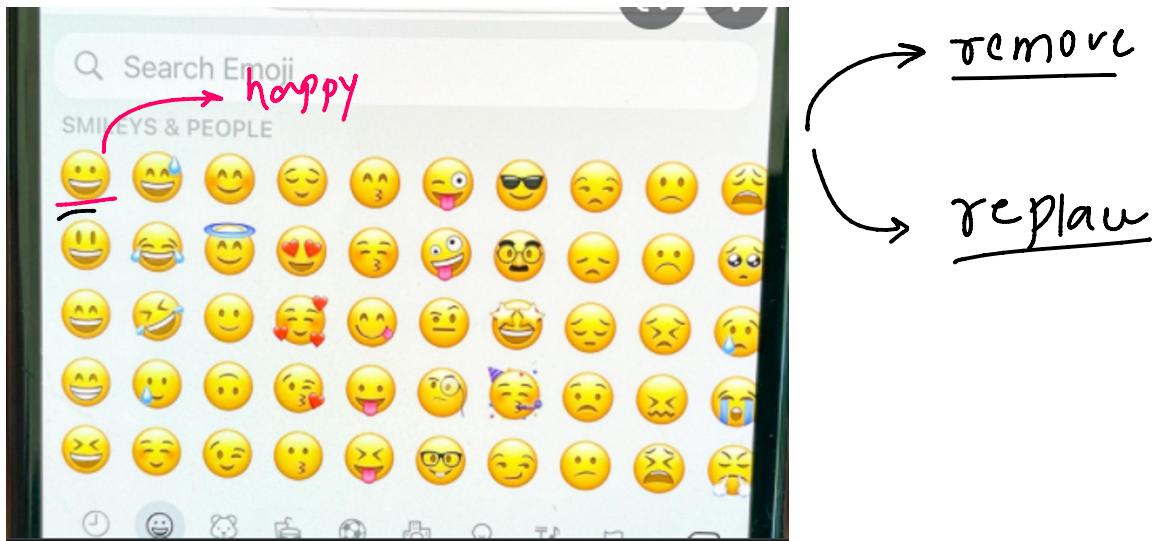
Removing Stop words

Thursday, December 9, 2021 3:18 AM



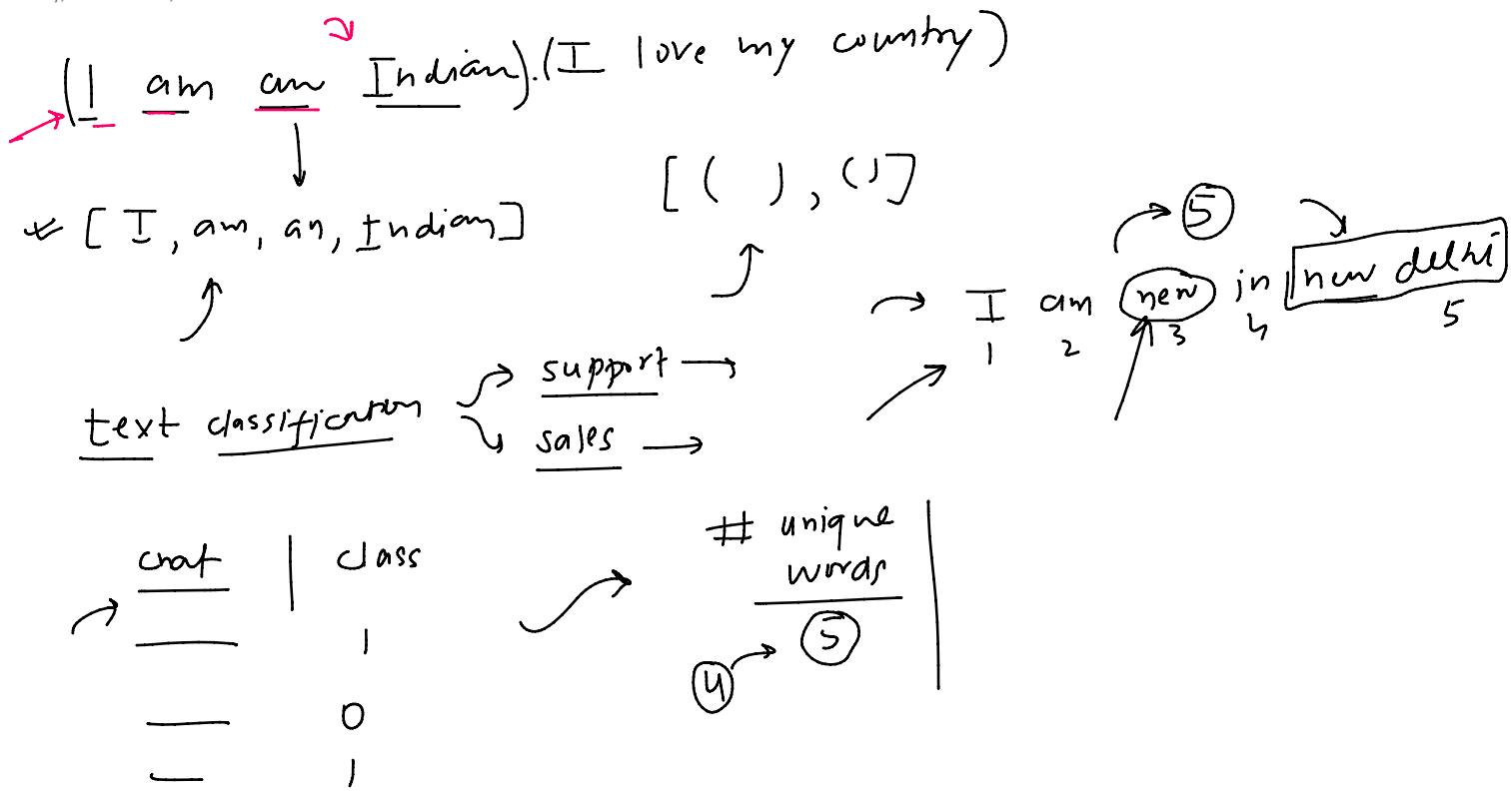
Handling Emojis

Thursday, December 9, 2021 12:07 PM



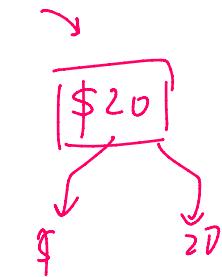
Tokenization

Friday, December 10, 2021 11:01 AM



- **Prefix:** Character(s) at the beginning
- **Suffix:** Character(s) at the end
- **Infix:** Character(s) in between
- **Exception:** Special-case rule to split a string into several tokens or prevent a token from being split when punctuation rules are applied

\$ ("i"
km), ..!"
--- / ...
let's U.S.



New-York
gr

let → ①
U e U.S.
e s

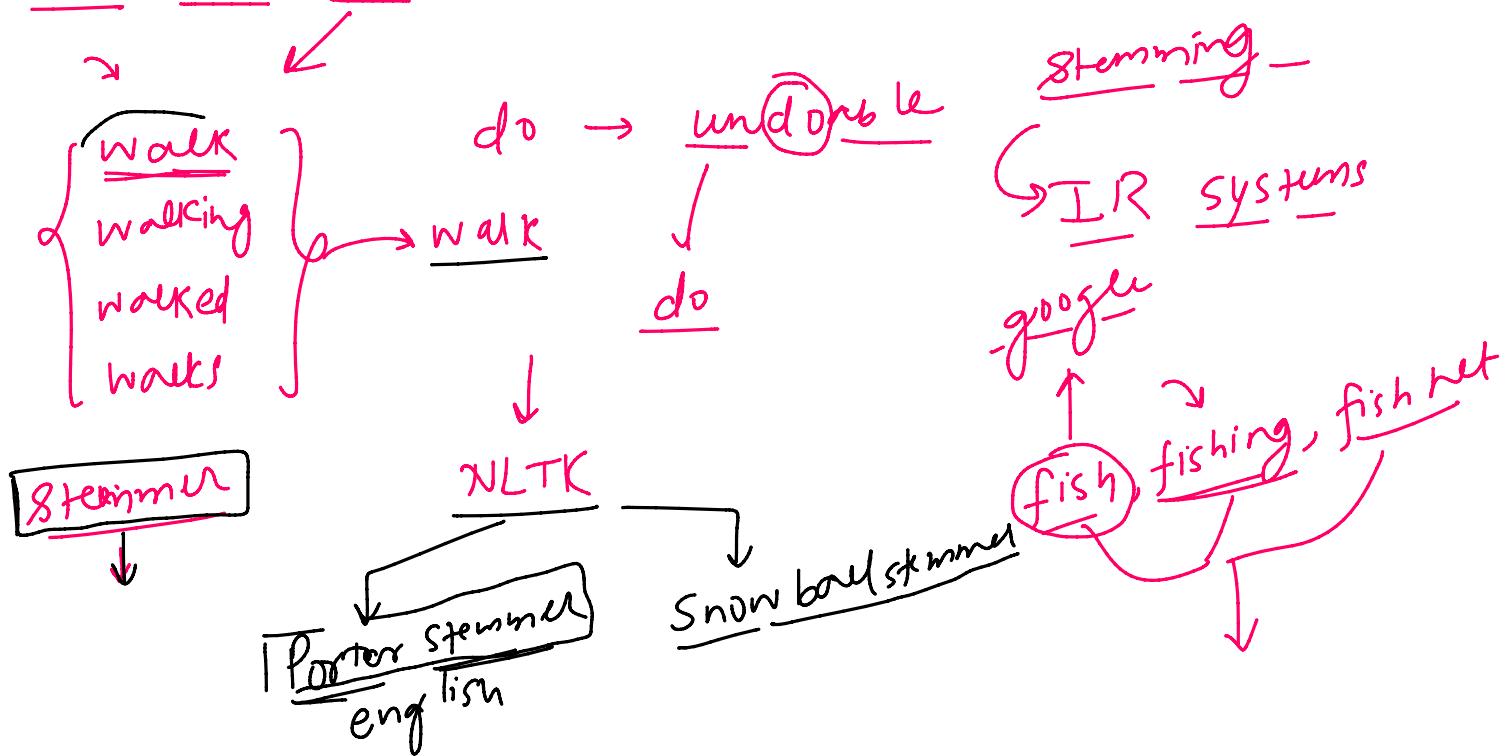
delhi!
det ?

10cm
10 km

Stemming

Friday, December 10, 2021 11:47 AM

"In grammar, inflection is the modification of a word to express different grammatical categories such as tense, case, voice, aspect, person, number, gender, and mood.



"Stemming is the process of reducing inflection in words to their root forms such as mapping a group of words to the same stem even if the stem itself is not a valid word in the Language."

→ n lization

→ lemmatization

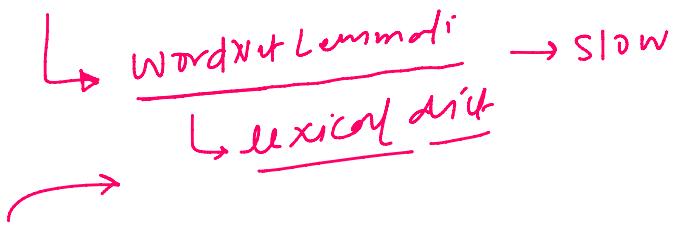
Lemmatization

Friday, December 10, 2021 12:23 PM

Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization root word is called Lemma. A lemma (plural lemmas or lemmata) is the canonical form, dictionary form, or citation form of a set of words.

