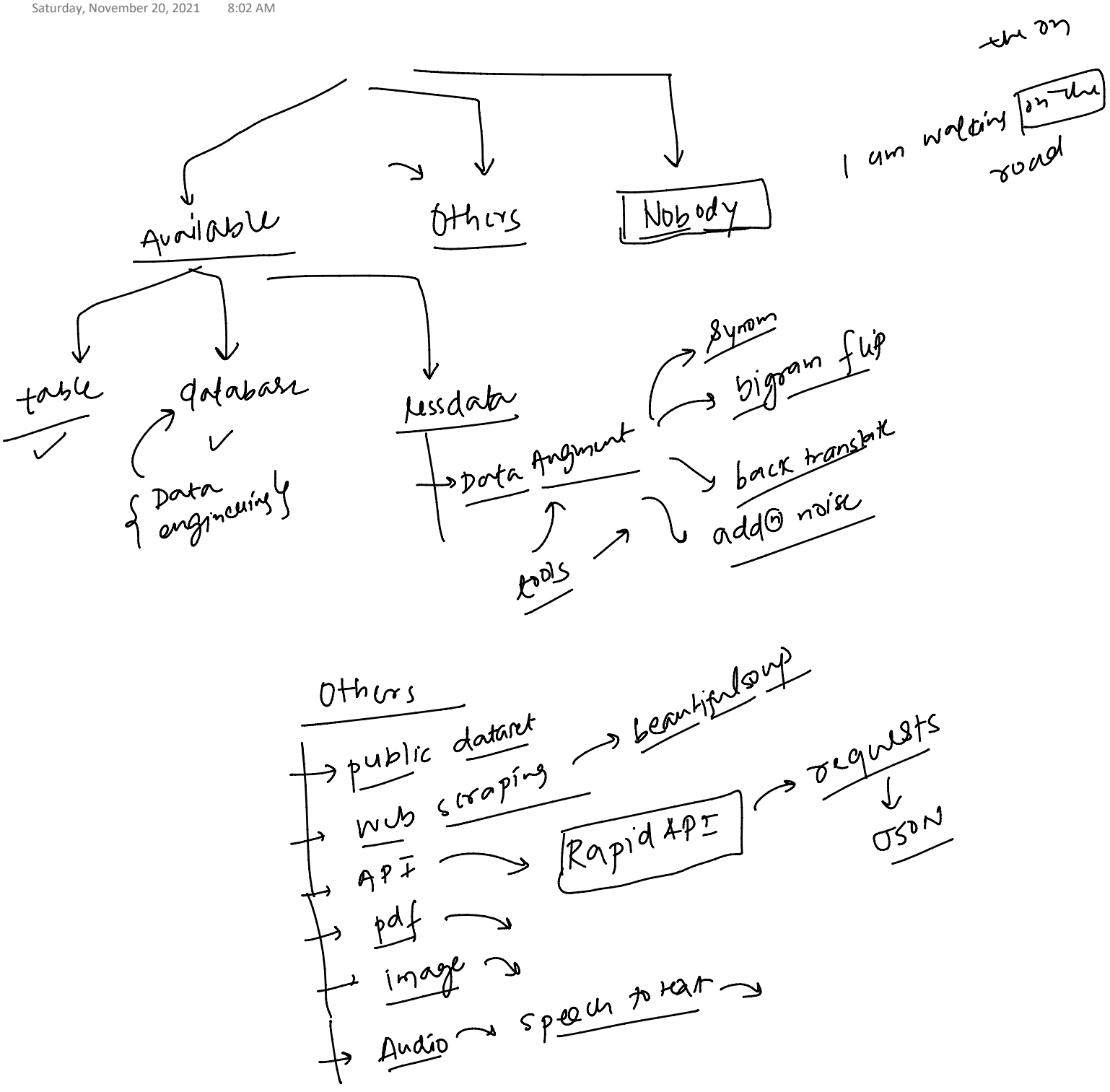
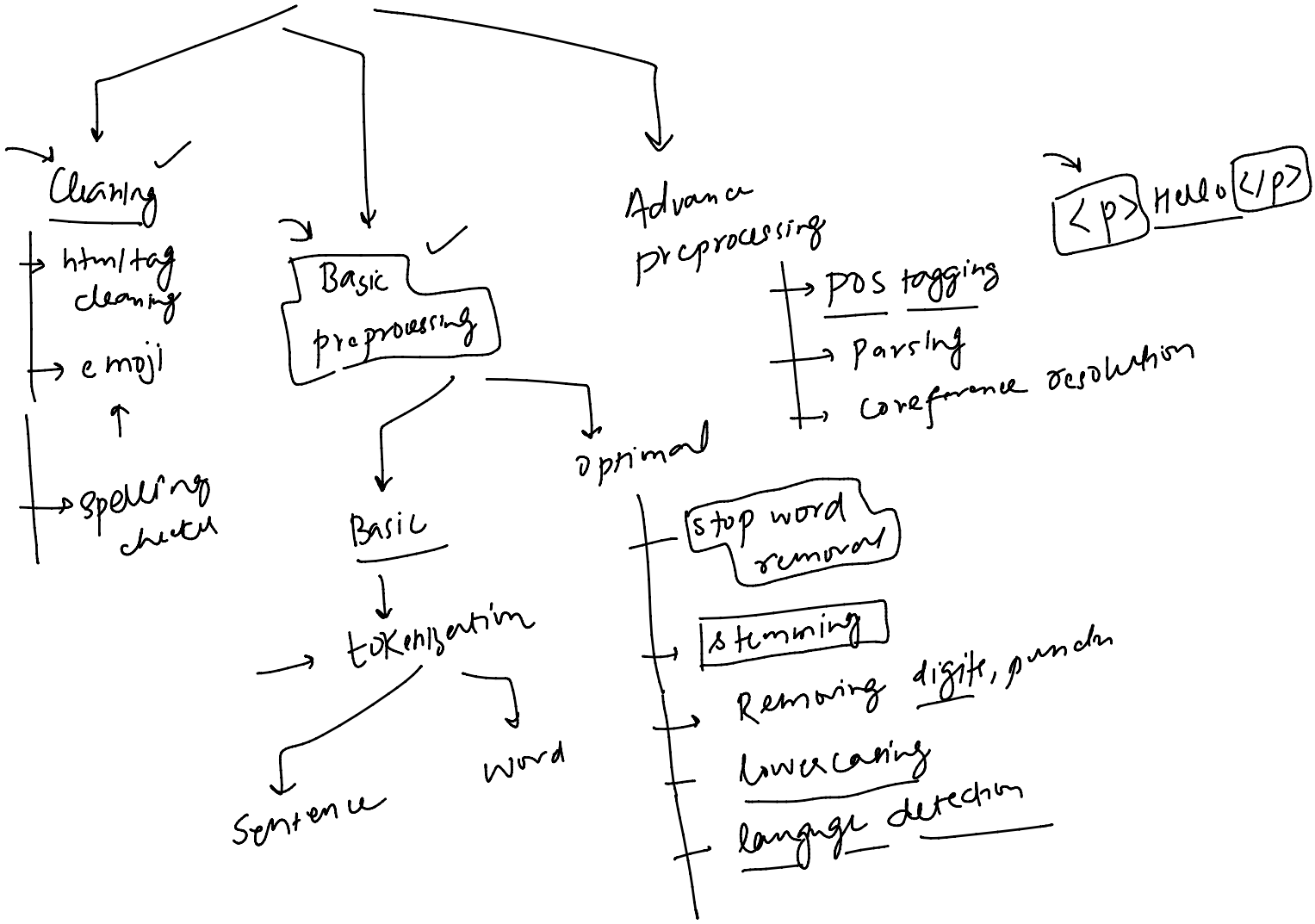
Data Acquisition

Saturday, November 20, 2021 8:02 AM



Text Preparation

Saturday, November 27, 2021 12:47 PM



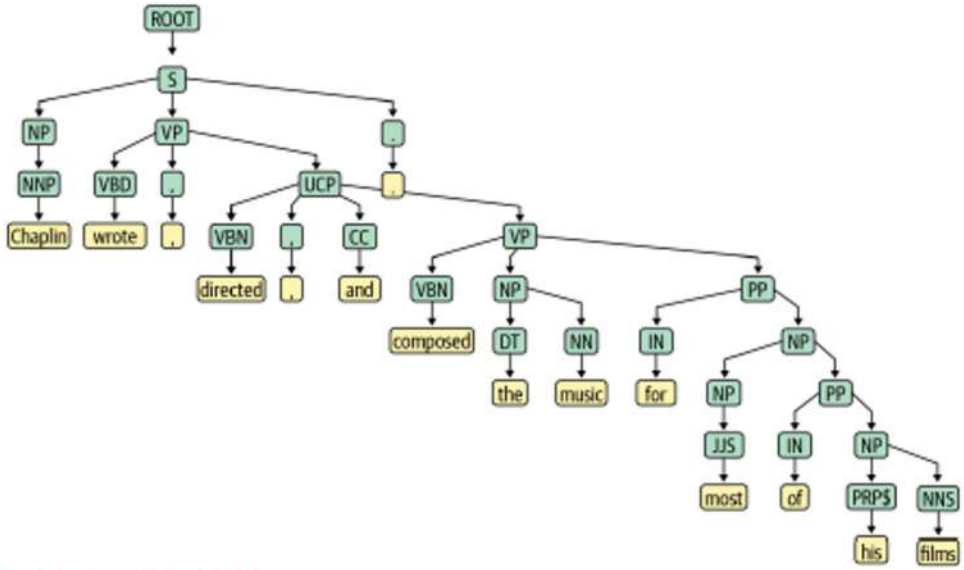
Input

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

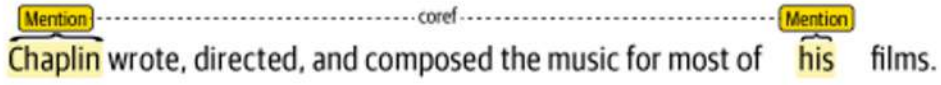
Tokenization with Lemmatization

POS Tagging

Parse Tree

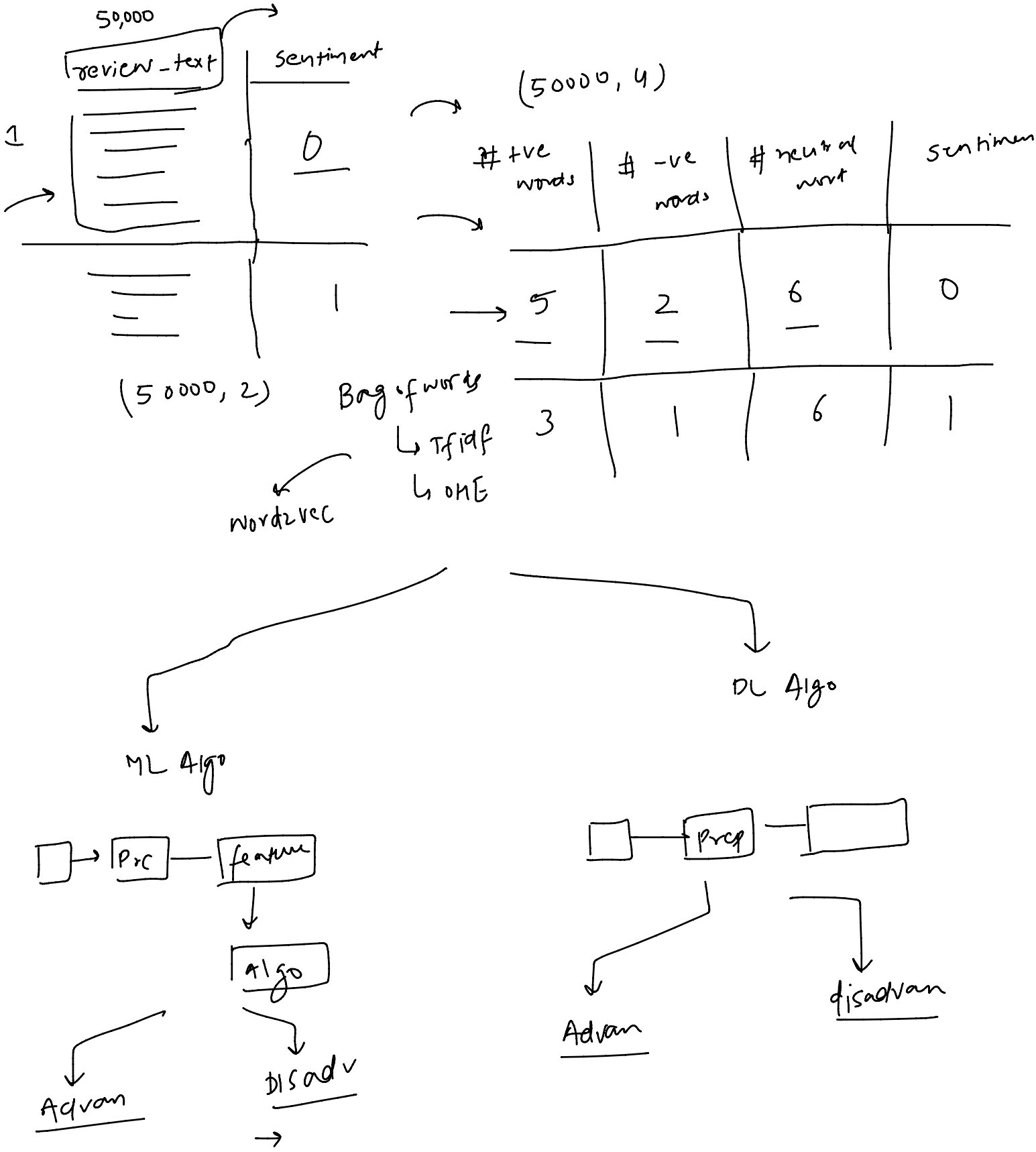


Coreference Resolution



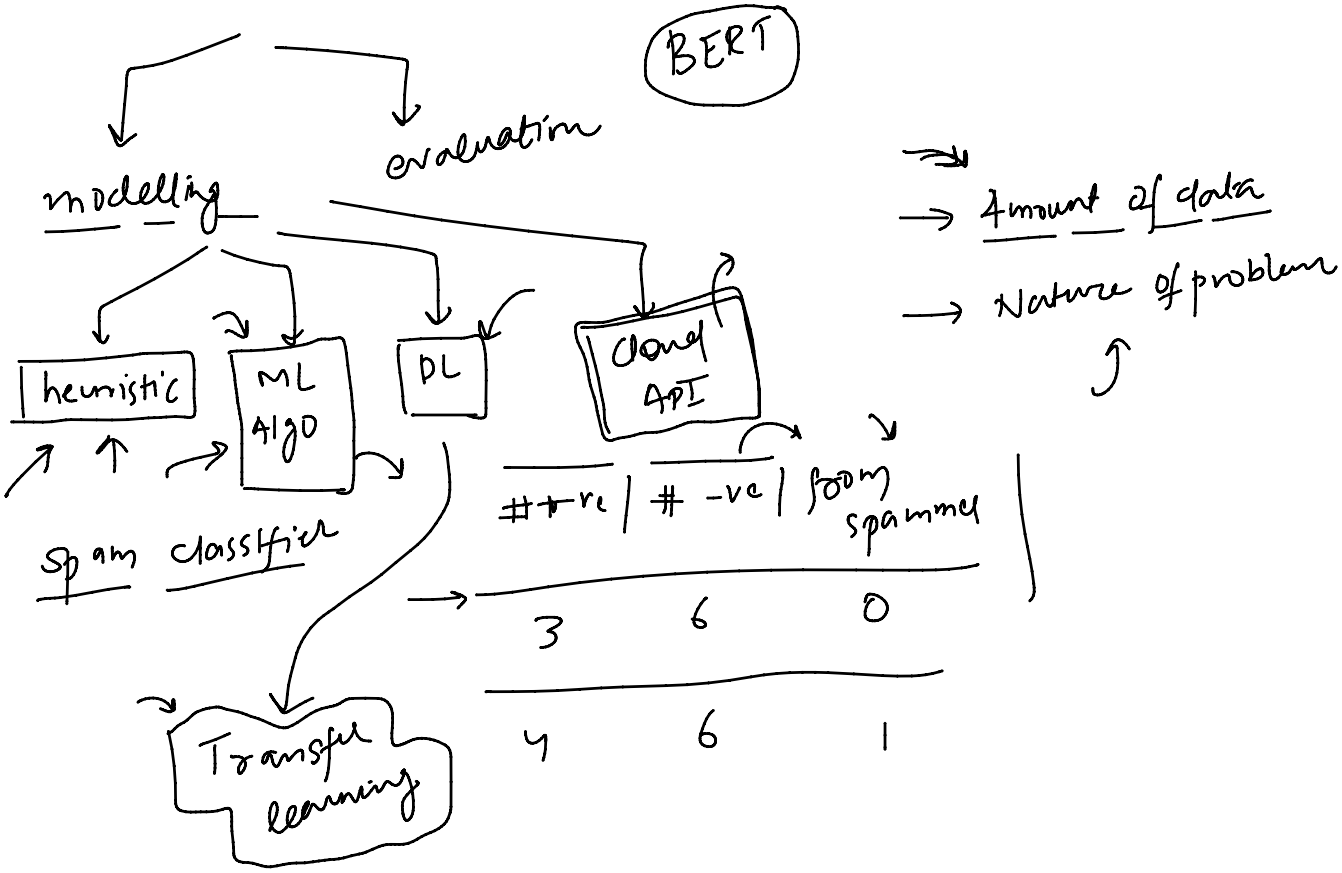
Feature Engineering

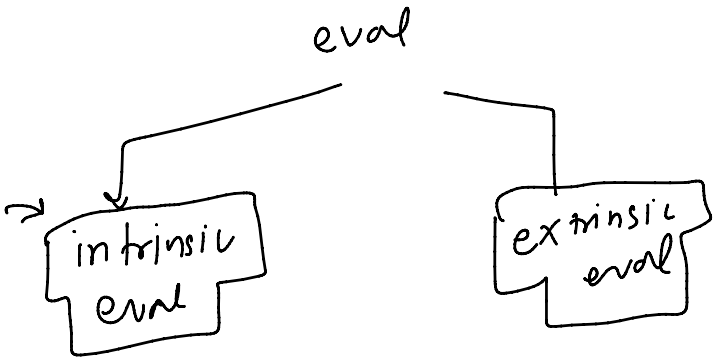
Tuesday, November 30, 2021 10:44 AM



Modelling

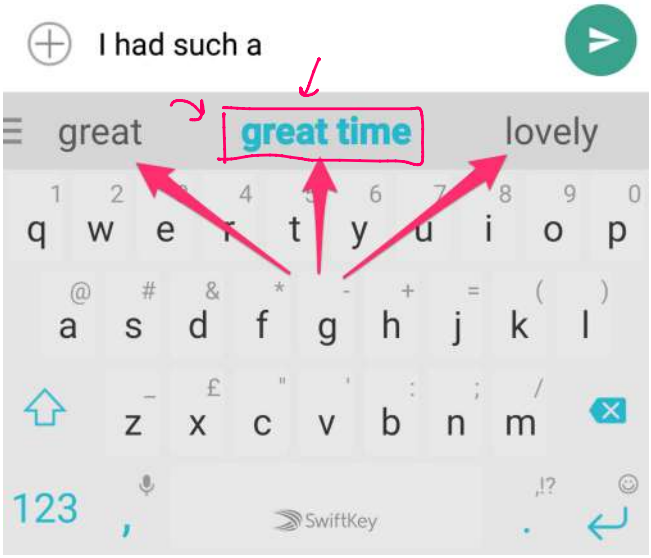
Tuesday, November 30, 2021 11:14 AM





text gen

SOS



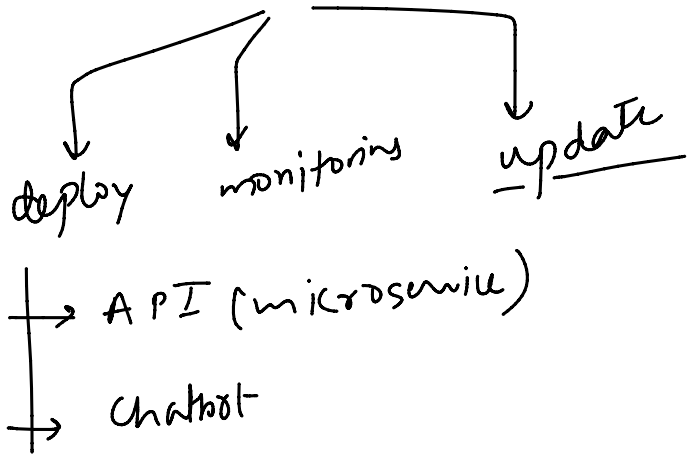
perplexity → NLP

extrinsic

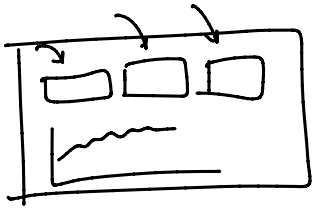
1000 → 0

Deployment

Tuesday, November 30, 2021 12:50 PM

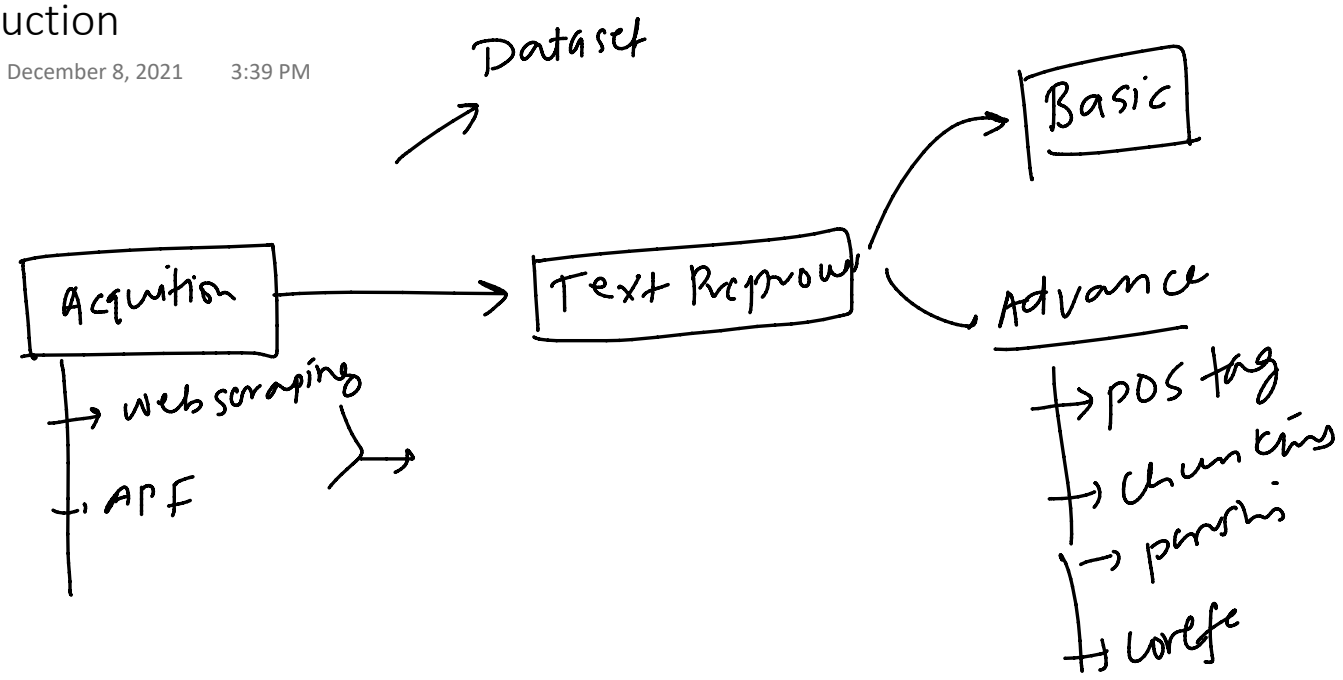


monitoring



Introduction

Wednesday, December 8, 2021 3:39 PM



Lowercasing

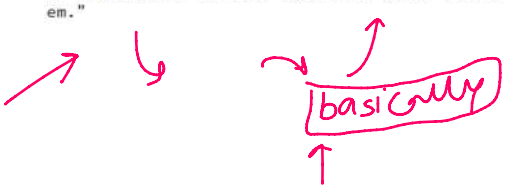
Wednesday, December 8, 2021 3:40 PM

"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."