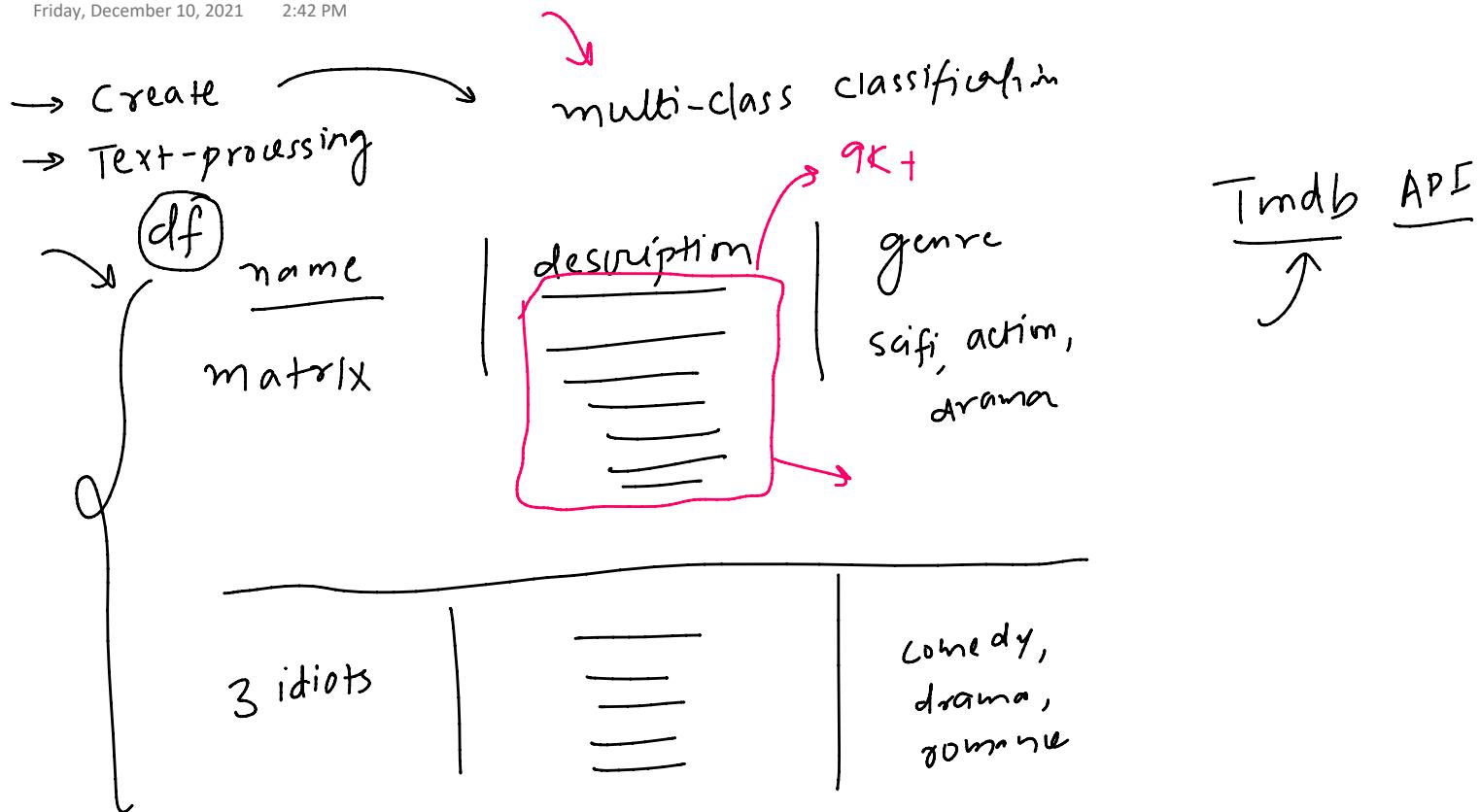
lemma

algorithm
suffix
fast



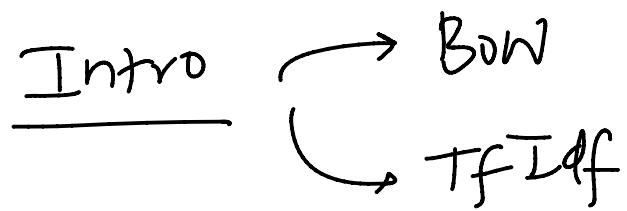
Assignment

Friday, December 10, 2021 2:42 PM



Plan of Attack

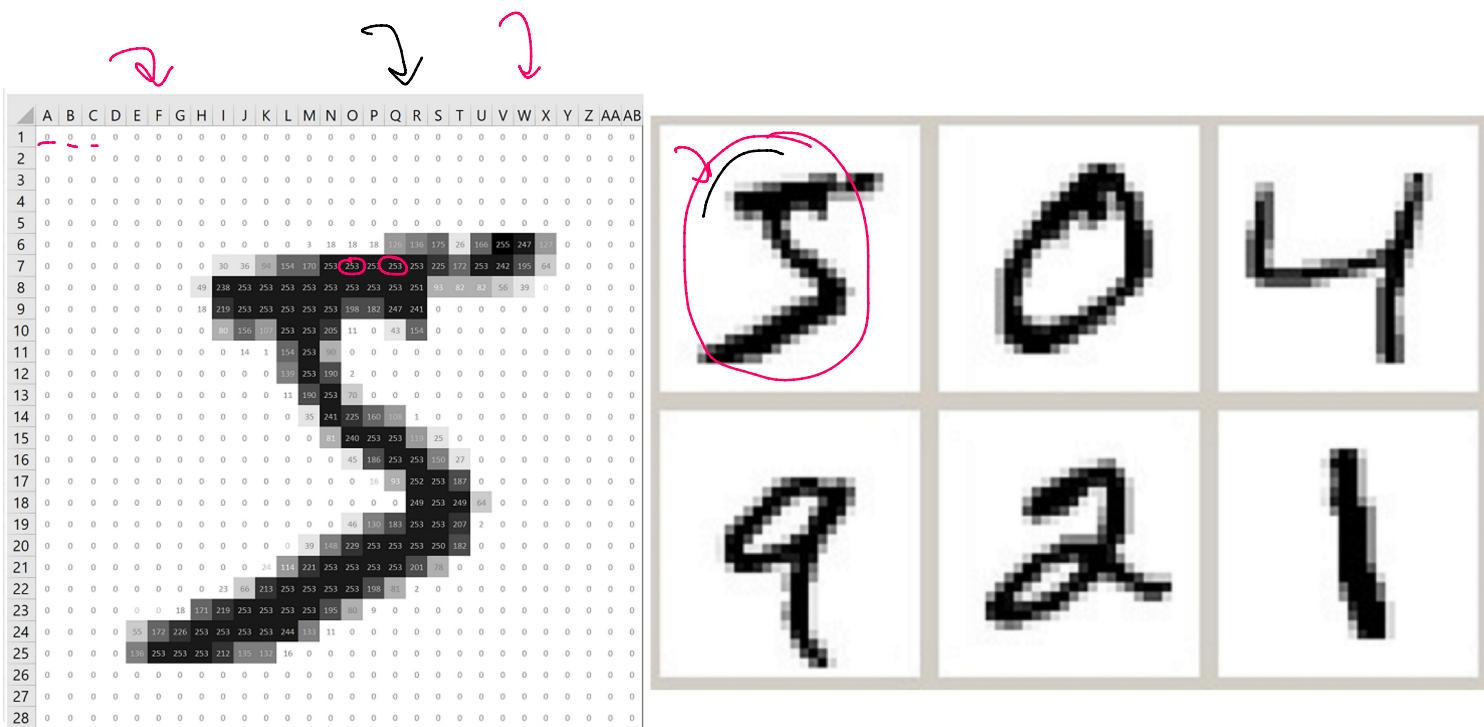
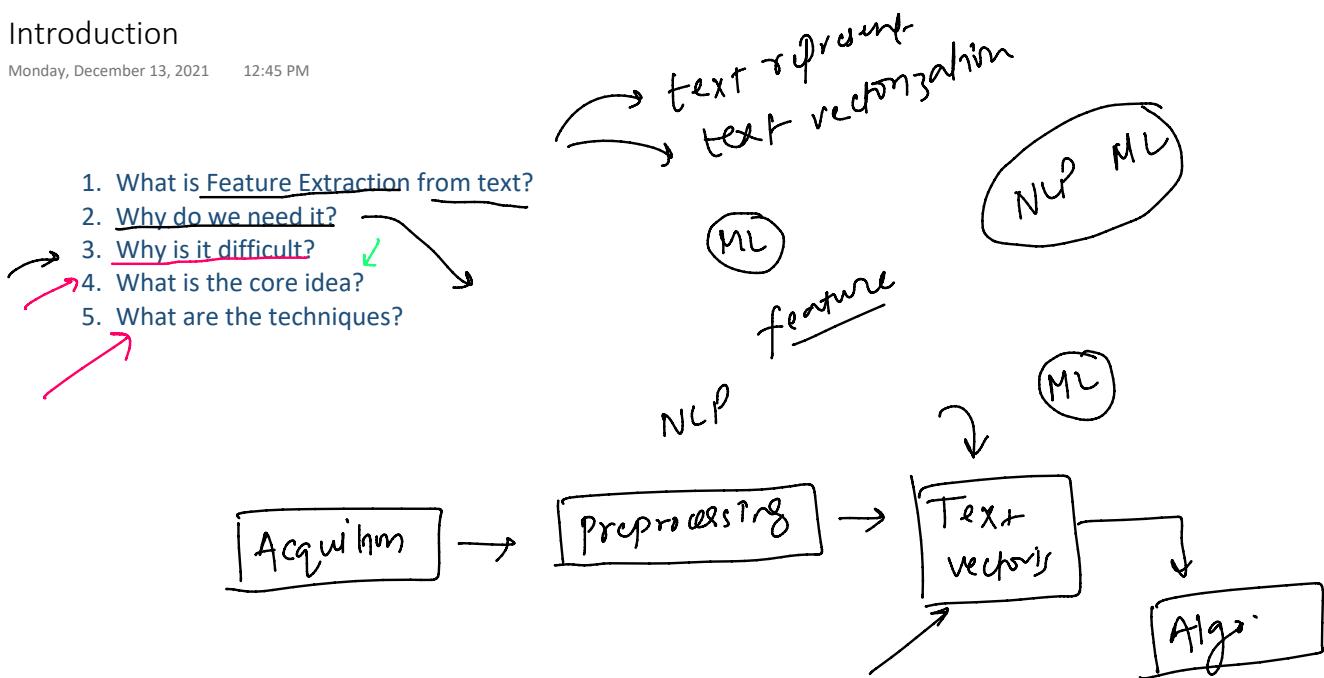
Monday, December 13, 2021 11:05 AM



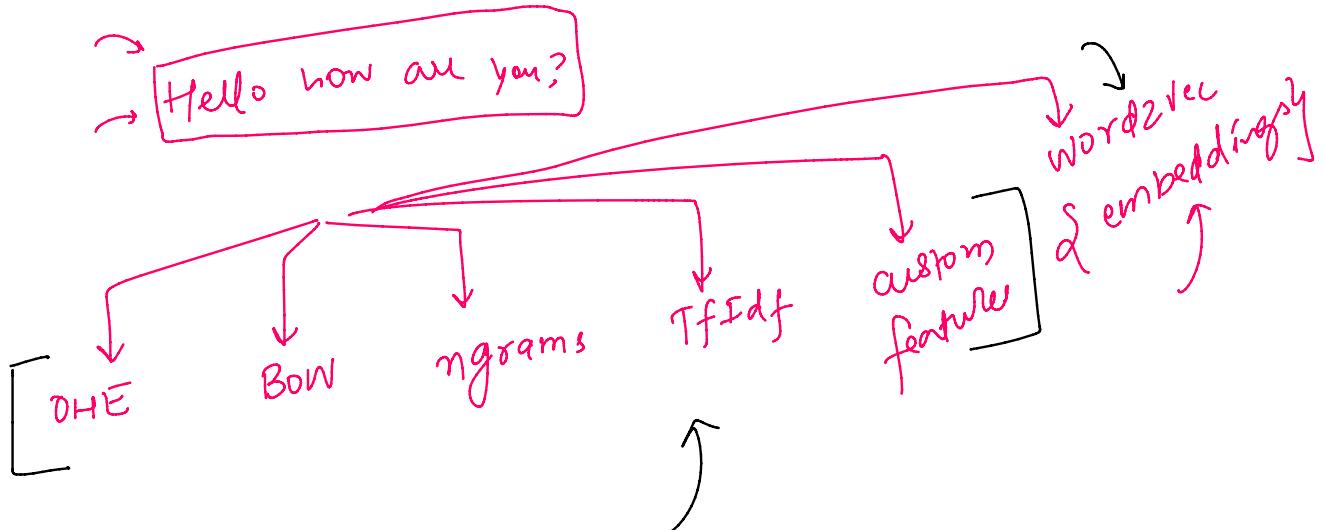
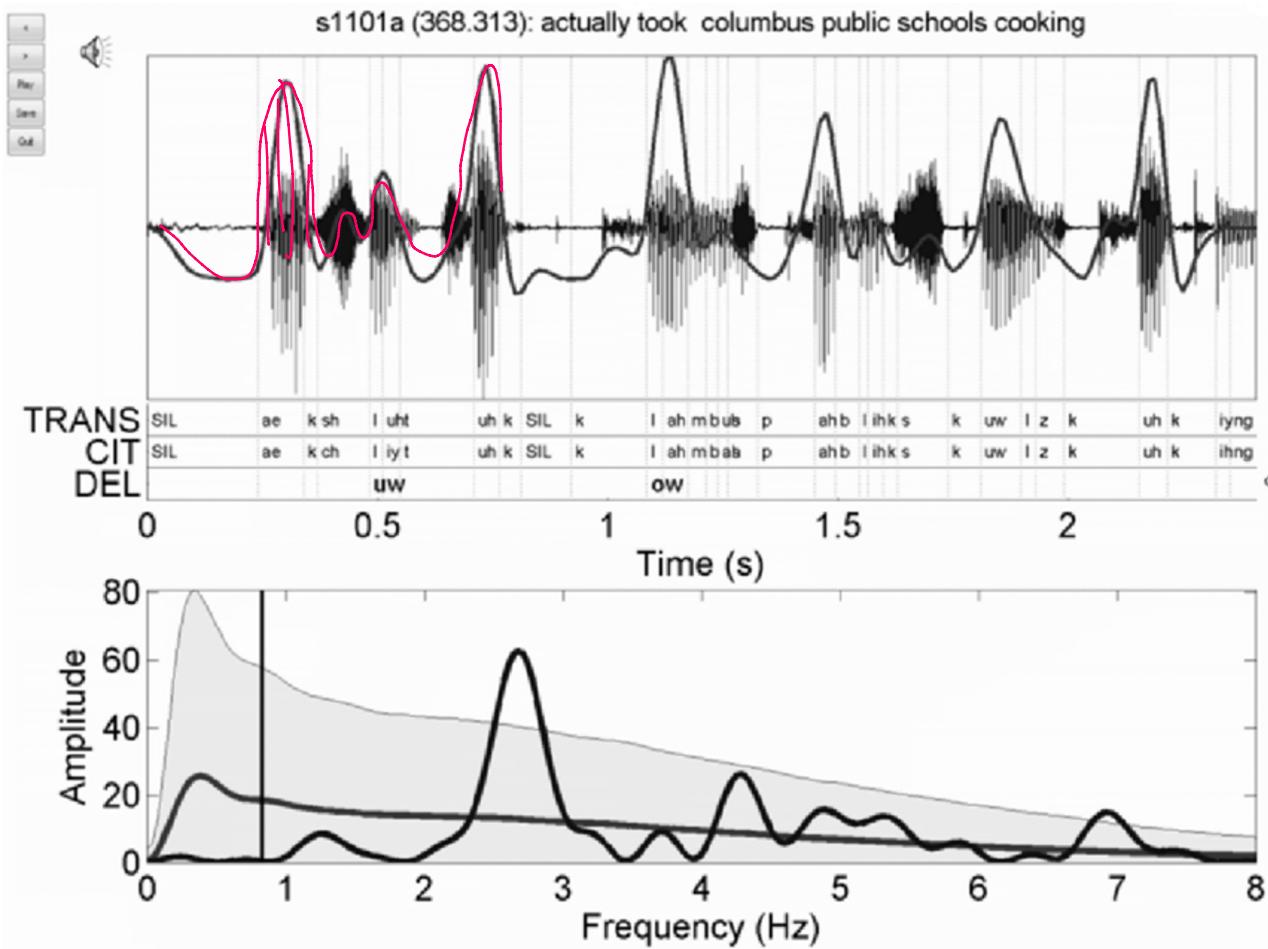
Assignment

Introduction

Monday, December 13, 2021 12:45 PM



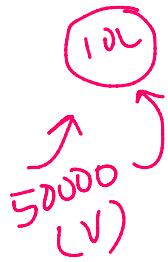
→ [23, 45, 64, 61, 72]



Common Terms

Monday, December 13, 2021 1:07 PM

1. Corpus C
2. Vocabulary V
3. Document D
4. Word w



Imdb 50K

A review	A sentiment
49582 unique values	2 unique values
<p>One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. The...</p>	positive ✓
A wonderful little production. The filming technique is very unassuming- very old-time-B...	positive
I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air con...	positive
Basically there's a family where a little boy (Jake) thinks there's a zombie in his closet & his par...	negative

One Hot Encoding

Monday, December 13, 2021 1:15 PM

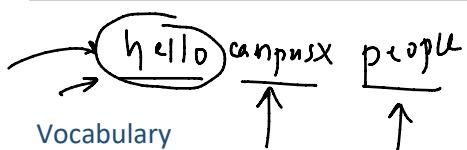
$5000 \rightarrow 100 \text{ words} \rightarrow V \rightarrow 50,000$

ML Algo

D1	people <u>watch campusx</u>	1
D2	campusx watch campusx	1
D3	people write comment	0
D4	campusx write comment	1

Corpus

people watch campusx campusx watch campusx
people write comment campusx write comment



$1 \rightarrow n \rightarrow V \text{ dim}$

$V = 50,000$] sparse array

people watch campusx write comment

$$V = 5$$

people	watch	campusx	write	comment
0	1	0	0	0
0	0	1	0	0

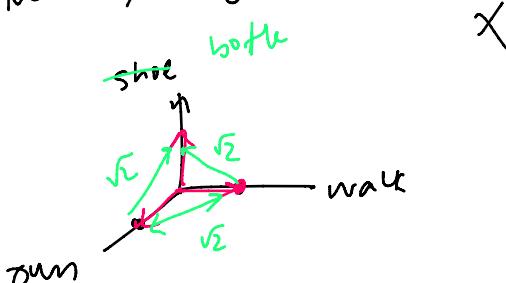
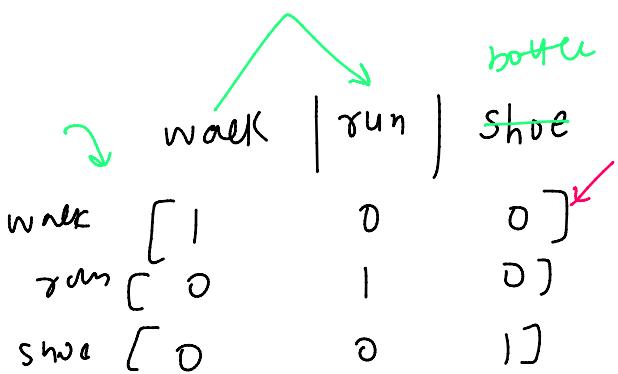
$$[001]$$

$$\underline{(3,5)} \quad D_1 = [\begin{bmatrix} 1, 0, 0, 0, 0 \\ 0, 1, 0, 0, 0 \\ 0, 0, 1, 0, 0 \end{bmatrix},]$$

$$D_2 = [\begin{bmatrix} 0, 0, 1, 0, 0 \\ 0, 1, 0, 0, 0 \\ 0, 0, 1, 0, 0 \end{bmatrix},] \rightarrow (3,5)$$

- Pros
- intuitive
 - easy to implement

- Flaws
- #1 sparsity →
 - #2 No fixed size →
 - #3 ODV →
 - #4 No capturing of semantic



Bag of Words

Monday, December 13, 2021 6:22 PM

50000

→ Text Classification

binary = True / False

vdim (5d)

		text	output
D1	people watch campusx	1	
D2	campusx watch campusx	1	
D3	people write comment	0	
D4	campusx write comment	0	

	people	watch	campusx	write	comment
D1	1	1	1	0	0
D2	0	1	2	0	0
D3	1	0	0	1	1
D4	0	0	1	1	1

core intuition

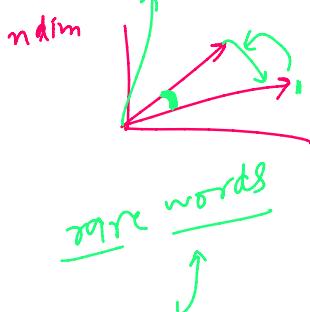
Bog

order of words

context

frequency

exist



of Count-rectly

mod-feature = ① / 2

Hello campw

Advantages

→ Simple and intuitive

→ ordering

Hello how are you?

RNN DTM

Disadvano

#1 sparsity

of This is a very good movie
This is not a very good movie

#2 OOV

→ Ingrams

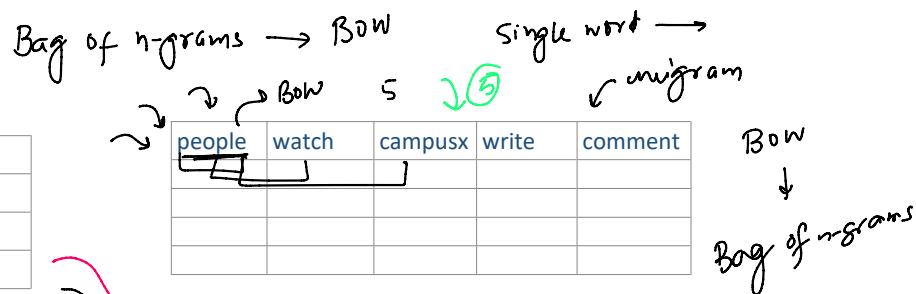
#3 ordering

#4

N-grams

Tuesday, December 14, 2021 10:57 AM

D1	people watch campusx	1
D2	campusx watch campusx	1
D3	people write comment	0
D4	campusx write comment	0



D1	people	watch	campusx	1
D2	campusx	watch	campusx	1
D3	people	write	comment	0
D4	campusx	write	comment	0

quad gram

→ error

n-grams

people watch campusx | 52 | 53 | 54 | 55
n grams

1	0	0	0	D1
0	1	0	0	D2
0	0	1	0	
0	0	0	1	

This movie is very good
This movie is not good

BOW ⑥

This movie is very good not

→ 1 1 1 1 1 0

→ 1 1 1 0 1 1

→ ... is very good & This movie | movie is | is very | very good | is not | not good