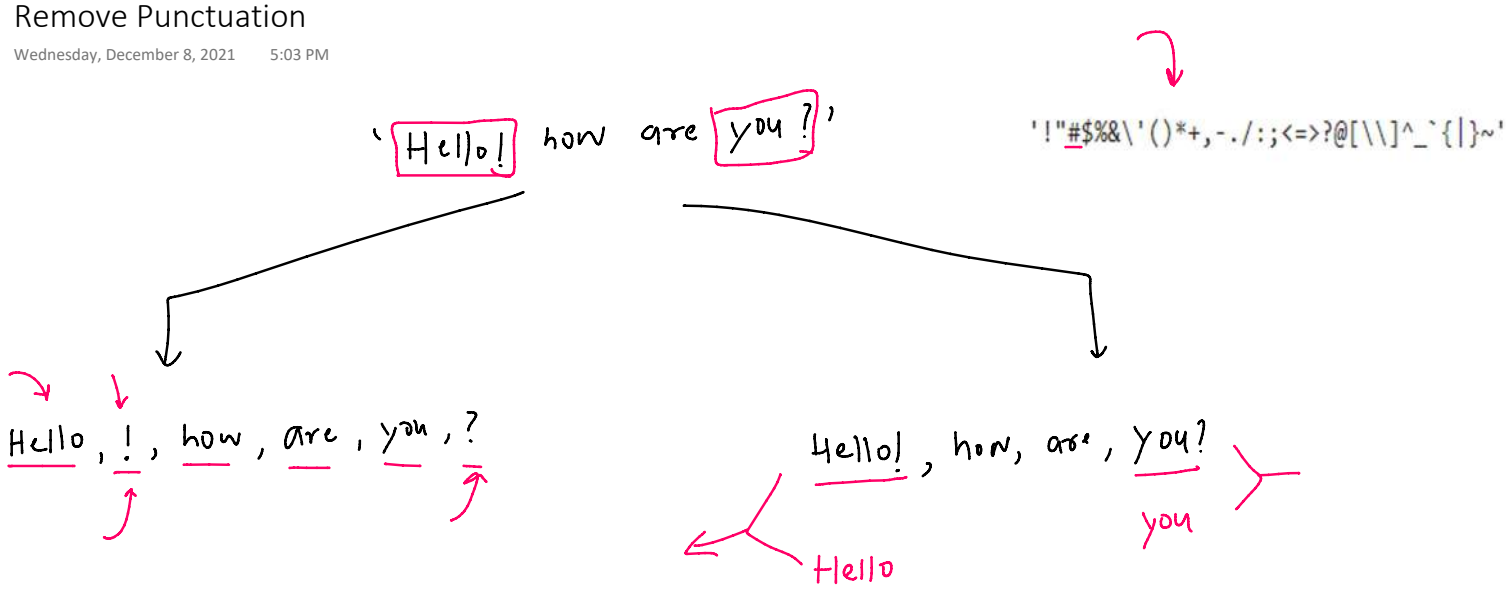
Remove URLs

Wednesday, December 8, 2021 4:23 PM

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

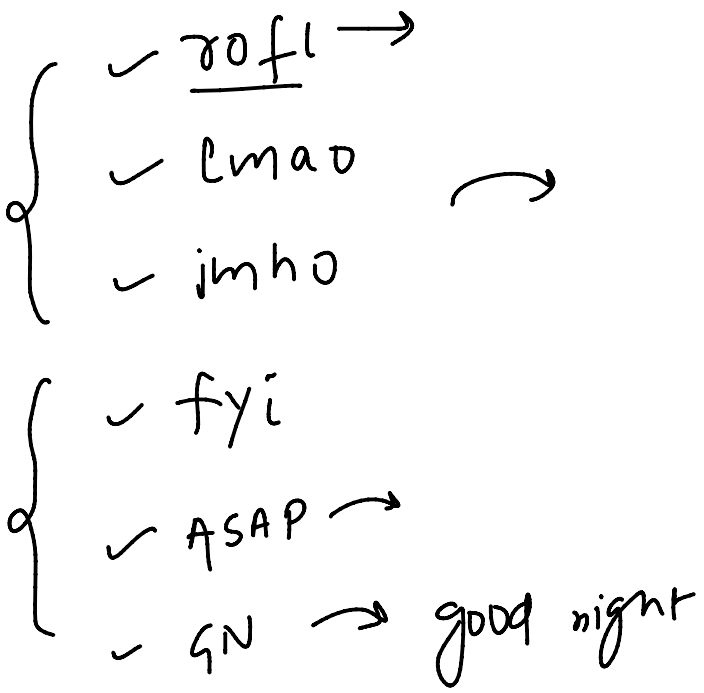
Remove Punctuation

Wednesday, December 8, 2021 5:03 PM



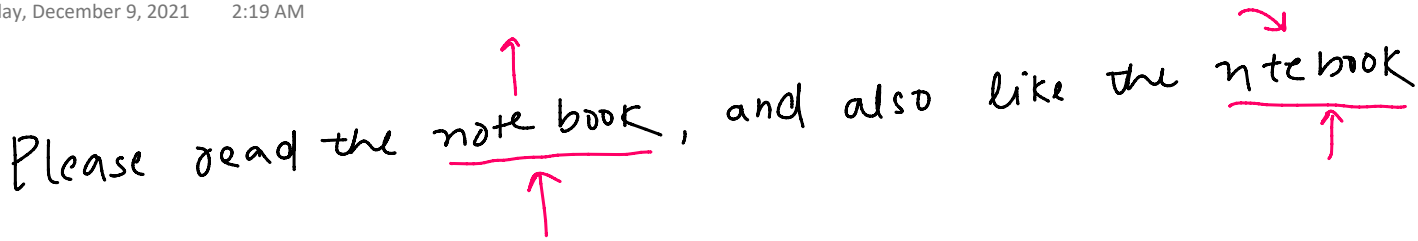
Chat word treatment

Thursday, December 9, 2021 1:54 AM



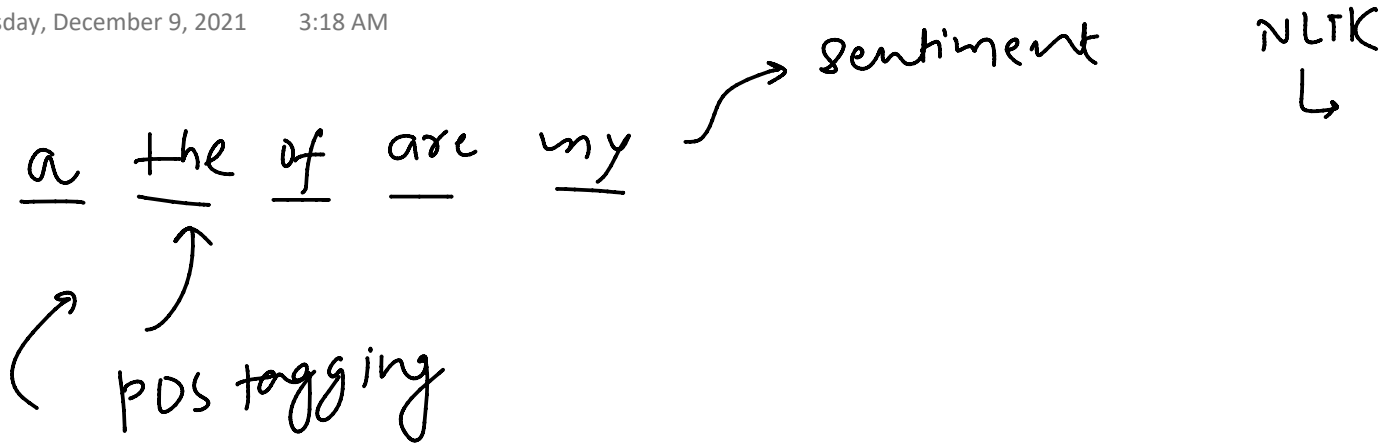
Spelling Correction

Thursday, December 9, 2021 2:19 AM



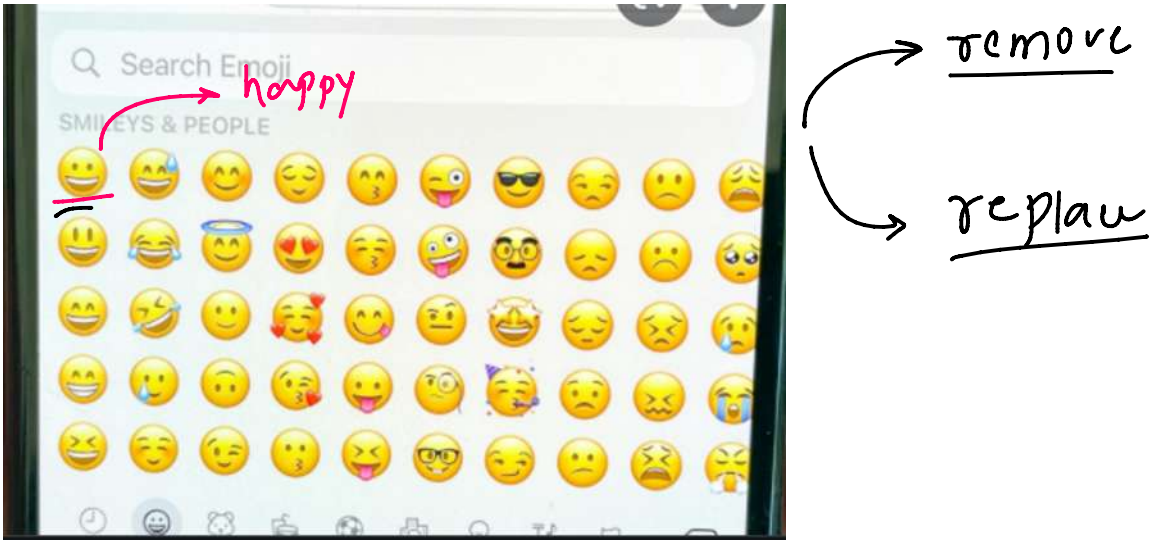
Removing Stop words

Thursday, December 9, 2021 3:18 AM



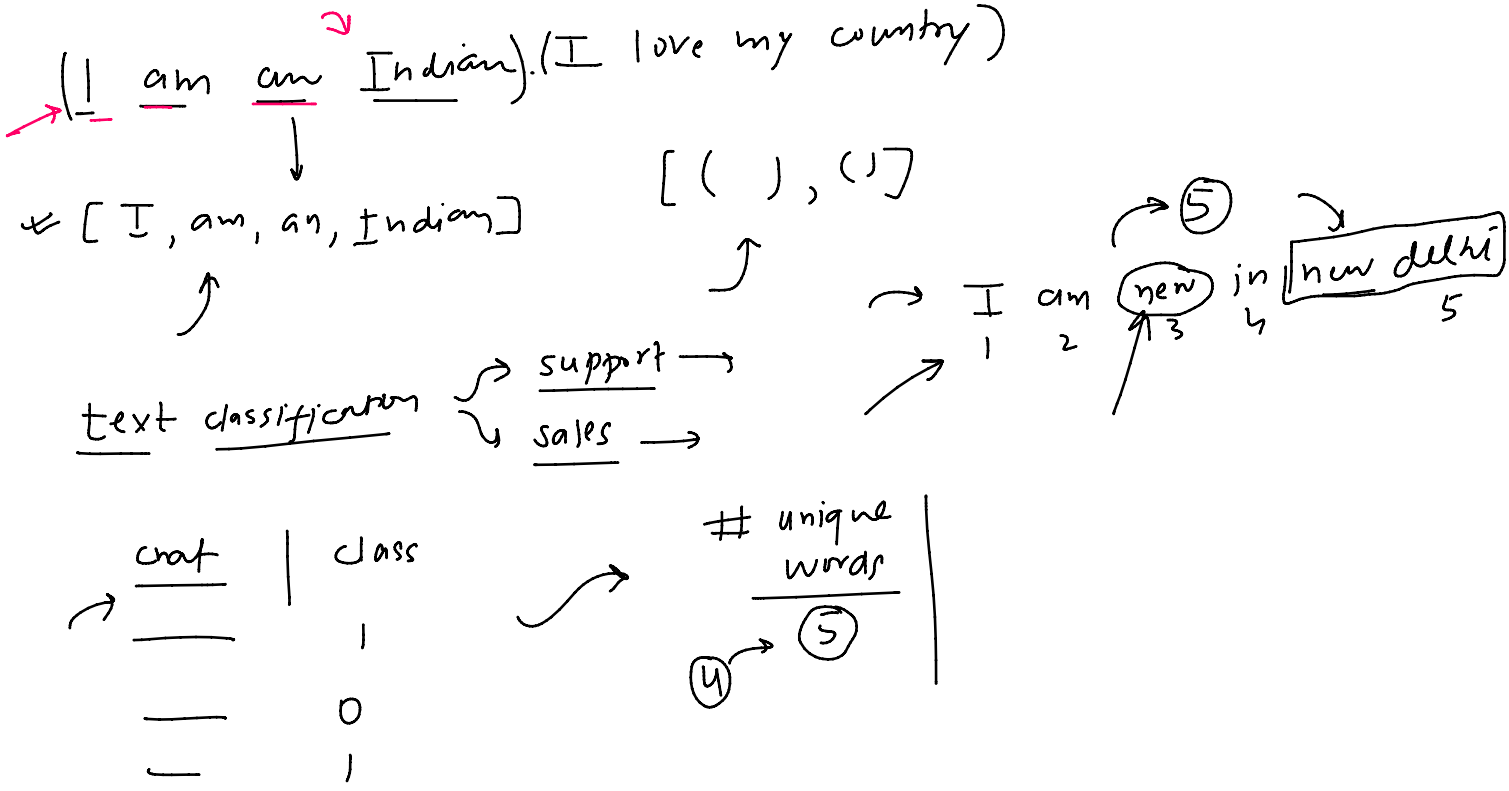
Handling Emojis

Thursday, December 9, 2021 12:07 PM

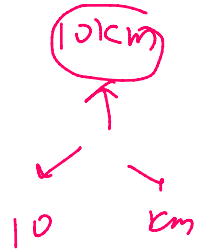


Tokenization

Friday, December 10, 2021 11:01 AM



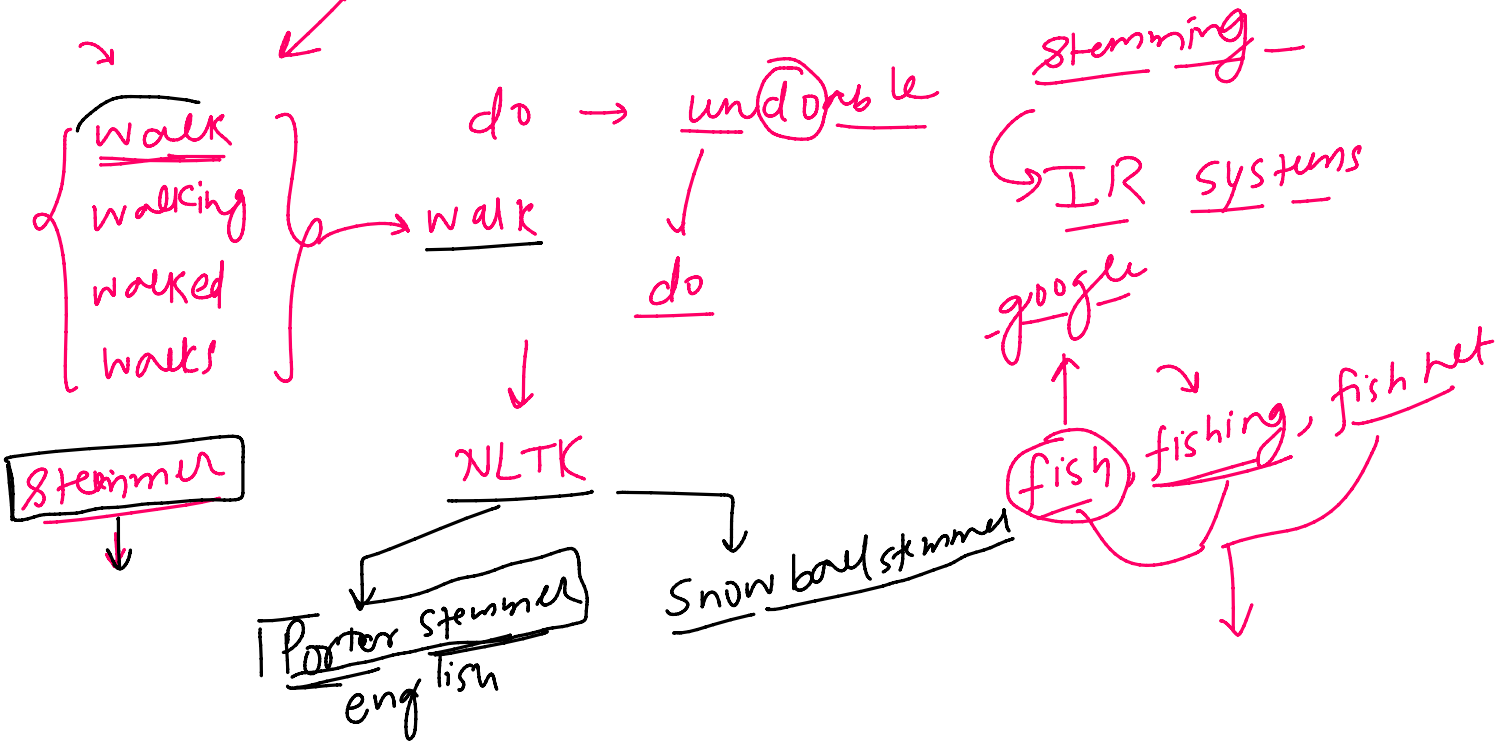
- **Prefix:** Character(s) at the beginning — $ (“ ¿
- **Suffix:** Character(s) at the end — km) , . ! ”
- **Infix:** Character(s) in between — - -- / ...
- **Exception:** Special-case rule to split a string into several tokens or prevent a token from being split when punctuation rules are applied — let’s U.S.