6D

Benefits

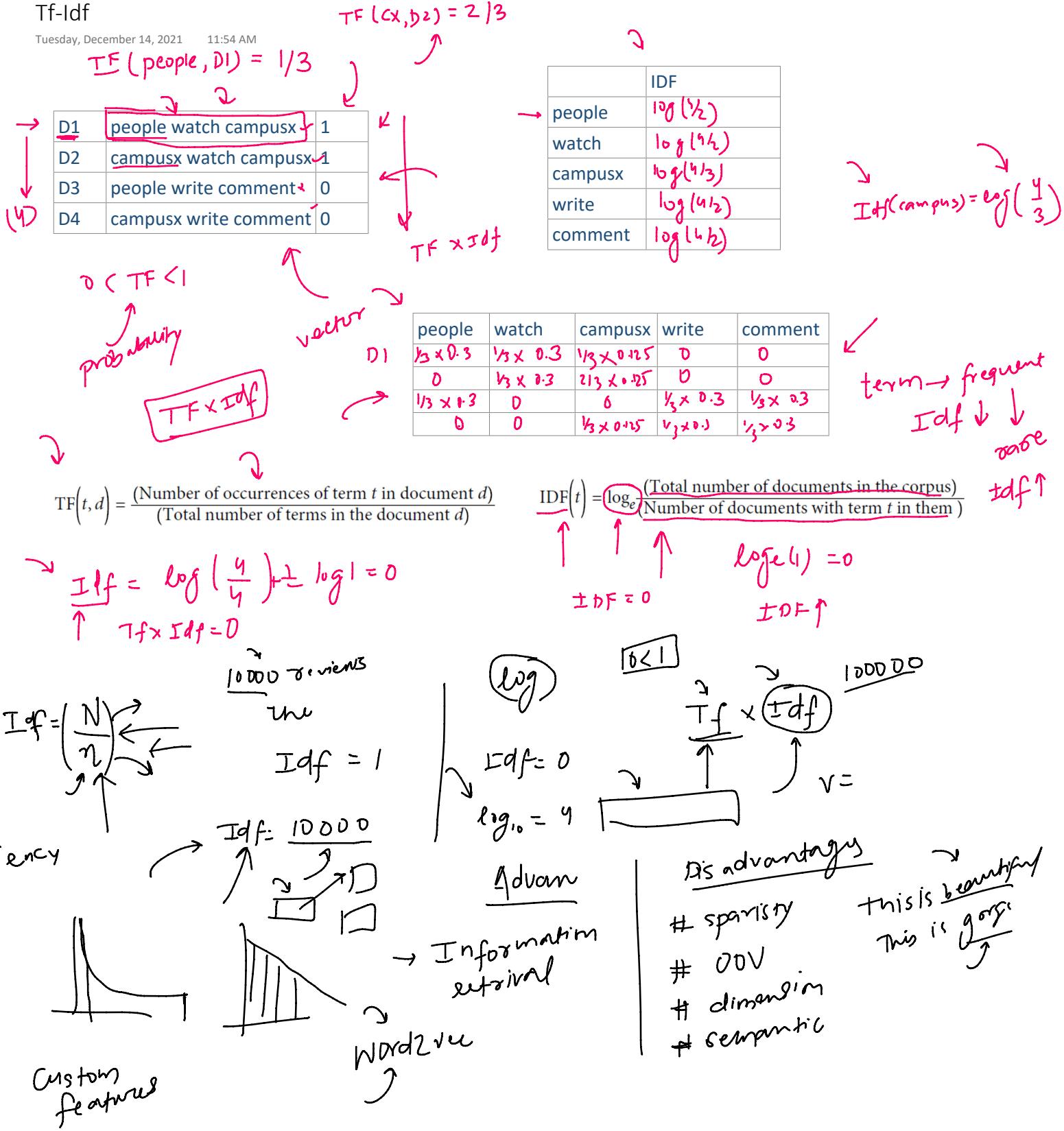
- 1) Able to capture semantic of the sentence
- 2) easy implement

Disadvantages
 → slows down the algo
 → dimension increased
 → Out of vocab

Large dataset
 unigrams → bi-gram
 bi-gram → tri-gram

Tf-Idf

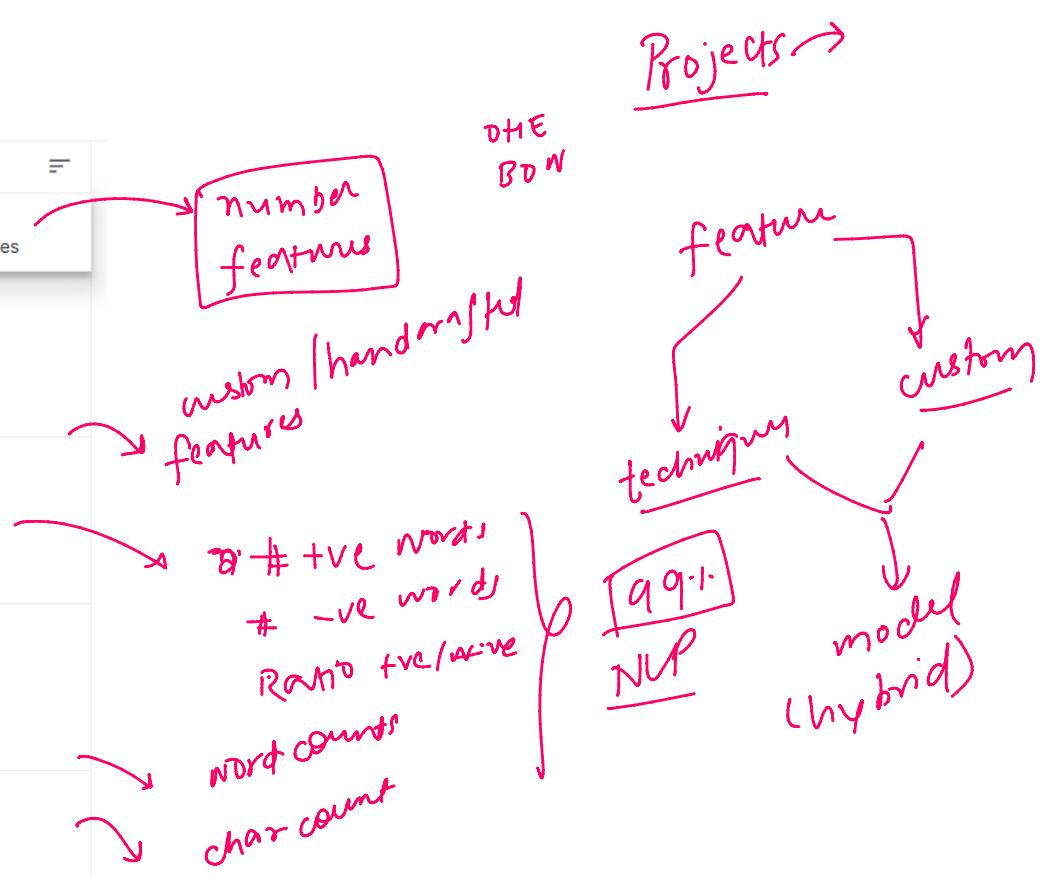
Tuesday, December 14, 2021 11:54 AM



Custom Features

Tuesday, December 14, 2021 2:59 PM

A review	A sentiment
49582 unique values	2 unique values
One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. The...	positive
A wonderful little production. The filming technique is very unassuming- very old-time-B...	positive
I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air con...	positive
Basically there's a family where a little boy (Jake) thinks there's a zombie in his closet & his par...	negative

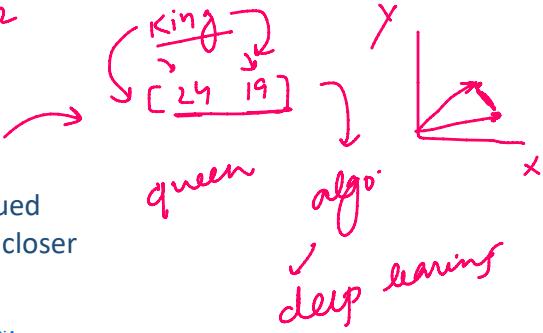


[Word Embeddings] →

Wednesday, December 22, 2021

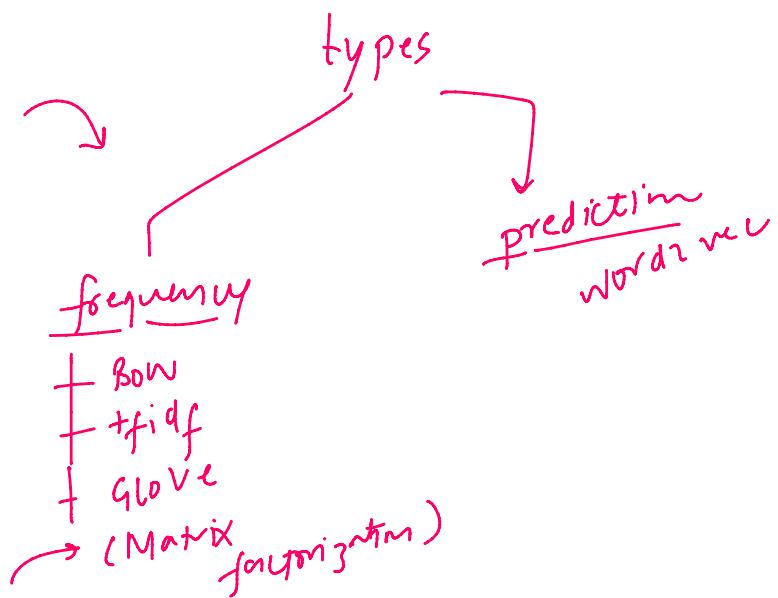
4:11 PM

tf if Bow, word



In natural language processing, word embedding is a term used for the representation of words for text analysis, typically in the form of a real-valued vector that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning.

From <<https://www.google.com/search?q=what+are+word+embeddings&oq=what+are+word+&aq=chrome.1.69i57j0i512l9.4201j0j4&sourceid=chrome&ie=UTF-8>>



What is Word2Vec

Wednesday, December 22, 2021 4:11 PM

→ what?

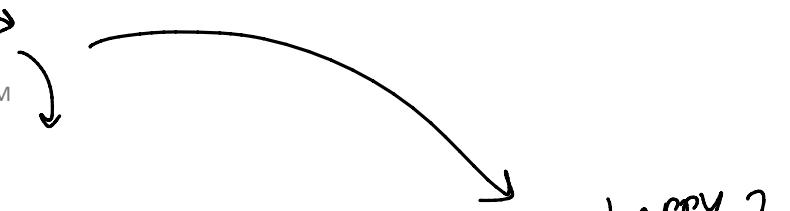
↳ google 2013

→ Bow / tfIdf

↓
high dim
100000dim

$\begin{bmatrix} 0 & 0 & 0 & \dots & 0 \end{bmatrix}$

sparse vector



1) semantic meaning ✓

2) low dim
 $[100 - 300]$ → computation

3) dense vector → non-zeros

→ overfitting

↳ deep learning

Demo

Wednesday, December 22, 2021 4:11 PM

- We will use the pre-trained weights of word2vec that was trained on Google News corpus containing 3 billion words. This model consists of 300-dimensional vectors for 3 million words and phrases.

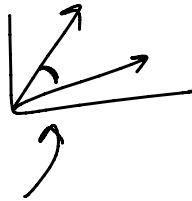
30 lac's

pretrained

self-trained

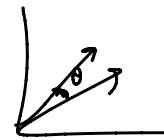
gensim

Vector → 30 lac words → vector → 300 numbers
↓ 300 dim space



[2]d

300 dim



Really Cool Stuff

Wednesday, December 29, 2021 2:08 PM

Intuition

Wednesday, December 22, 2021 4:12 PM

	1	2	3	4	5
gender	1	0	1	0	1
wealth	1	1	0.3	0.3	0
power	1	0.7	0.2	0.2	0
weight	0.8	0.4	0.6	0.5	0.3
speak	1	1	1	1	0

↳ vocab (5)

→ 30 features

automated

neural network

features

f_1 f_2 f_3 ...

0.5

[1 0 0 0.3 0]

King = man + woman

$$1 - 1 + 0 = 0$$

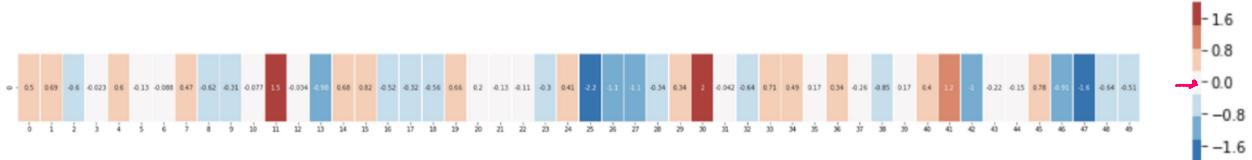
$$1 - 0.3 + 0.3 = 1$$

$$1 - 0.2 + 0.2 = 1$$

$$0.8 - 0.1 + 0.5 = 0.7$$

$$1 - 1 + 1 = 1$$

→ 50 numbers



f_{ff}

50 features

→ entire

f_{ff}

woman

girl

youth

boy

man

king

queen

water

the youth all played soccer on
soccer field
soccer player soccer and

The underlying assumption of Word2Vec is that two words sharing similar contexts also share a similar meaning and consequently a similar vector representation

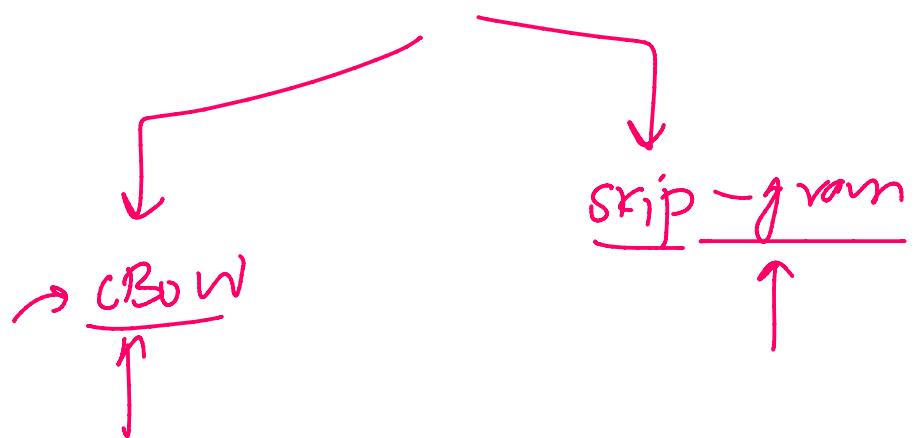
The underlying assumption of Word2Vec is that two words sharing similar contexts also share a similar meaning and consequently a similar vector representation from the model.

the L-
the hockey player took
goal.



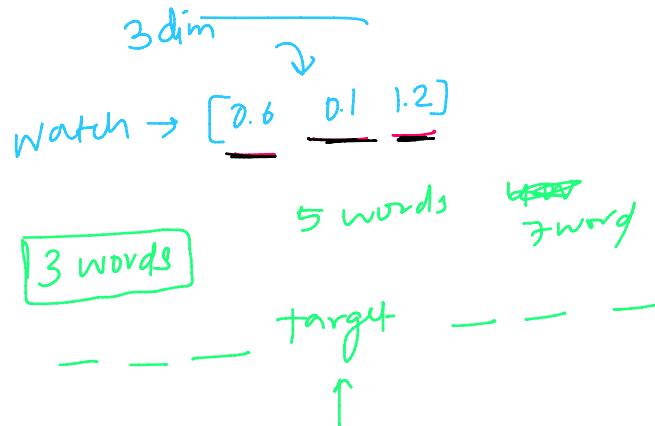
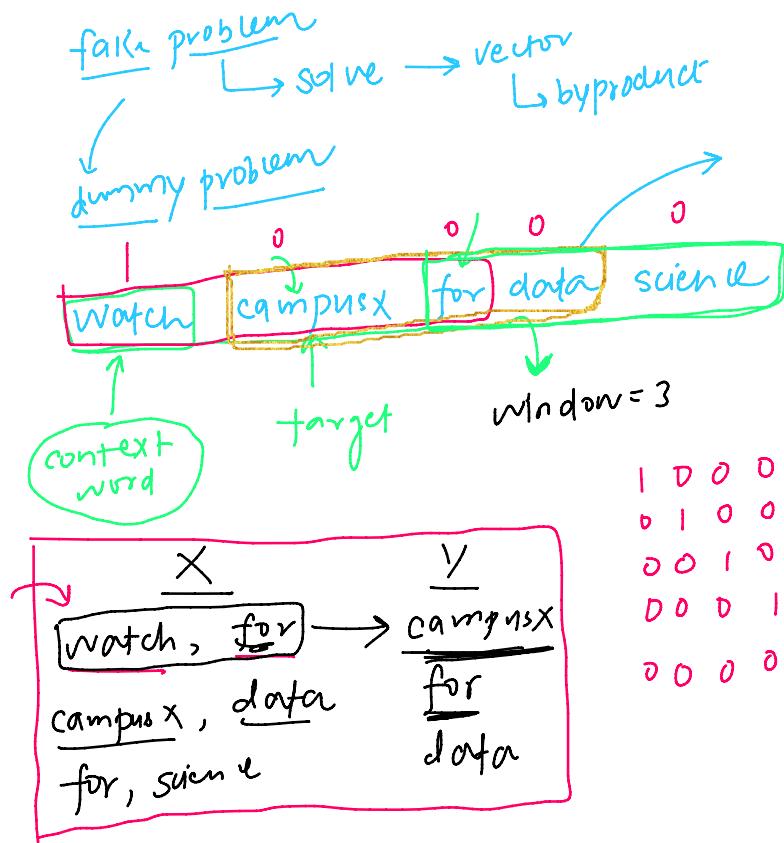
Types of Word2Vec

Wednesday, December 22, 2021 4:11 PM

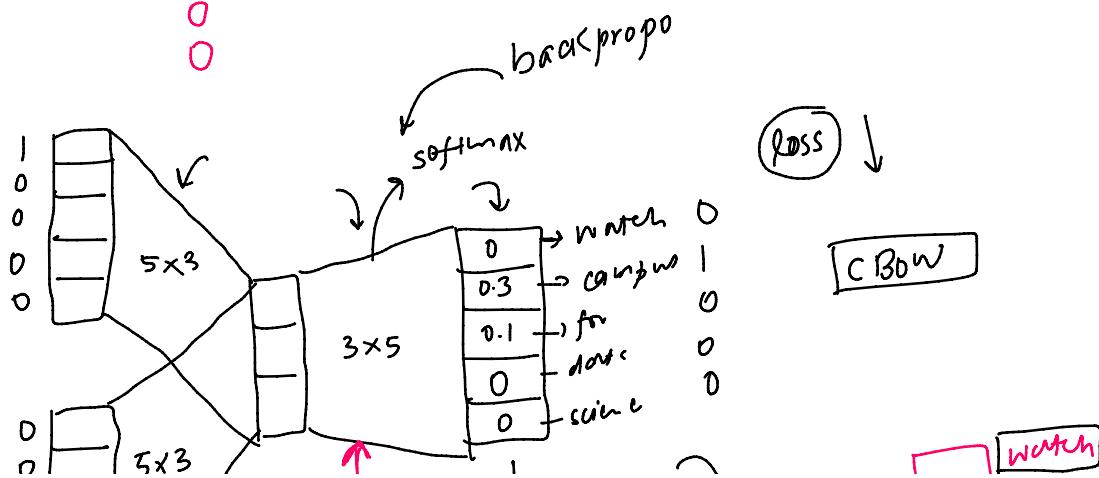
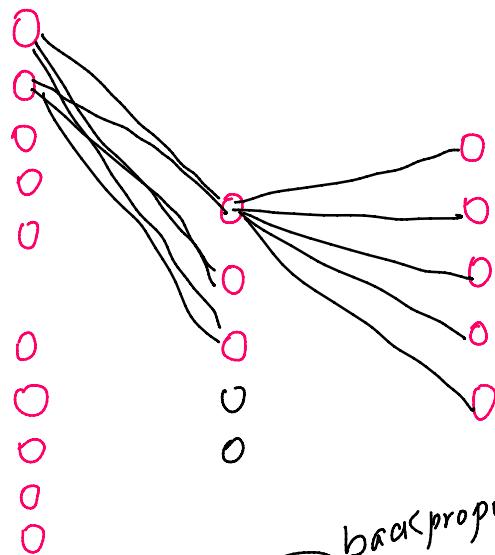


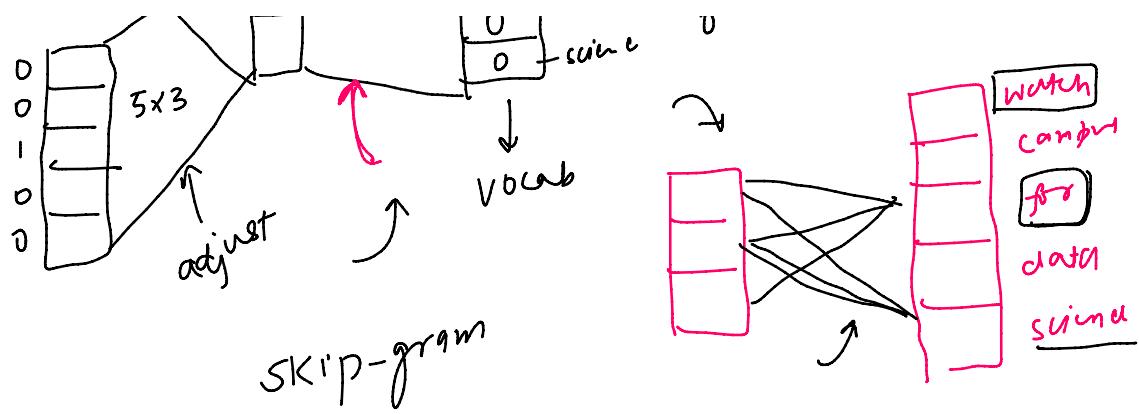
CBOW

Wednesday, December 22, 2021 4:12 PM



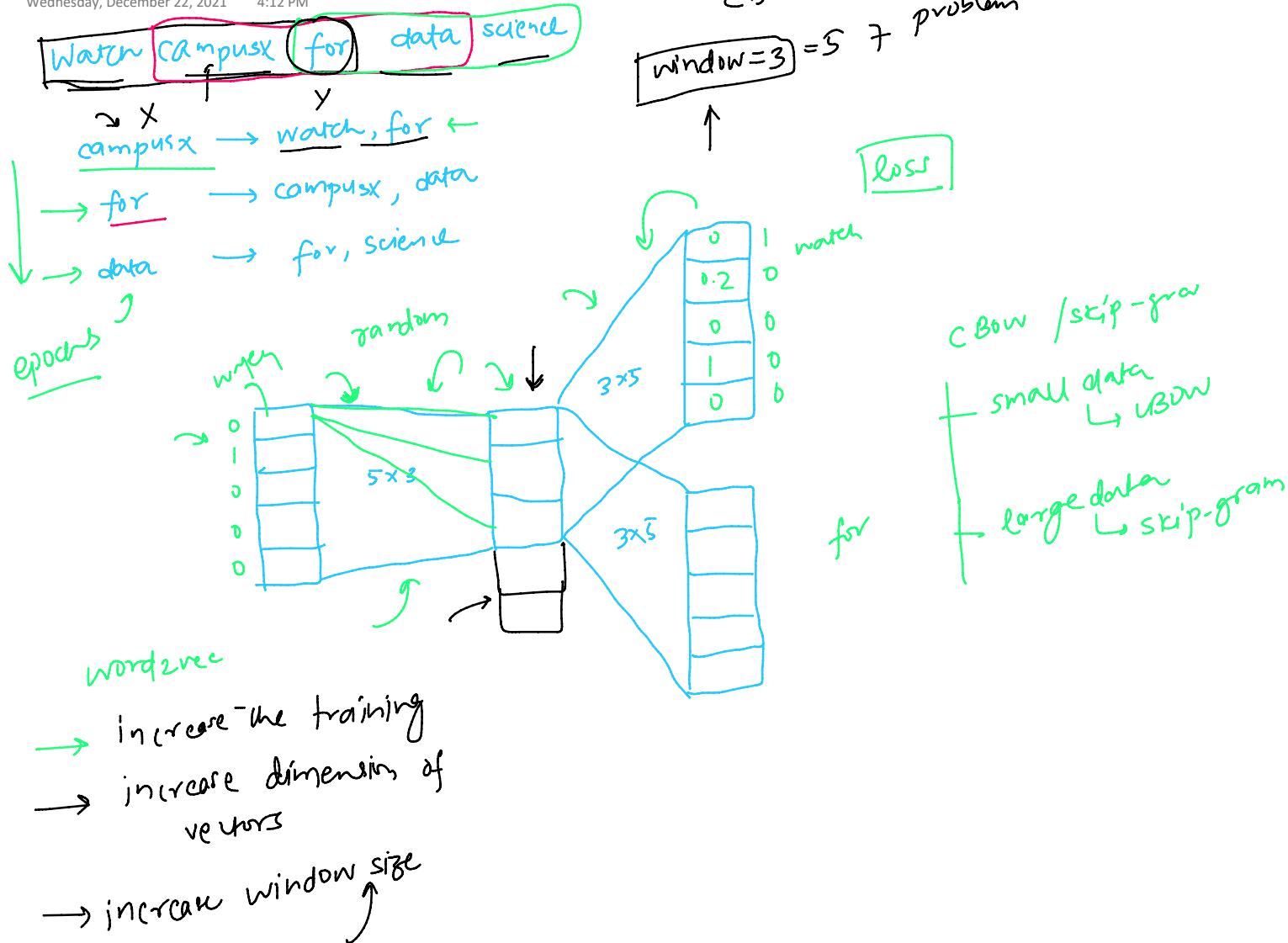
$$\begin{matrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{matrix}$$