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."

→ ...atization

→ 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.*

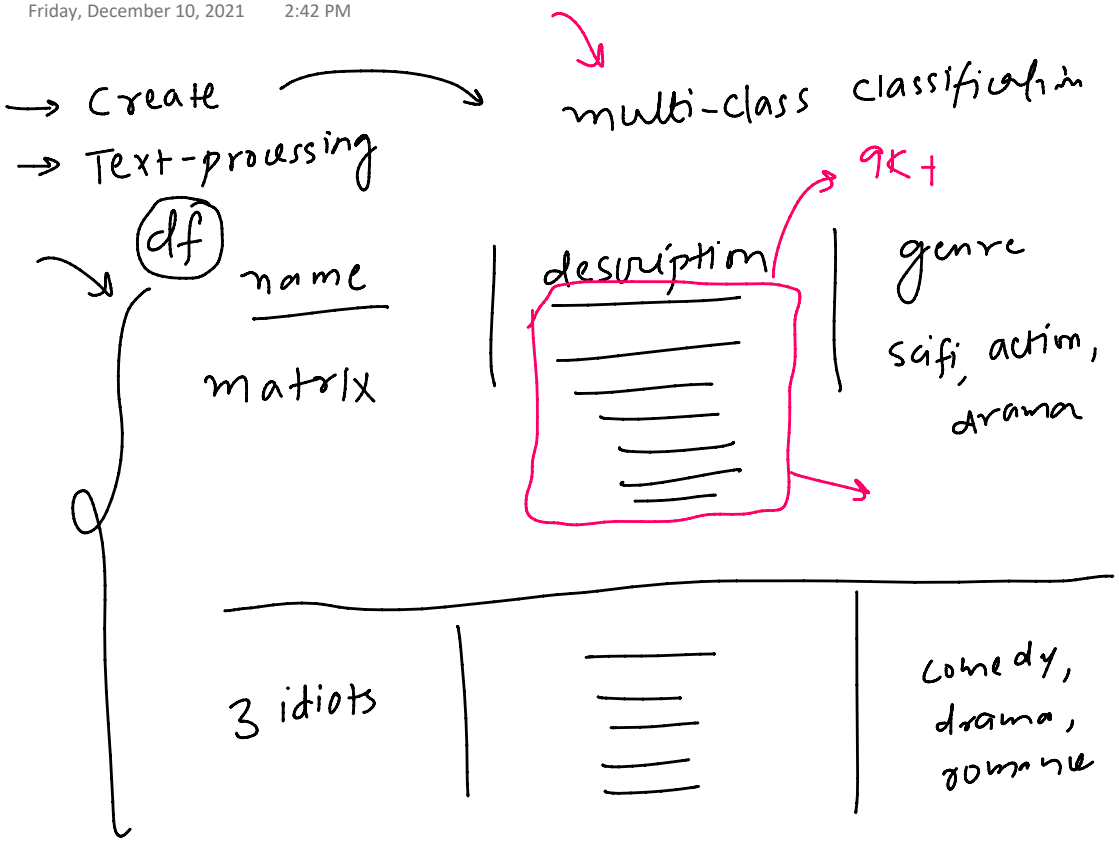
lemma

algorithm
suffix
fast

WordNet Lemmatizer → slow
lexical db

Assignment

Friday, December 10, 2021 2:42 PM



Imdb API

Plan of Attack

Monday, December 13, 2021 11:05 AM

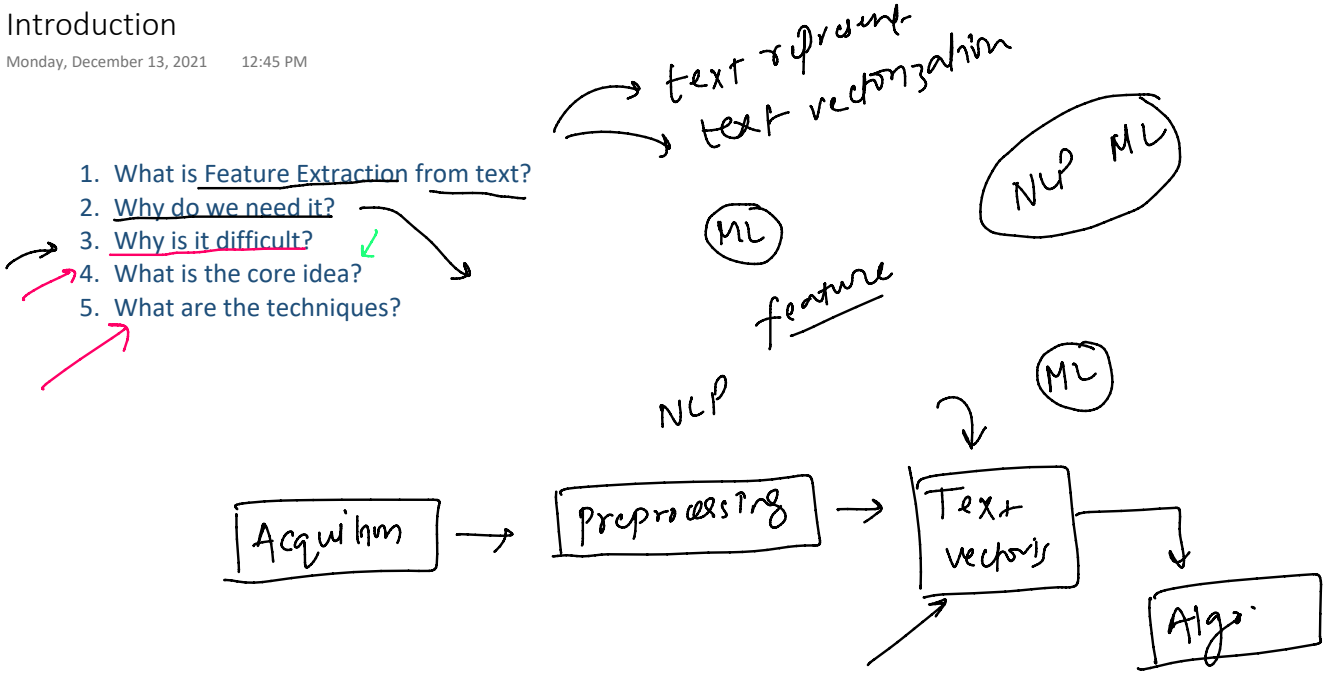
Intro → BoW
→ TfIdf

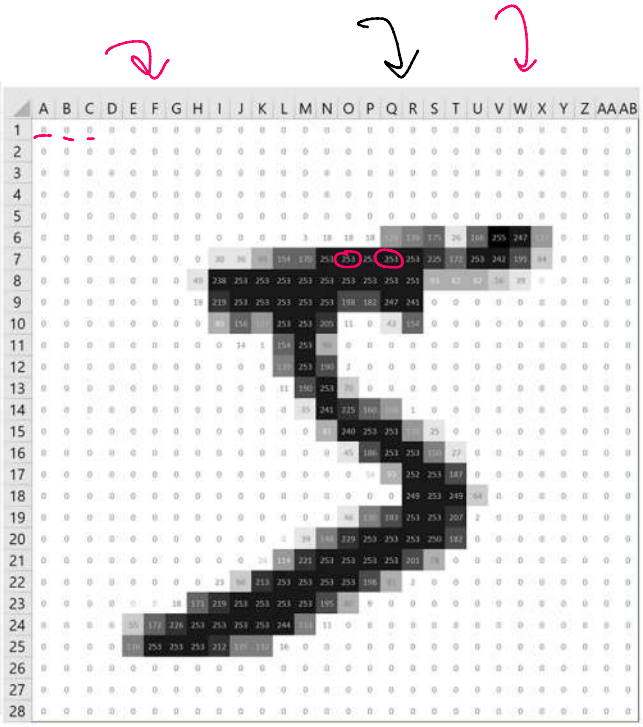
Assignment

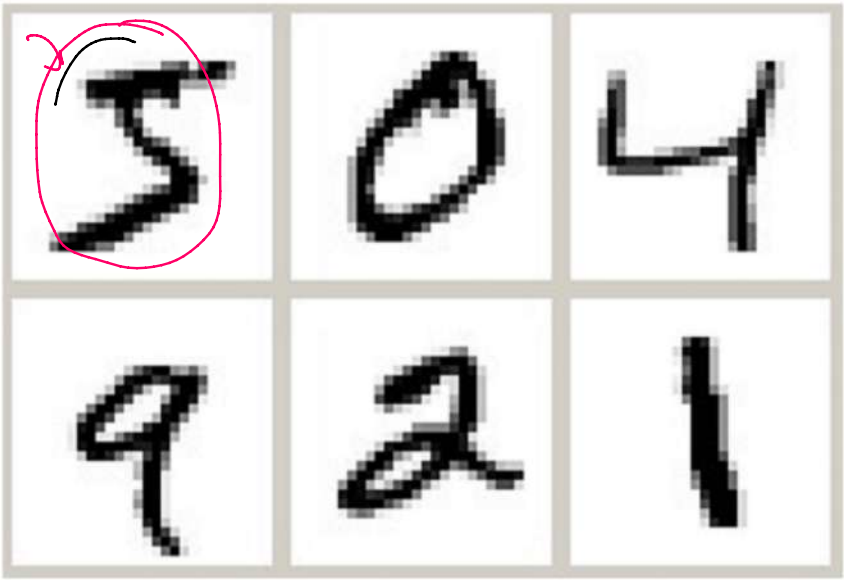
Introduction

Monday, December 13, 2021 12:45 PM

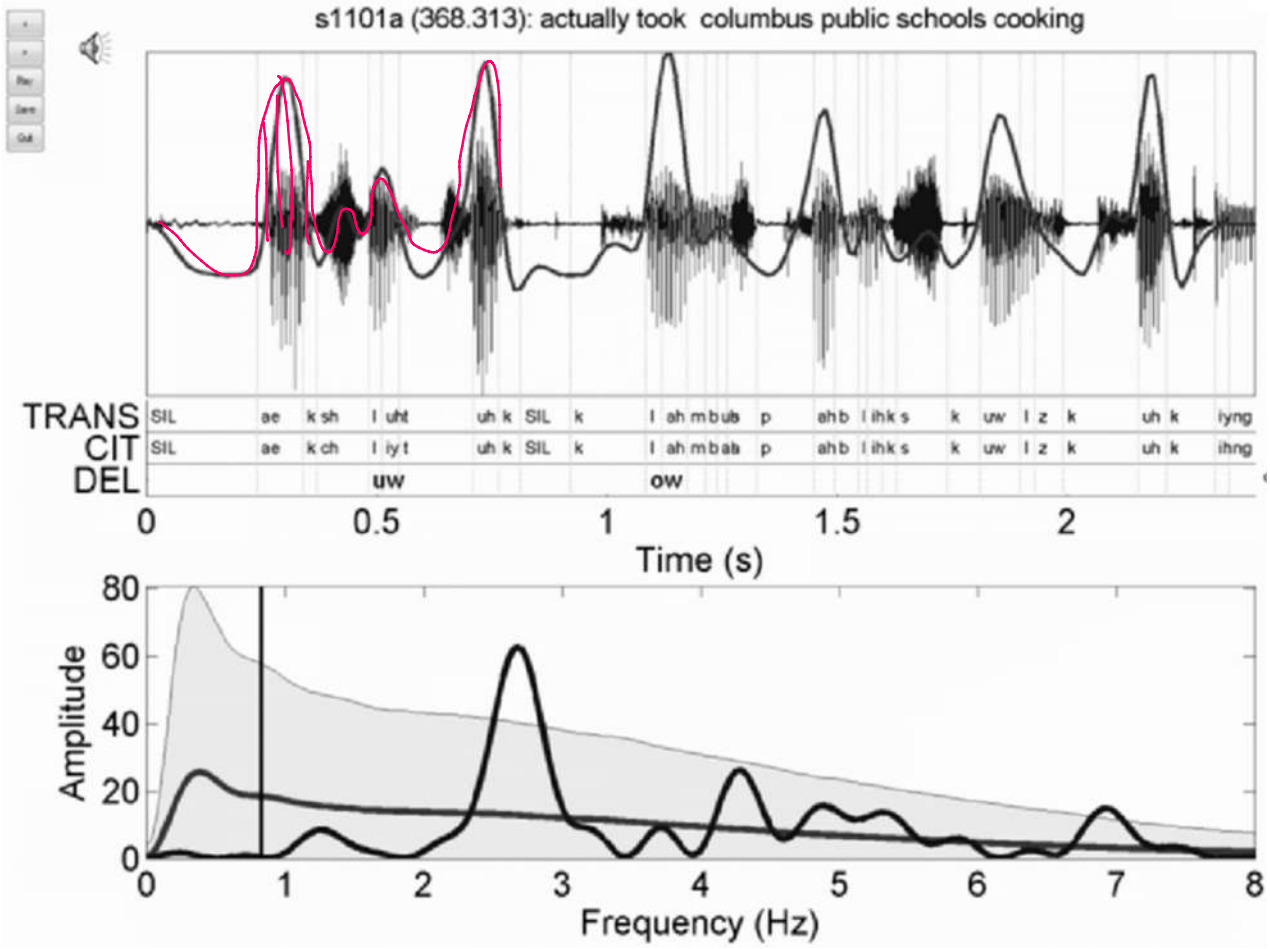
1. What is Feature Extraction from text?
2. Why do we need it?
3. Why is it difficult?
4. What is the core idea?
5. What are the techniques?