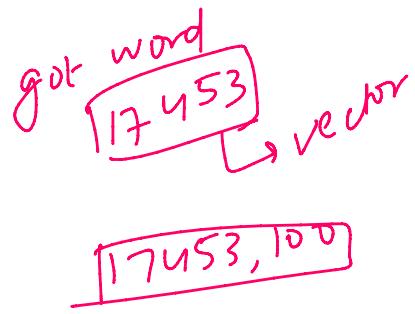
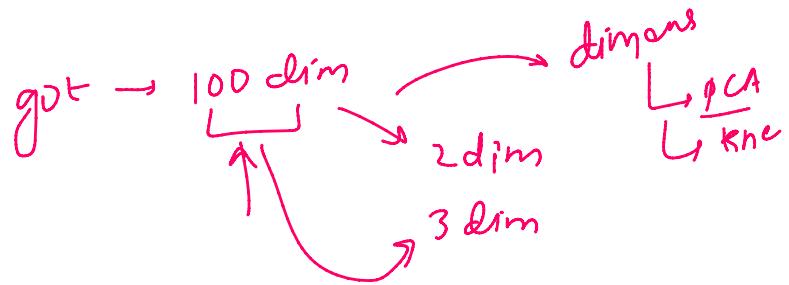
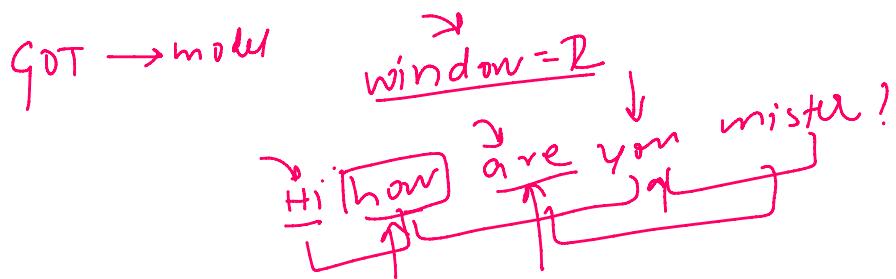
Skip-gram

Wednesday, December 22, 2021 4:12 PM



Training your own model

Wednesday, December 22, 2021 4:12 PM



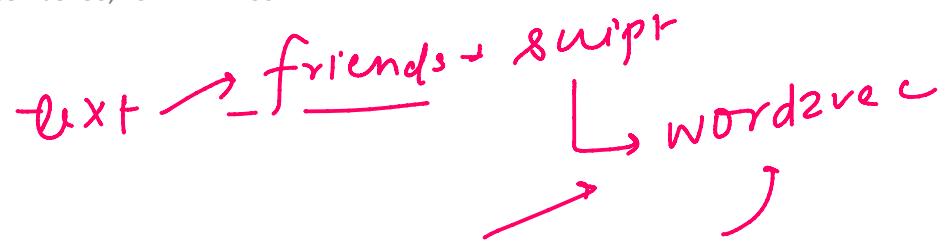
Visualization

Wednesday, December 29, 2021

12:24 PM

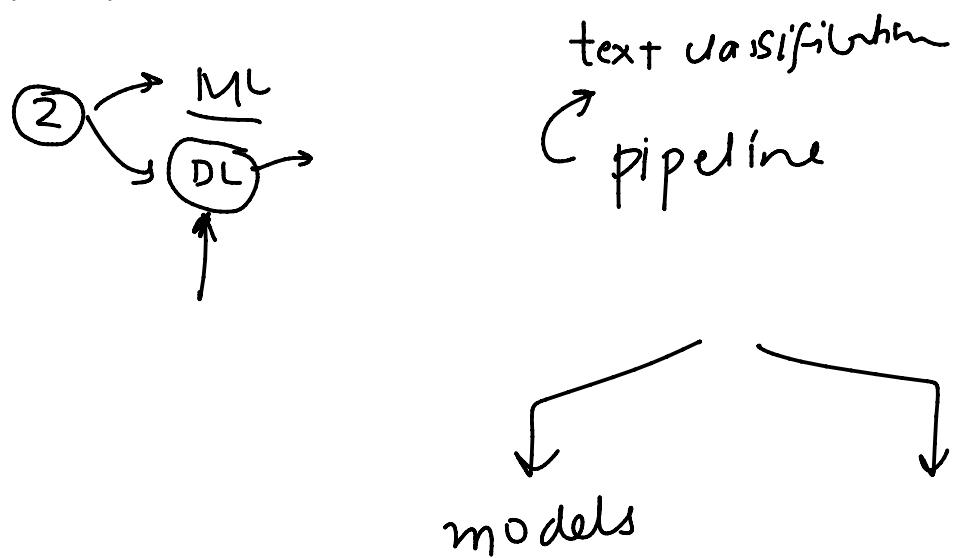
Assignment

Thursday, December 30, 2021 4:00 PM



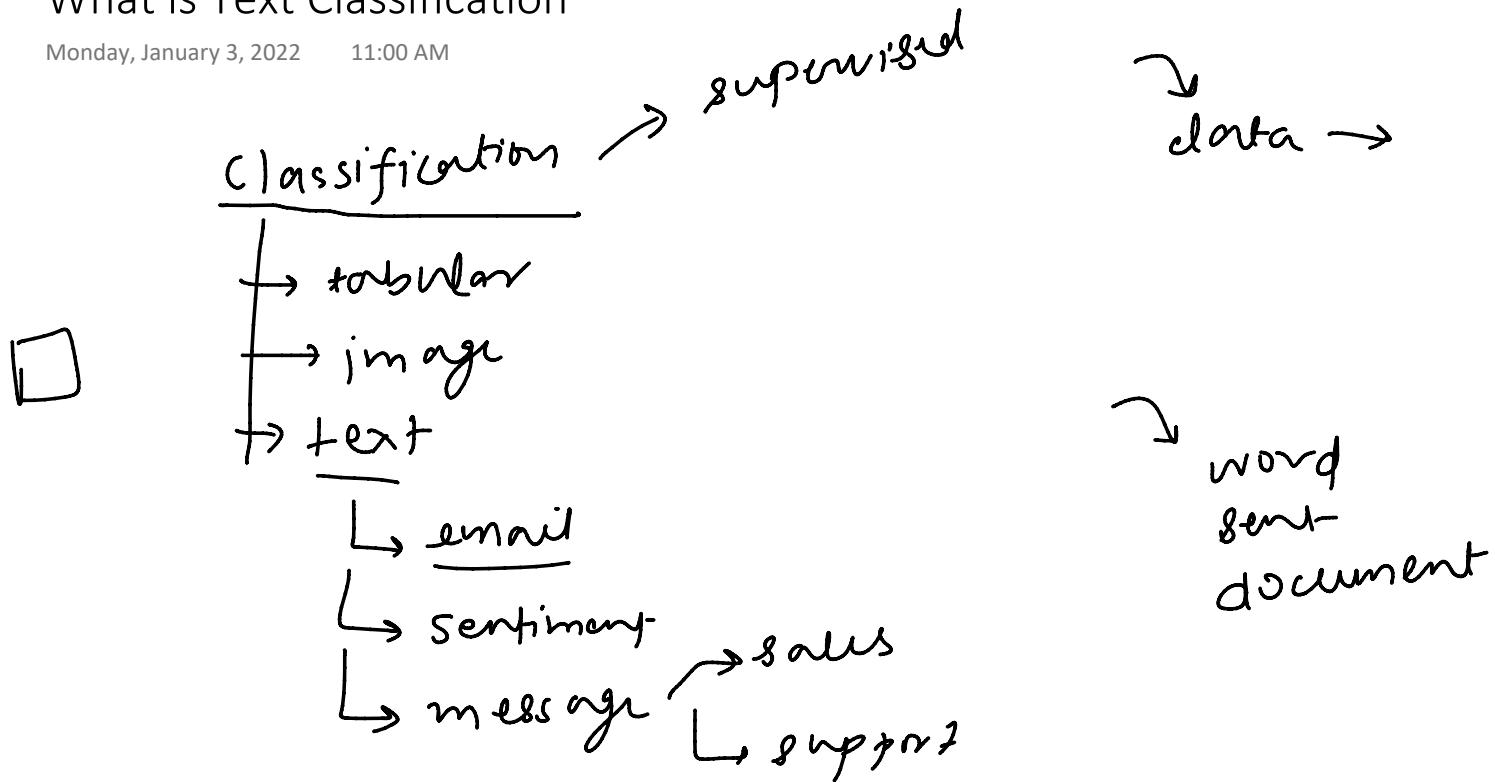
Plan of Attack

Wednesday, January 5, 2022 5:39 PM



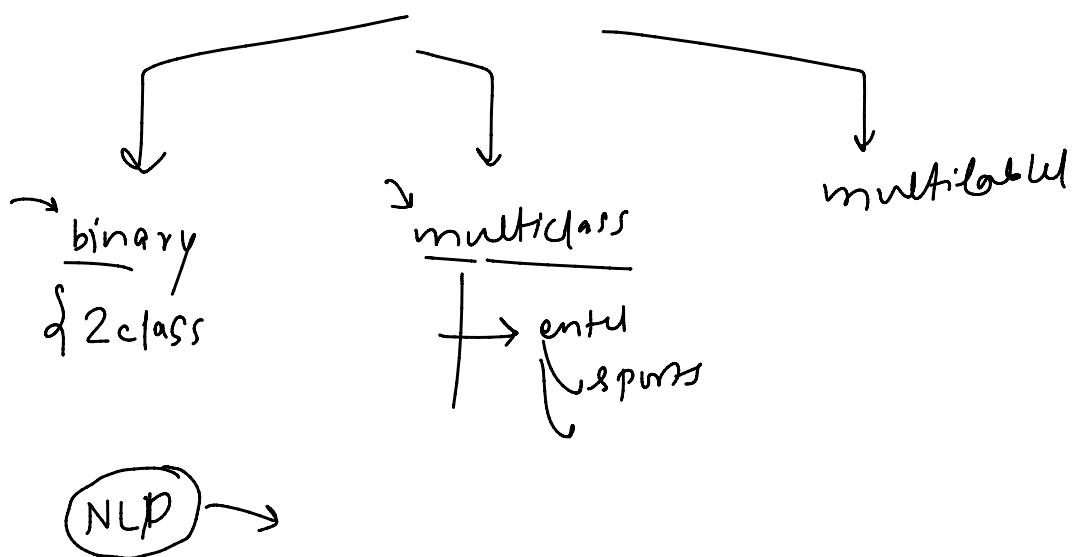
What is Text Classification

Monday, January 3, 2022 11:00 AM



Types of Text Classification

Wednesday, January 5, 2022 2:36 PM



Applications

Wednesday, January 5, 2022 2:40 PM

swiggy

70%

Email
filtering

Customer
support

Sentiment
analysis

Language
detection

Fake news
detection

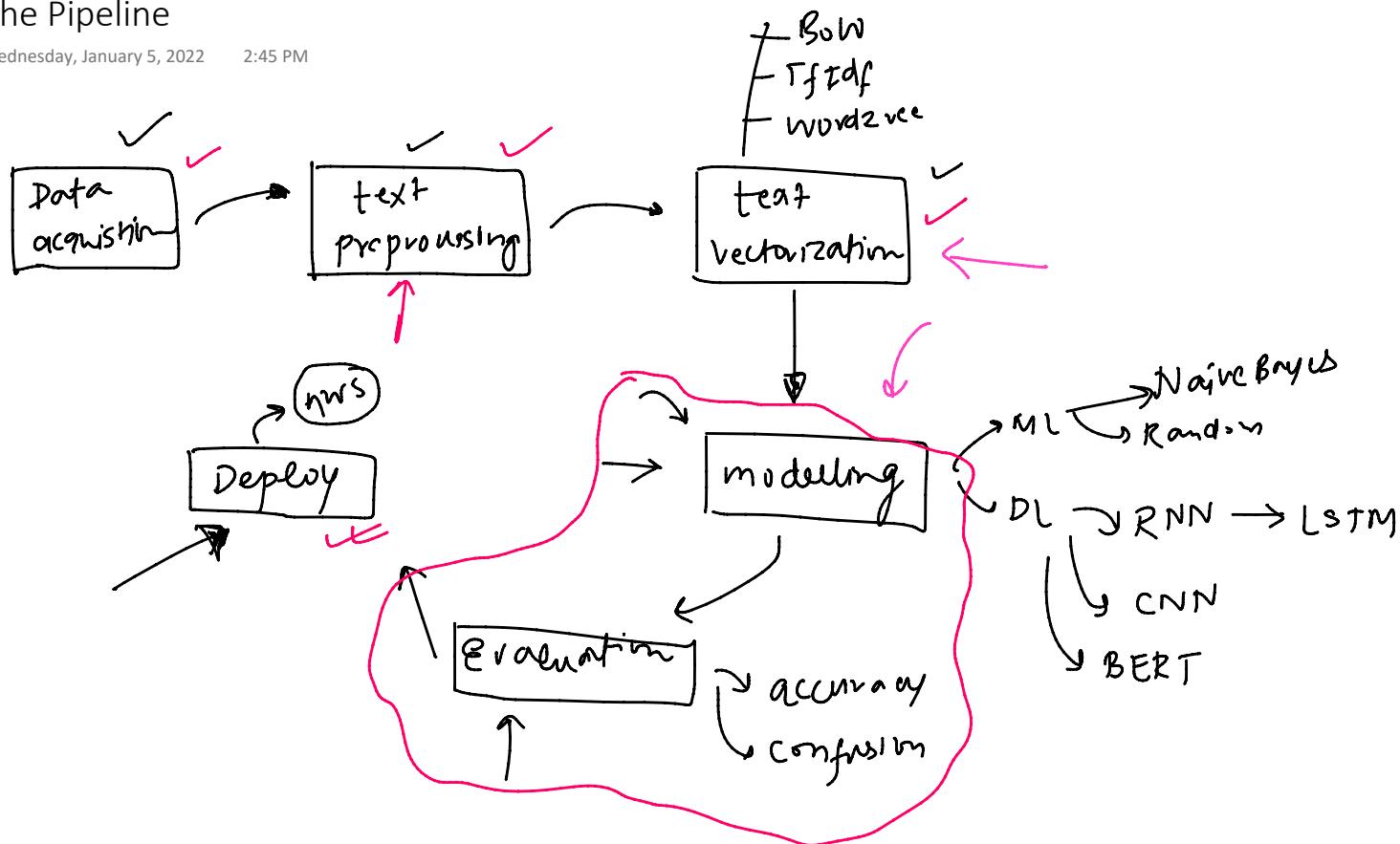
tweet

e-commerce
Review



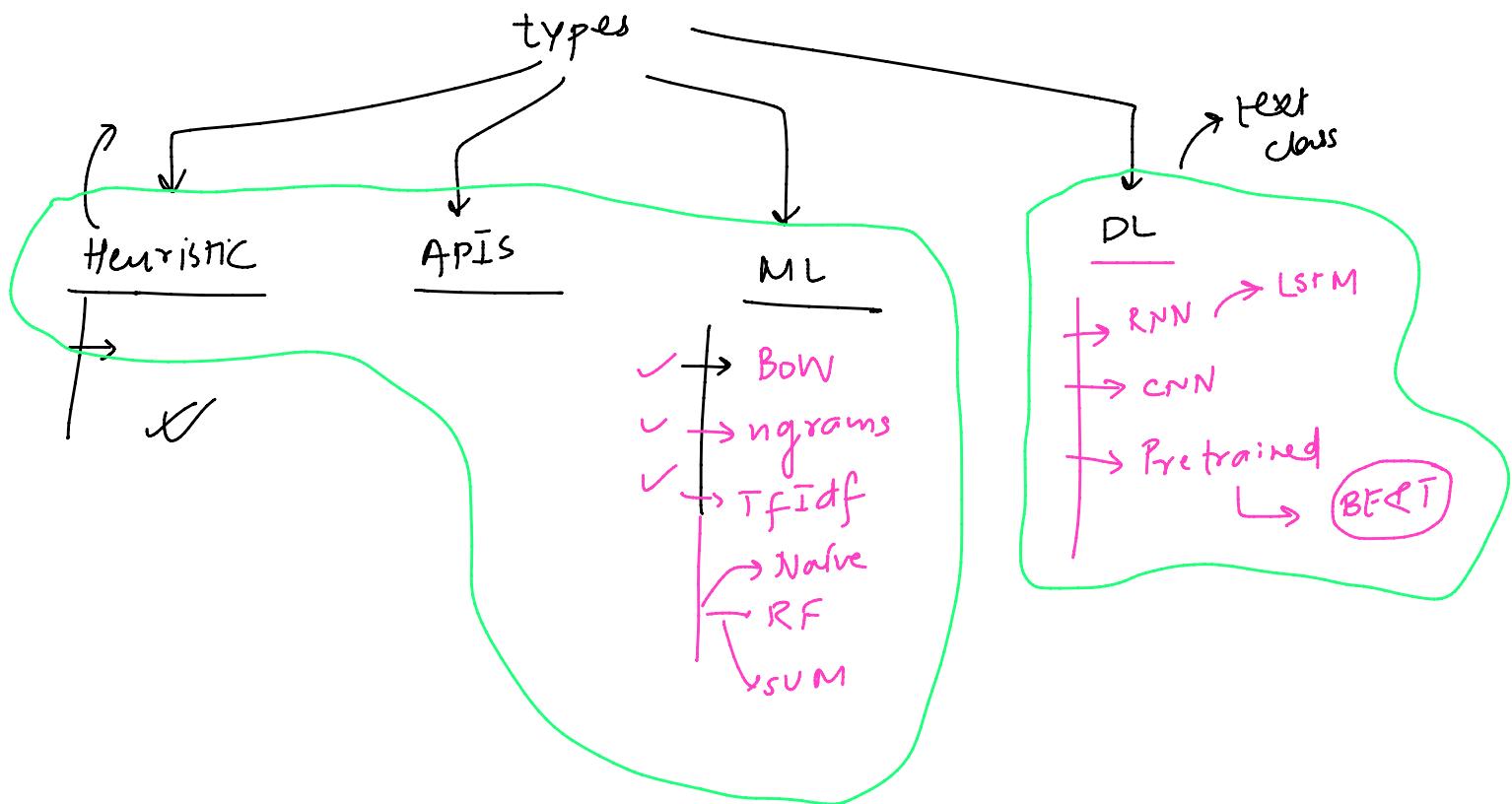
The Pipeline

Wednesday, January 5, 2022 2:45 PM



Different Approaches

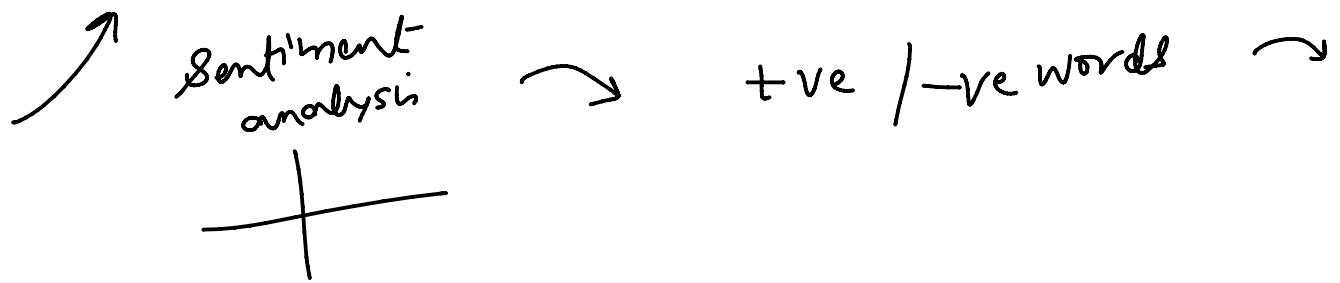
Wednesday, January 5, 2022 2:36 PM



Heuristic Approach

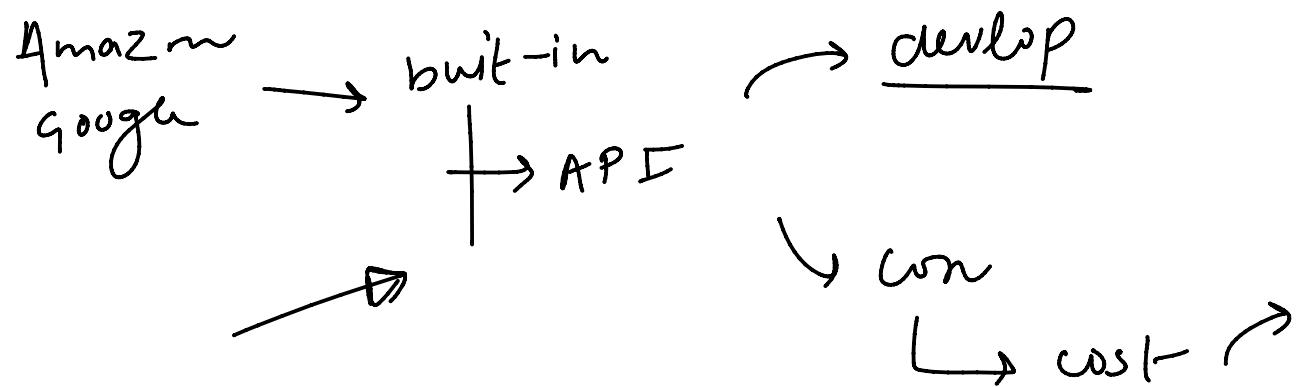
Wednesday, January 5, 2022 5:46 PM

1980's



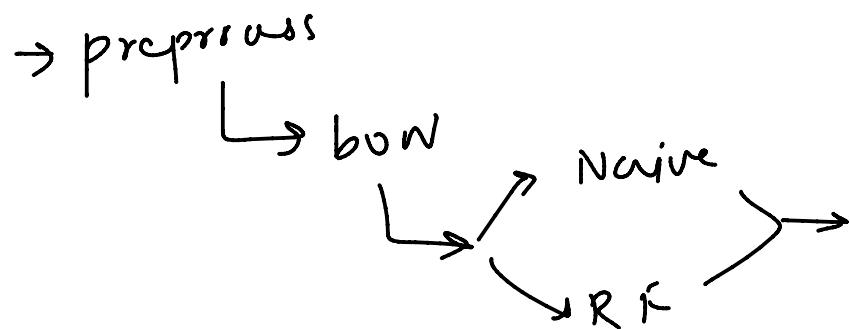
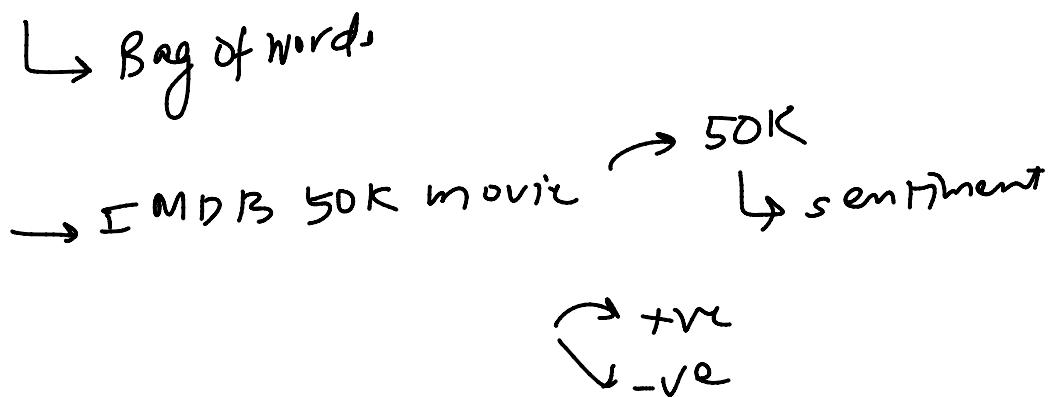
Using API