[23, 45, 64, 61, 12]



Hello how are you?

word2vec & embeddings

OHE

BOW

ngrams

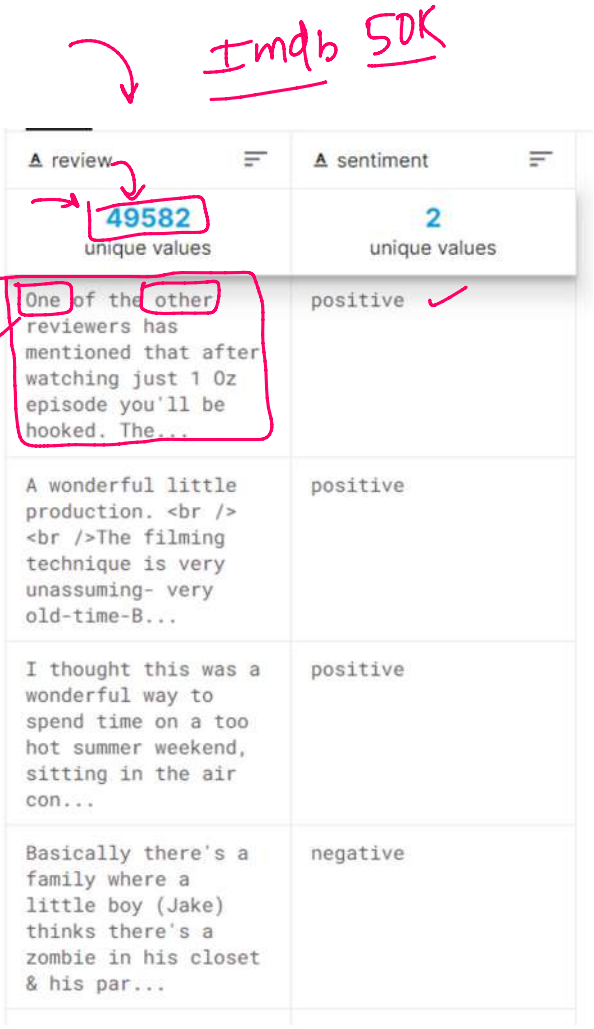
TfIdf

custom features

Common Terms

Monday, December 13, 2021 1:07 PM

1. Corpus
2. Vocabulary
3. Document
4. Word



review	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

One Hot Encoding

Monday, December 13, 2021 1:15 PM

50000 → 100 words → V → 50,000

→ ML Algo

D1	people watch campusx	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

hello campusx people 1

Vocabulary

people watch campusx write comment

V = 5

D1 = [[1,0,0,0,0], [0,1,0,0,0], [0,0,1,0,0]] → (3,5)

W → V dim

V = 50,000

[ ] → sparse array

Overfitting

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

OOV

D2 = [[0,0,1,0,0], [0,1,0,0,0], [0,0,1,0,0]] → (3,5)

pros

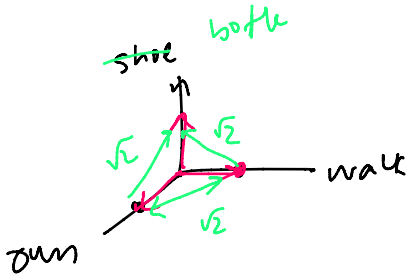
→ intuitive
→ easy to implement

flaws

#1 sparsity →
#2 No fixed size →
#3 OOV →
#4 No capturing of semantic X

walk | run | shoe

	walk	run	shoe
walk	1	0	0
run	0	1	0
shoe	0	0	1



5 50000

Bag of Words

Monday, December 13, 2021 6:22 PM

→ Text Classification

binary = True / False

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

V = 5

Vdim (5d)

	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

(0, 1)

Bag

core intuition → Order of words

→ Context

frequency → exist

sentiment analysis

ndim

of Count vect

more words

max_feature = (1) / 2

Hello campus

Advantages

- Simple and intuitive

Hell how are you?

ordering

RNN LSTM

Disadvantages

#1 sparsity

#2 OOV

#3 Ordering

#4

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

ngrams

N-grams

Tuesday, December 14, 2021 10:57 AM

Bag of n-grams → BOW

Single word → unigram

BOW → Bag of n-grams

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

BOW 5

people	watch	campusx	write	comment

bi-gram | tri-gram | n-gram

1 word

Bag of bi-gram

V = 6

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

(D1, D2, D3, D4)

quad gram → error

n-grams

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

n grams

people watch campusx	s2	s3	s4	
1	0	0	0	D1
0	1	0	0	
0	0	1	0	D2
0	0	0	1	

D1: This movie is very good

D2: This movie is not good

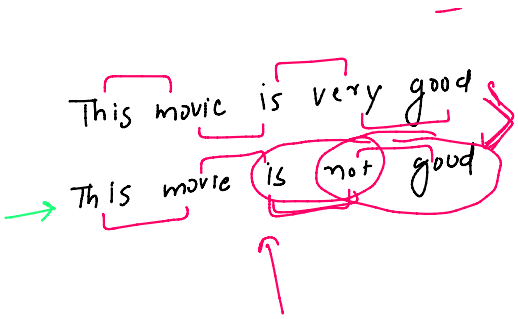
BOW (6D)

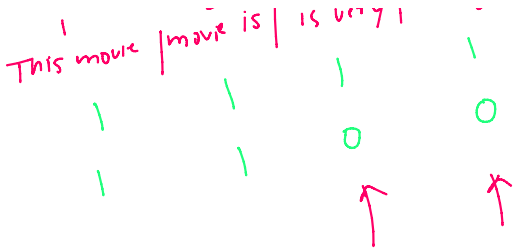
	This	movie	is	very	good	not
→	1	1	1	1	1	0
→	1	1	1	0	1	1

6D

This movie | movie is | is very | very good | is not | not good

This movie is very good





Large dataset

unigrams → bi-gram

tri-gram

Benefits

1) Able to capture semantic of the sentence

2) easy implement

Disadvantages

→ slows down the algo

|V| dimension increases

→ Out of vocab

Tf-Idf

Tuesday, December 14, 2021 11:54 AM

TF(people, D1) = 1/3

TF(cx, D2) = 2/3

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

(4)

TF x Idf

	IDF
people	log(4/2)
watch	log(4/2)
campusx	log(4/3)
write	log(4/2)
comment	log(4/2)

Idf(campus) = log(4/3)

0 < TF < 1

probability

vector

	people	watch	campusx	write	comment
D1	1/3 x 0.3	1/3 x 0.3	1/3 x 0.125	0	0
	0	1/3 x 0.3	2/3 x 0.125	0	0
	1/3 x 0.3	0	0	1/3 x 0.3	1/3 x 0.3
	0	0	1/3 x 0.125	1/3 x 0.3	1/3 x 0.3

term → frequent → Idf ↓ ↓ rare → idf ↑

TF x Idf

$$\mathrm{TF}(t,d) = \frac{\text{(Number of occurrences of term } t \text{ in document } d)}{\text{(Total number of terms in the document } d)}$$

$$\mathrm{IDF}(t) = \log_e \frac{\text{(Total number of documents in the corpus)}}{\text{(Number of documents with term } t \text{ in them)}}$$

Idf = log(4/4) = log 1 = 0

Tfx Idf = 0

IDF = 0

$\log_e(1) = 0$

IDF ↑

Idf = (N/n)

10000 reviews

the

Idf = 1

Idf = 10000

log

Idf = 0

$\log_{10} = 4$

0<1

Tf x Idf

10000

V =

ency

Advan

→ Information retrival

Word2vec

Custom features

Disadvantages

sparsity

OOV

dimension

semantic

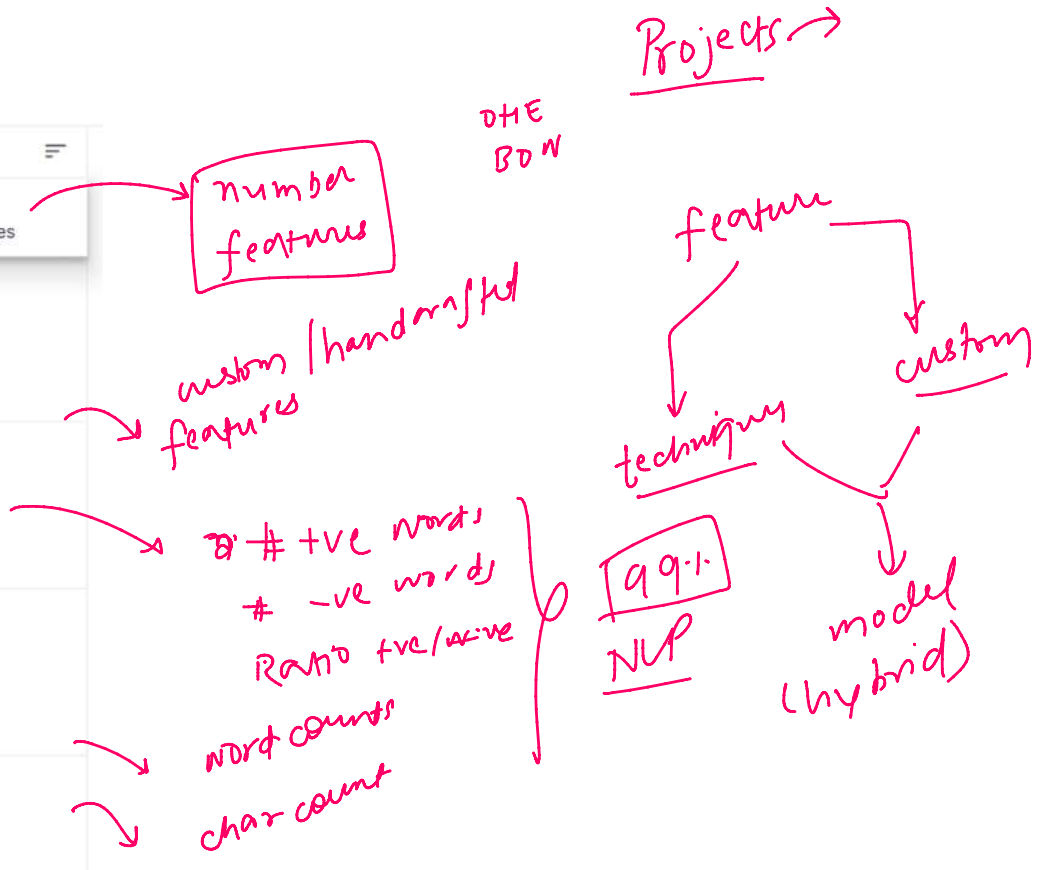
this is beautiful

this is gorgi

Custom Features

Tuesday, December 14, 2021 2:59 PM

review	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

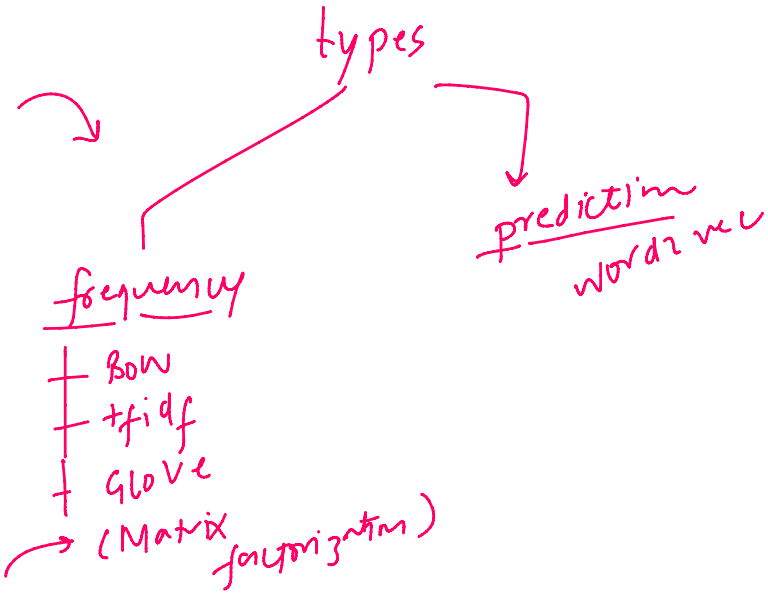


Word Embeddings

Wednesday, December 22, 2021 4:11 PM

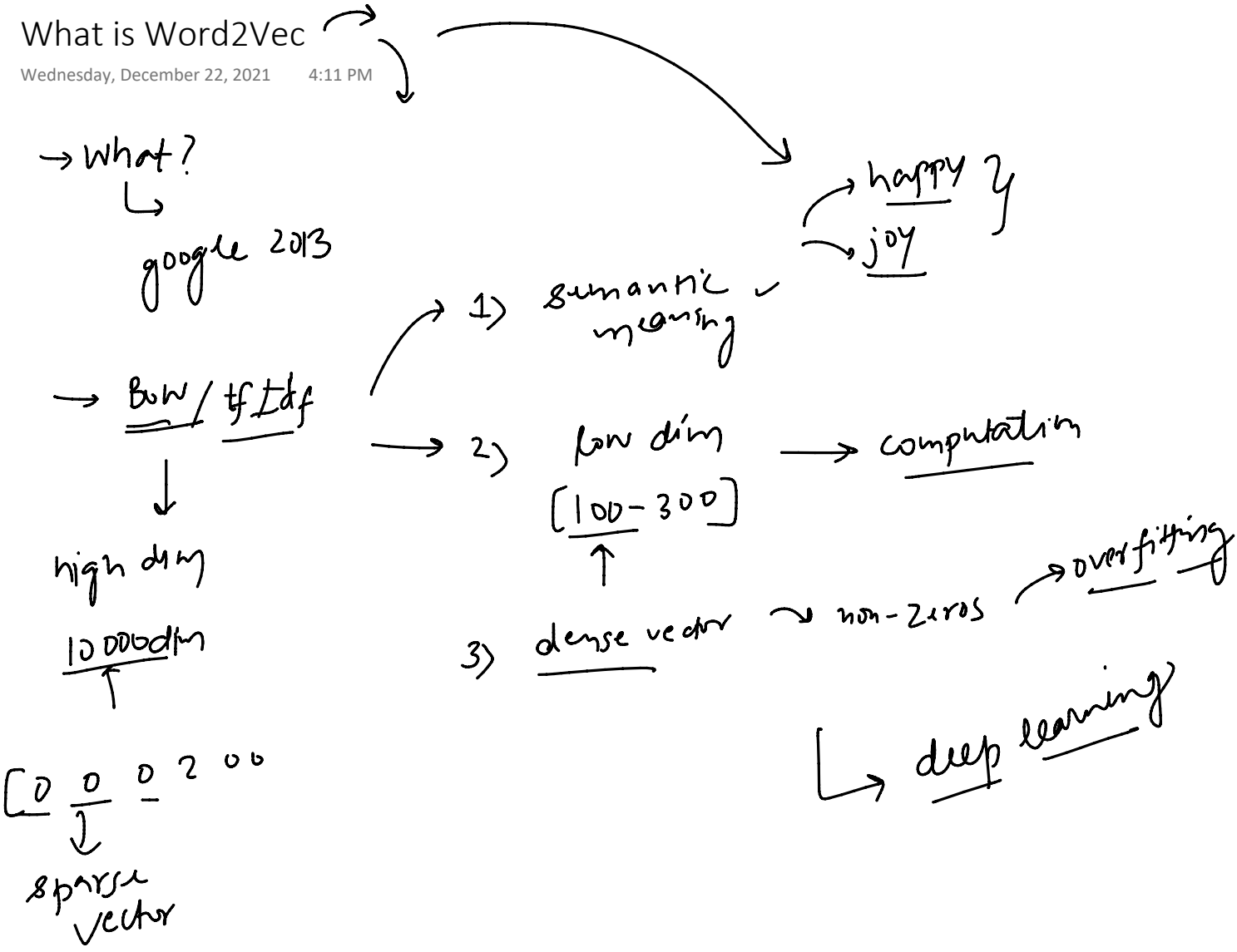
In natural language processing, word embedding is a term used for the representation of words for text analysis, typically in the form of a real-valued vector that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning.

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



What is Word2Vec

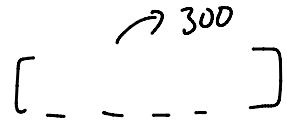
Wednesday, December 22, 2021 4:11 PM

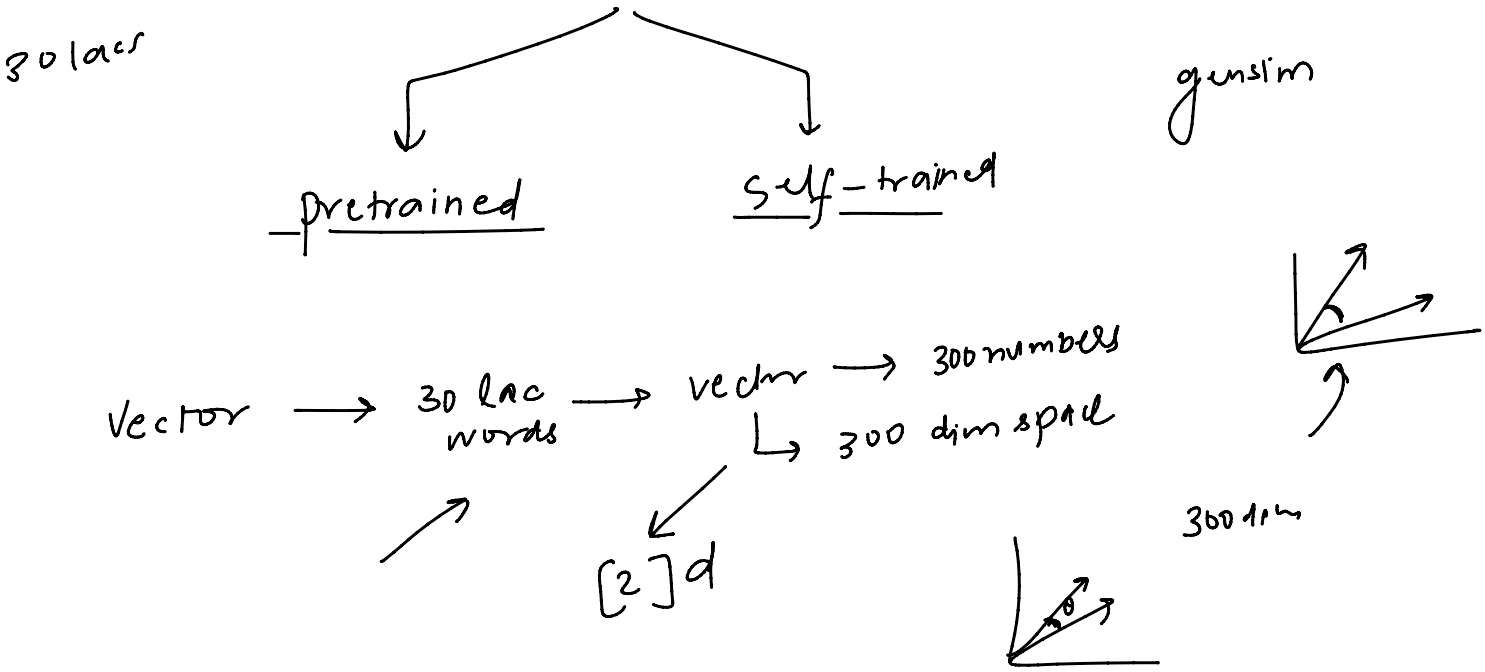


Demo

Wednesday, December 22, 2021 4:11 PM

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





Really Cool Stuff

Wednesday, December 29, 2021 2:08 PM

Intuition

Wednesday, December 22, 2021 4:12 PM

	King	Queen	Man	Woman	Monkey
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

[1 1 1 0.8 1]

↳ vocab (5)

Papers → 30 lacs

automated → neural network

features

f1 f2 f3 ...

0.5

[1 0 0 23 0]

King − man + woman

1 − 1 + 0 = 0

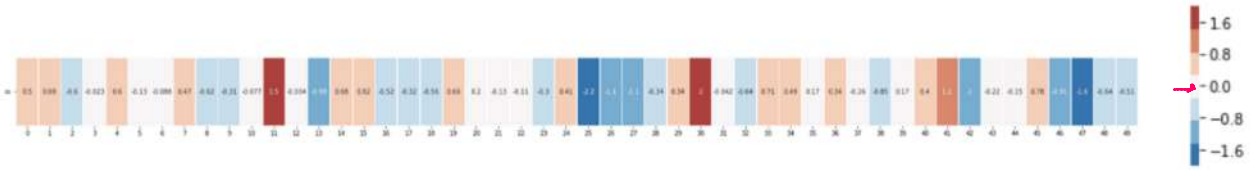
1 − 0.3 + 0.3 = 1

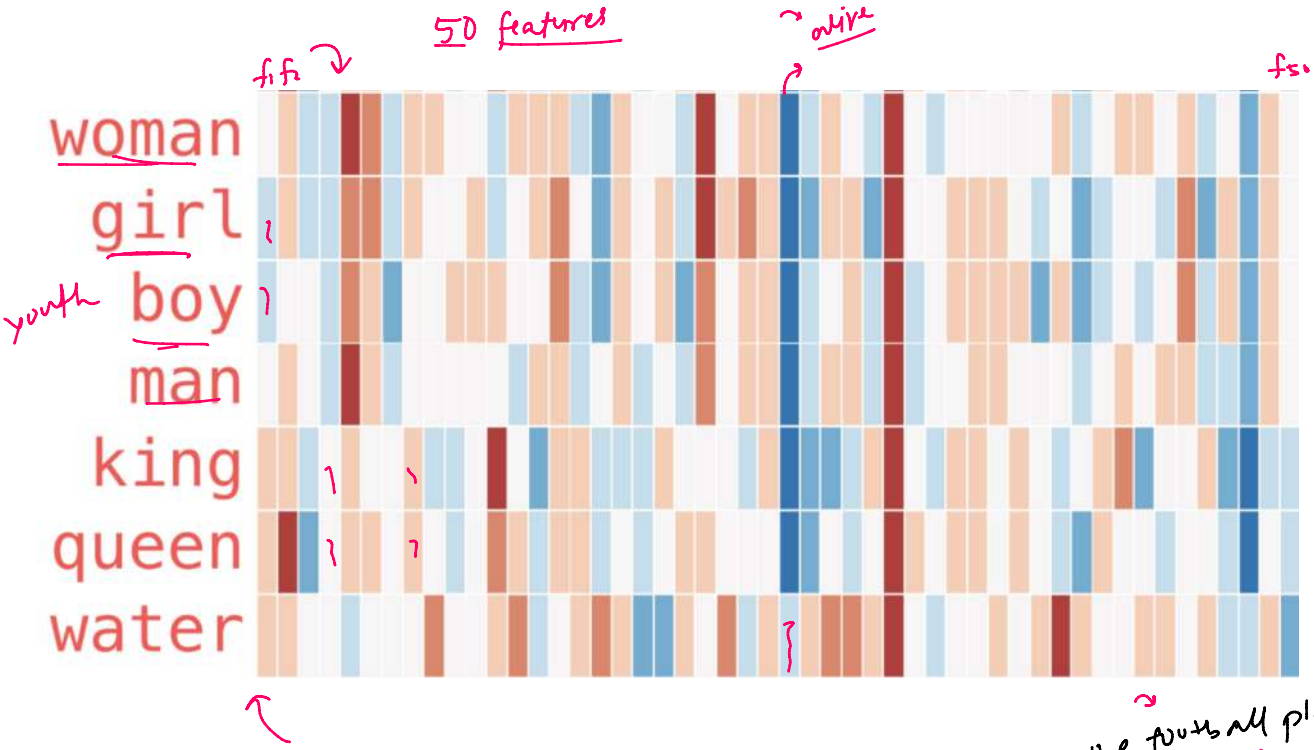
1 − 0.2 + 0.2 = 1

0.8 − 0.1 + 0.5 = 0.7

1 − 1 + 1 = 1

50 numbers

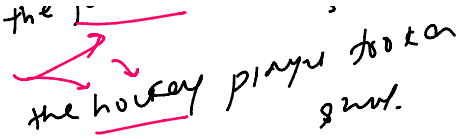




The underlying assumption of Word2Vec is that two words sharing similar contexts also share a similar meaning and consequently a similar vector representation

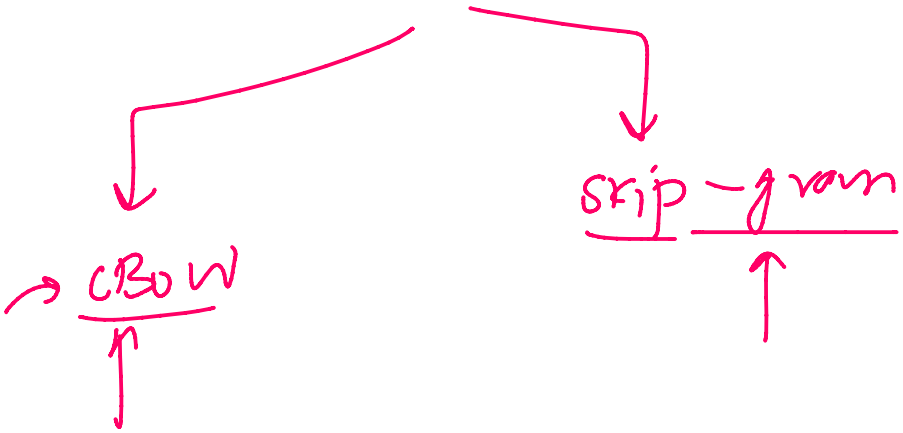
the football player took a shot

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.