Wednesday, January 5, 2022 5:46 PM



Using BoW and n-grams

Wednesday, January 5, 2022 3:24 PM

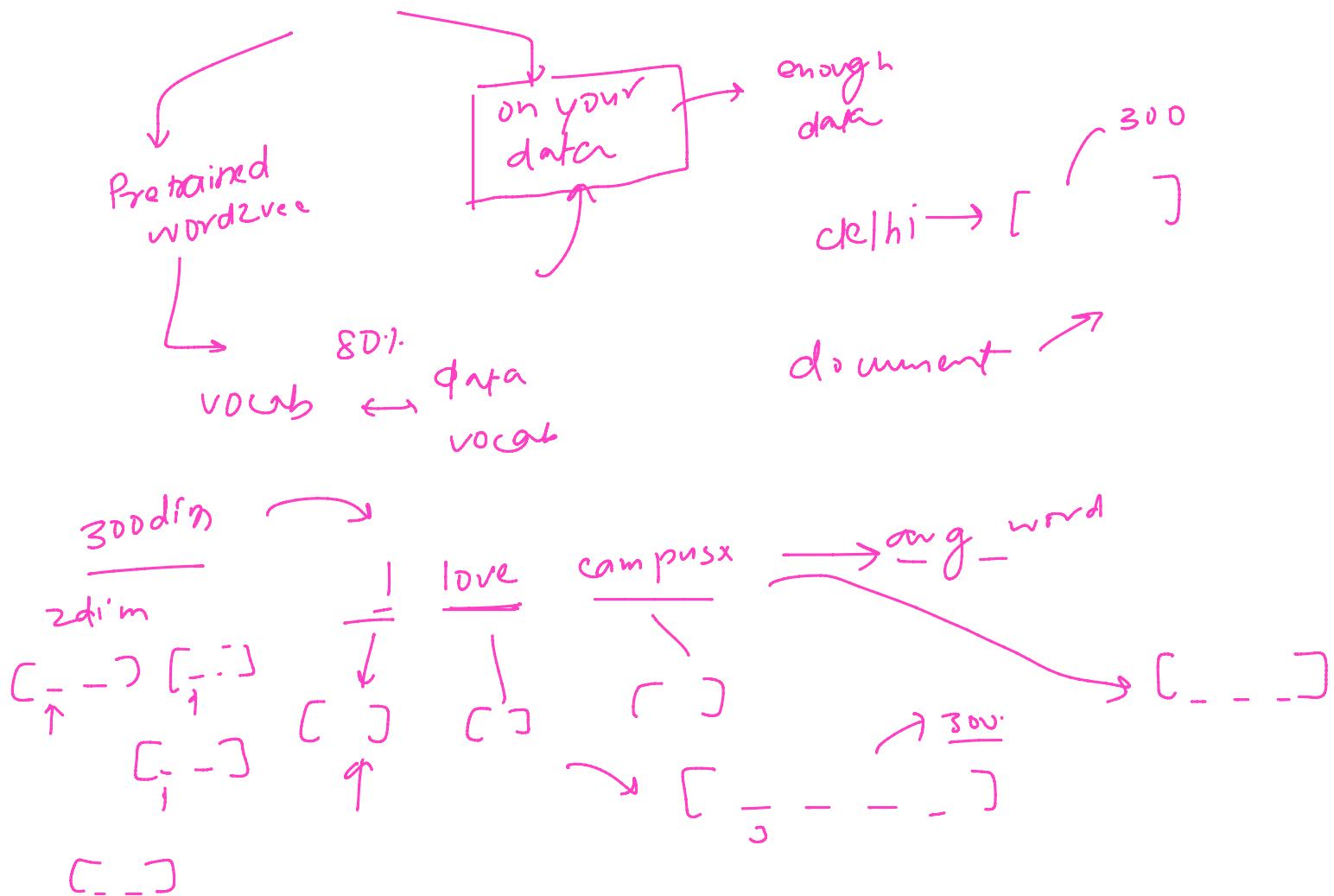


Using Tf-Idf

Wednesday, January 5, 2022 3:24 PM

Using Word2Vec

Wednesday, January 5, 2022 3:24 PM



Interpreting models

Wednesday, January 5, 2022 3:45 PM

Practical Advise

Wednesday, January 5, 2022 3:45 PM

- 1) Ensemble techniques
- 2) Heuristic features
- 3) Deep learning →
- 4) Imbalanced
 - { yes or no }
 - No 70%
- 5) To solve many project

Plan of Attack

Thursday, January 20, 2022 1:44 PM

- { 1) What POS tagging
- 2) Applications
- 3) Spacy → code demo
- 4) HMM → Viterbi }

What is Parts of Speech Tagging?

Thursday, January 20, 2022 1:44 PM

In simple words, we can say that POS tagging is a task of labelling each word in a sentence with its appropriate part of speech.

In traditional grammar, a part of speech or part-of-speech is a category of words that have similar grammatical properties.

Noun
pron

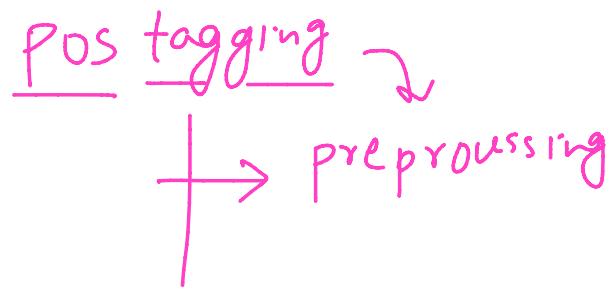
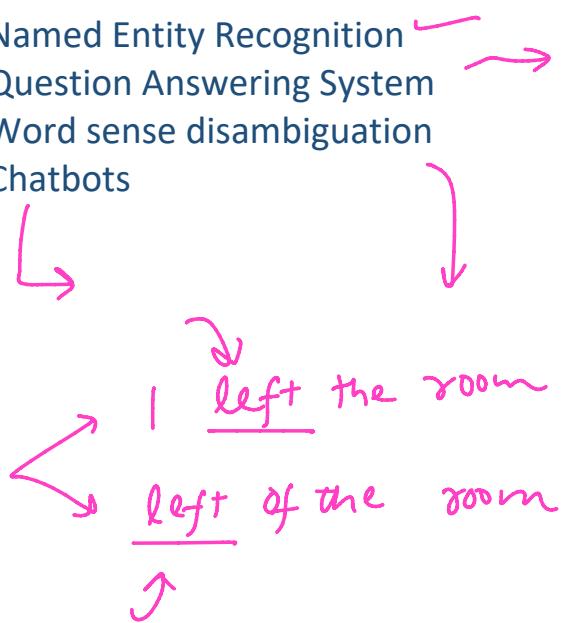
Why not tell someone ?
adverb adverb verb noun punctuation mark,
 sentence closer

Application

Applications of POS Tagging

Thursday, January 20, 2022 1:49 PM

1. Named Entity Recognition
2. Question Answering System
3. Word sense disambiguation
4. Chatbots



Code Demo

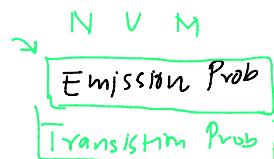
Thursday, January 20, 2022 1:49 PM

How POS Tagging works?

Thursday, January 20, 2022 1:50 PM

HMM →

[will will google campusx]



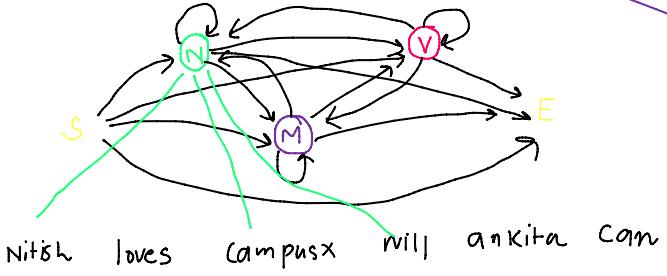
S N Nitish V loves N campusx E

S M can N Nitish V google N campusx E

S M will N Ankita V google N campusx E

S N Ankita V loves N will E

S N will V loves N google E



$\boxed{3^5 = 81}$ → exponential



→ Emission

	N	M	V	E
Nitish	2/10	0	0	
loves	0	0	3/15	
campusx	3/10	0	0	
google	1/10	0	2/15	
will	2/10	1/2	0	
Ankita	2/10	0	0	
can	0	1/2	0	

→ Transition

	N	M	V	E
S	3/5	2/5	0	0
N	1/4	1/4	5/10	5/10
M	2/2	0	0	0
V	5/5	0	0	0



$$\frac{1}{5} \times \frac{1}{5} = 0$$

$$S \rightarrow N \rightarrow N \rightarrow N \rightarrow N \rightarrow M \rightarrow E = 0$$

$$S \rightarrow M \rightarrow N \rightarrow N \rightarrow V \rightarrow N \rightarrow E \neq 0$$

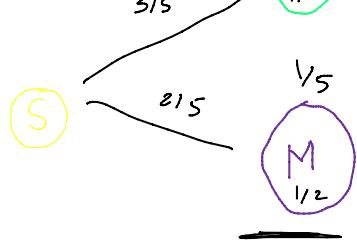
Optimization → Viterbi Algo

Viterbi Algorithm

$$\frac{3}{25} \times \frac{1}{4} \times \frac{1}{5} = \frac{3}{500}$$

$$\frac{1}{5} \times 1 \times \frac{1}{5} = \frac{1}{25}$$

$$\frac{3}{25} \times \frac{1}{4} \times \frac{1}{2} = \frac{3}{200}$$



will
3/25

3/500

will
3/25

3/500

$$\frac{3}{500} \times \frac{1}{4} \times \frac{1}{10} = \frac{3}{2000}$$

V
0

V
0

V
2/5

$$\frac{3}{2500} \times \frac{1}{4} \times \frac{3}{10} = \frac{9}{80000}$$

V
0

$$\frac{3}{2500} \times 1 \times \frac{3}{10} = \frac{9}{25000}$$

$$\frac{3}{500} \times \frac{1}{4} \times \frac{1}{10} = \frac{3}{2000}$$

(V)
0

(V)
0

(V)
215

(V)
0

$$= \frac{9}{25000}$$

$$\frac{3}{200} \times 1 \times \frac{1}{10} = \frac{3}{2000}$$

$$\frac{3}{200} \times \frac{1}{x} \times \frac{2}{5} = \frac{3}{2500}$$

19 December 2023 14:43