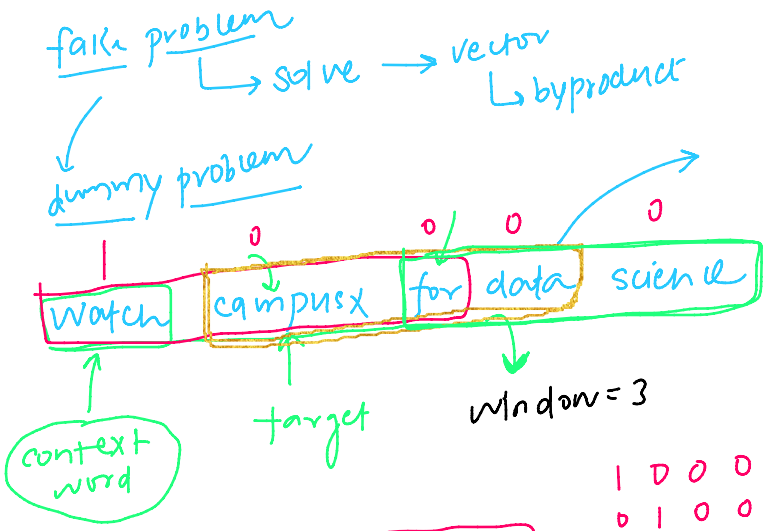
Types of Word2Vec

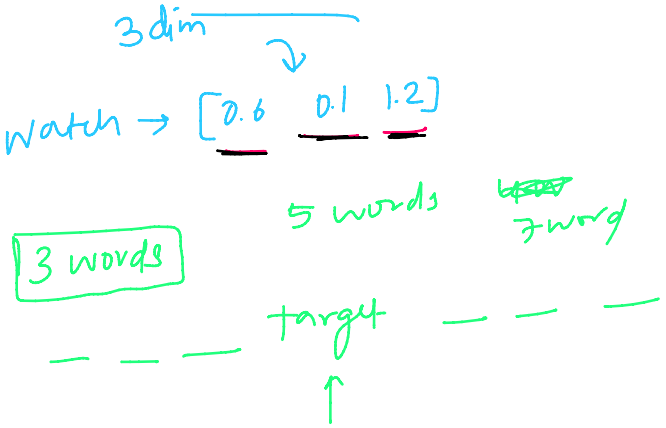
Wednesday, December 22, 2021 4:11 PM

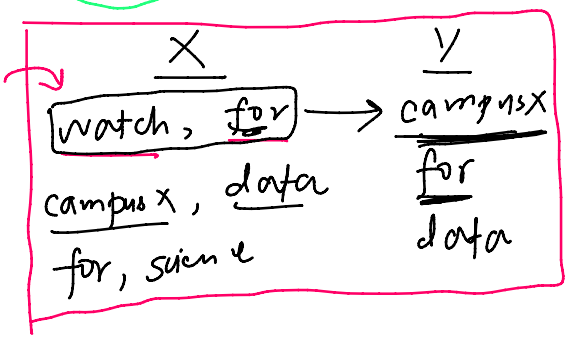


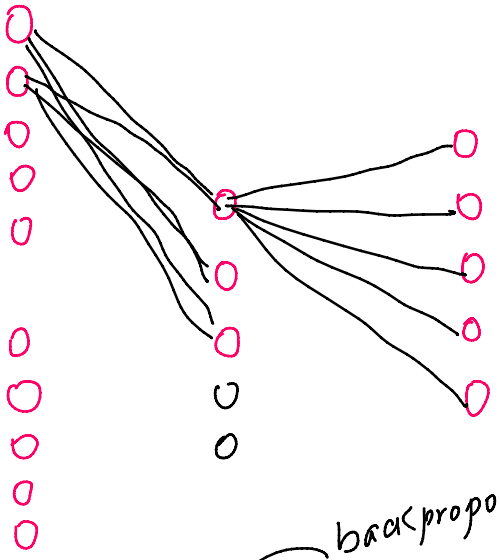
CBOW

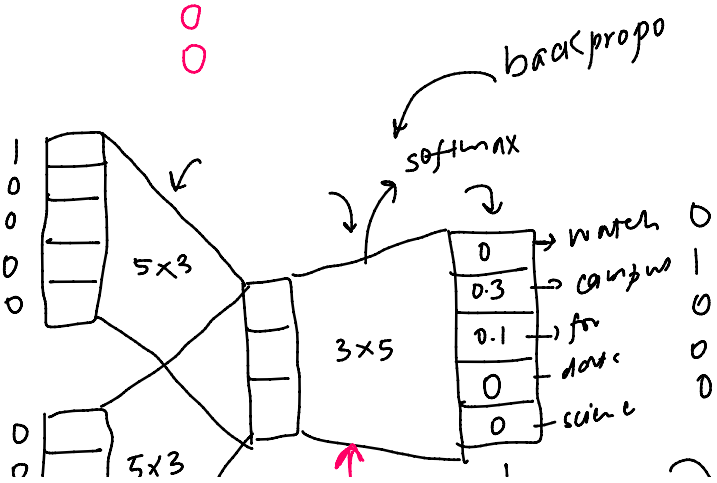
Wednesday, December 22, 2021 4:12 PM







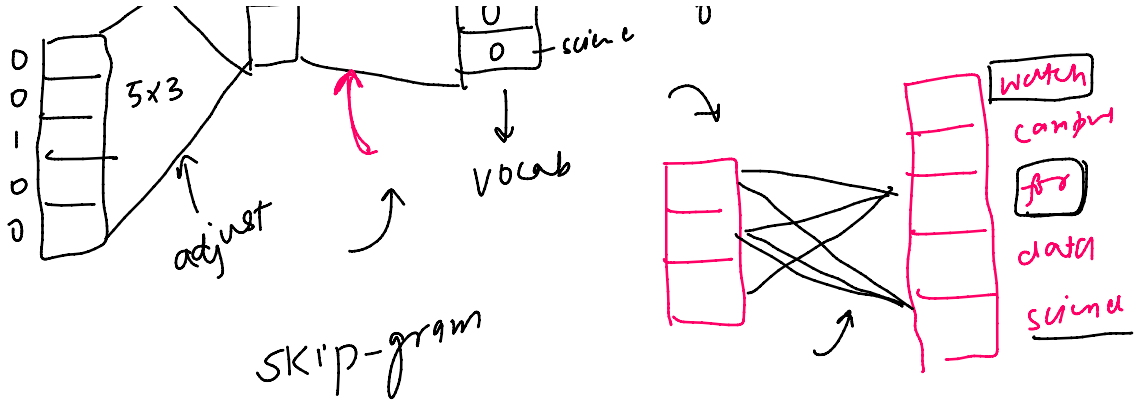




Ross ↓

CBOW

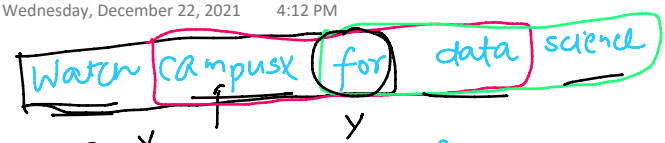
watch


0
0
1
0
0
3x3
adjust
0
0
science
vocab
watch
campus
for
data
science

skip-gram

Skip-gram

Wednesday, December 22, 2021 4:12 PM



campusx → watch, for ←

→ for → campusx, data

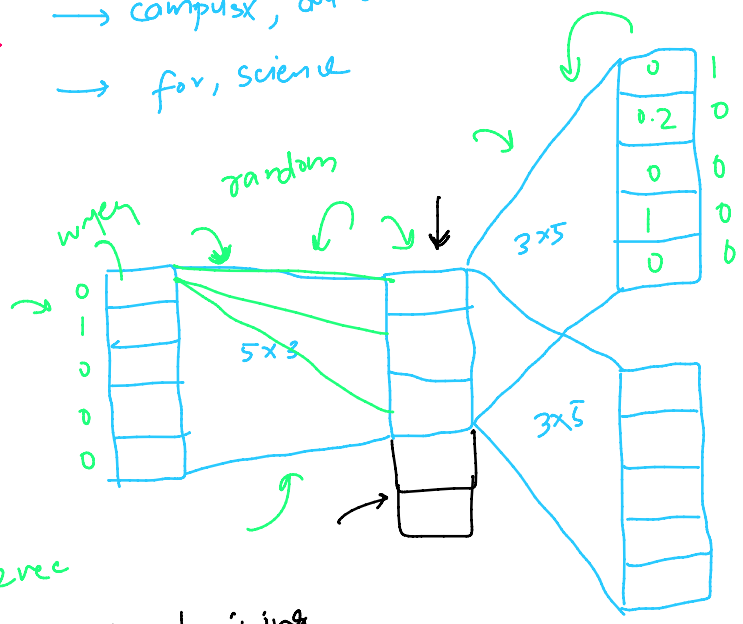
→ data → for, science

epochs

CBOW → dummy problem

window=3 =5 ?

loss



CBOW / skip-gram

- small data → CBOW
- large data → skip-gram

word2vec

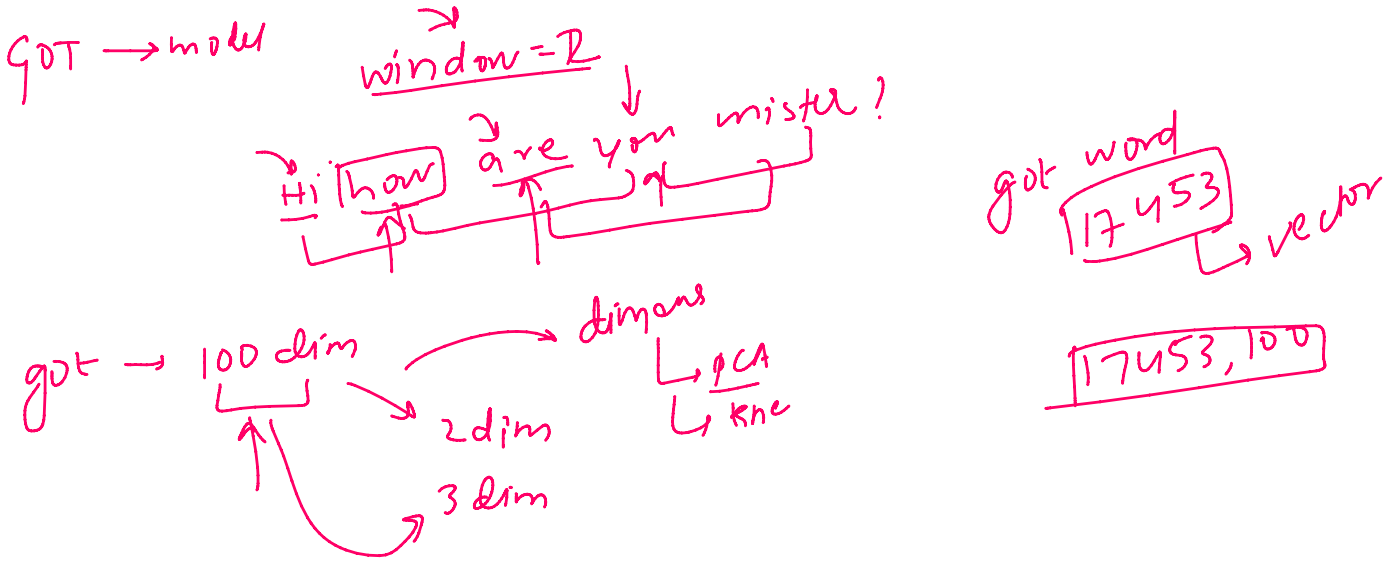
→ increase the training

→ increase dimension of vectors

→ increase window size

Training your own model

Wednesday, December 22, 2021 4:12 PM

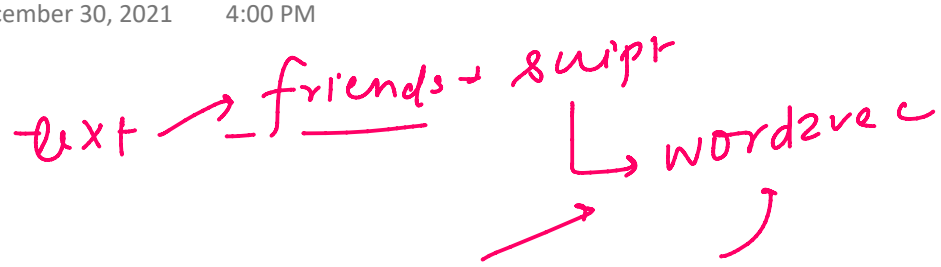


Visualization

Wednesday, December 29, 2021 12:24 PM

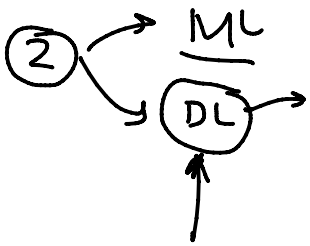
Assignment

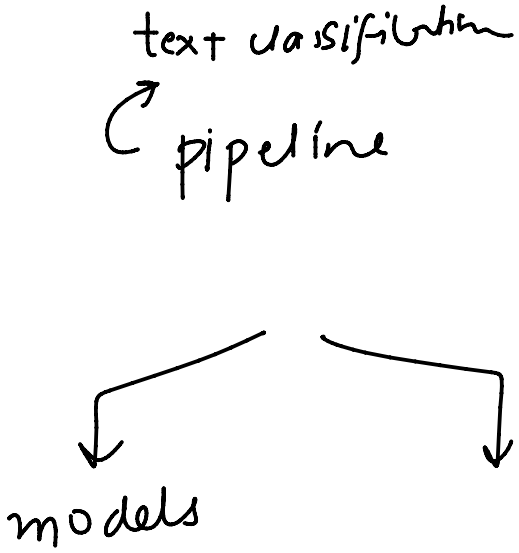
Thursday, December 30, 2021 4:00 PM



Plan of Attack

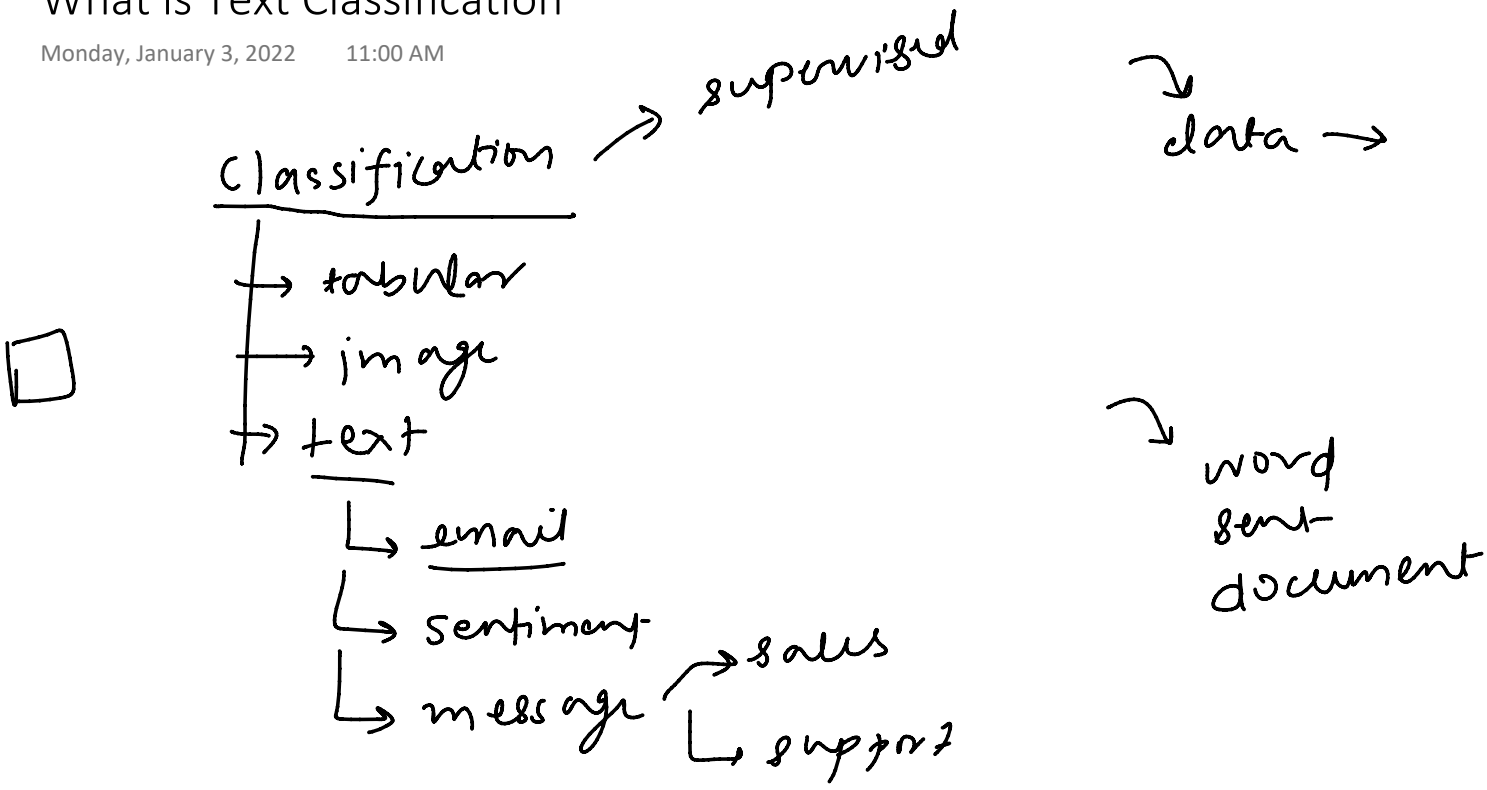
Wednesday, January 5, 2022 5:39 PM





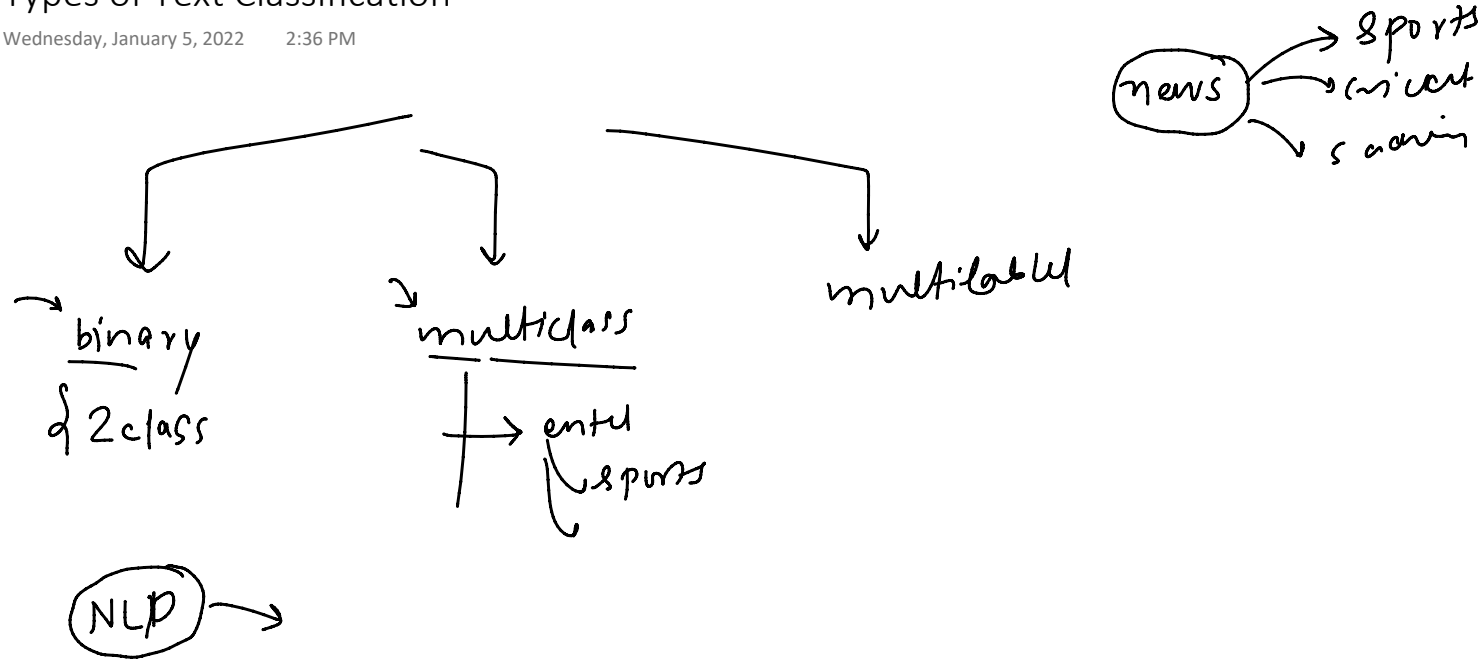
What is Text Classification

Monday, January 3, 2022 11:00 AM



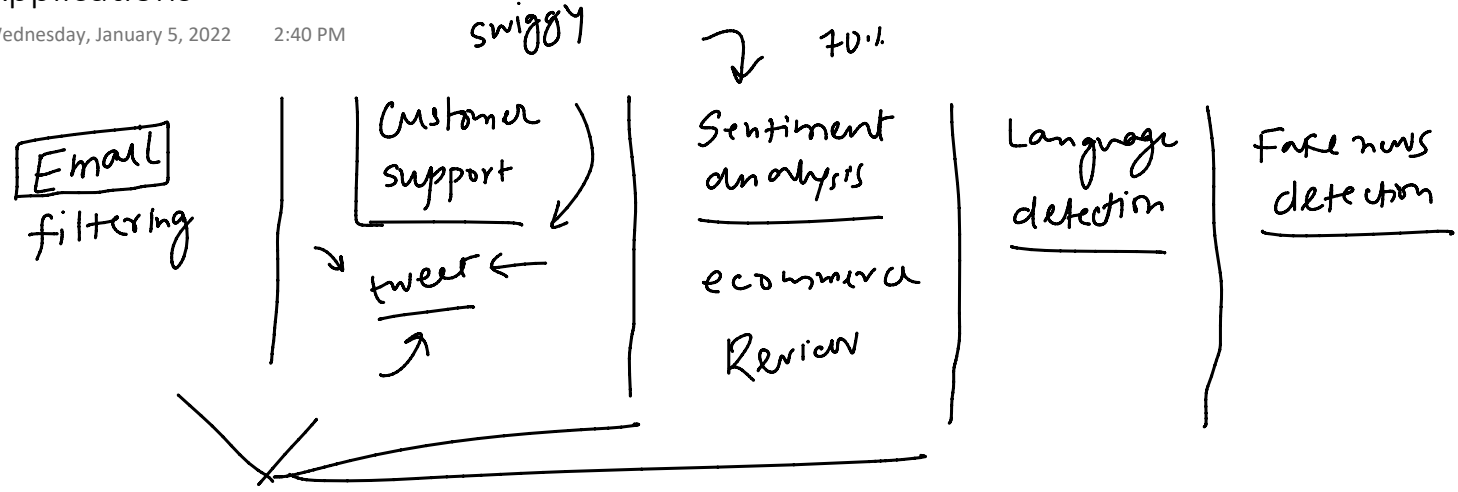
Types of Text Classification

Wednesday, January 5, 2022 2:36 PM



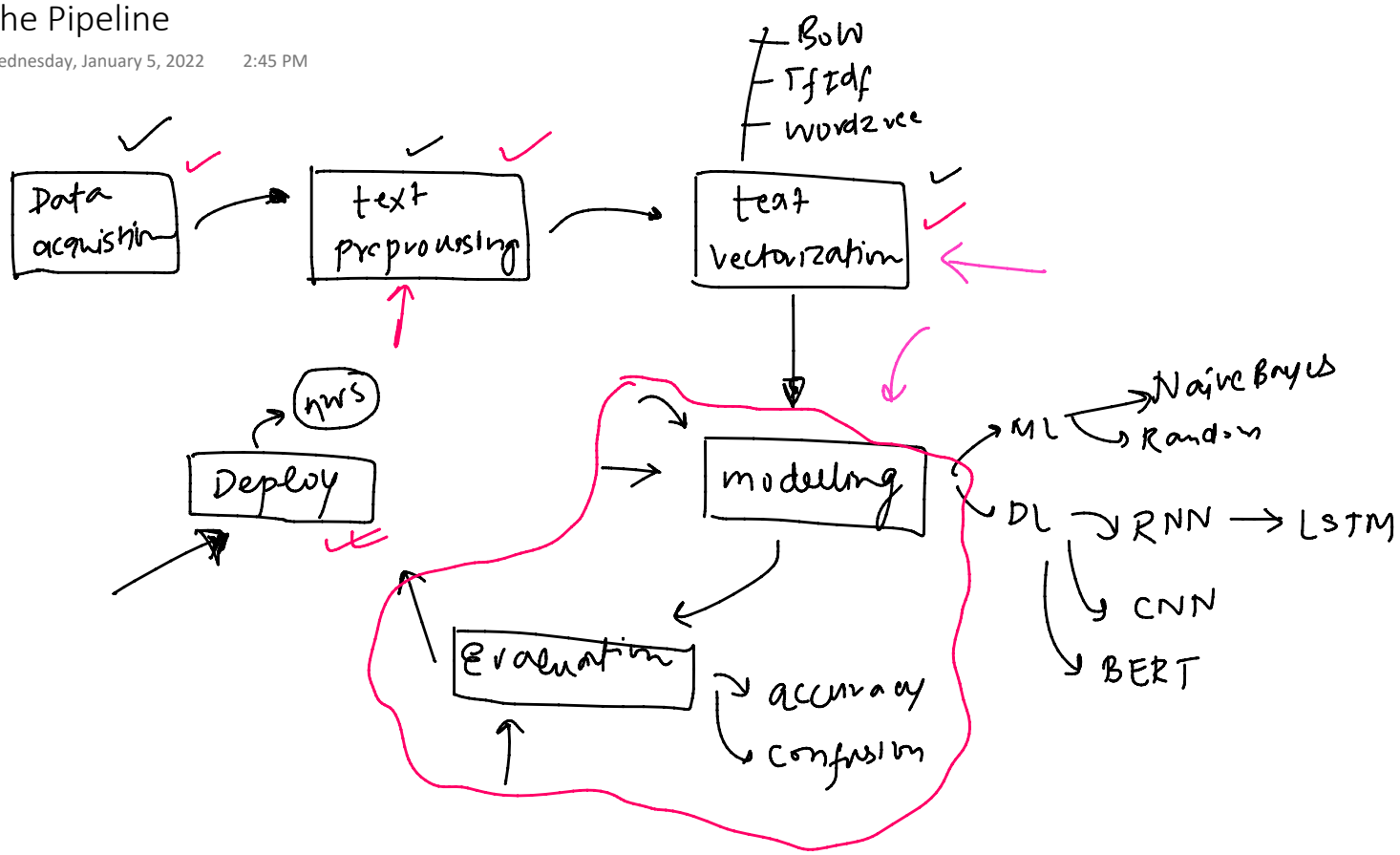
Applications

Wednesday, January 5, 2022 2:40 PM



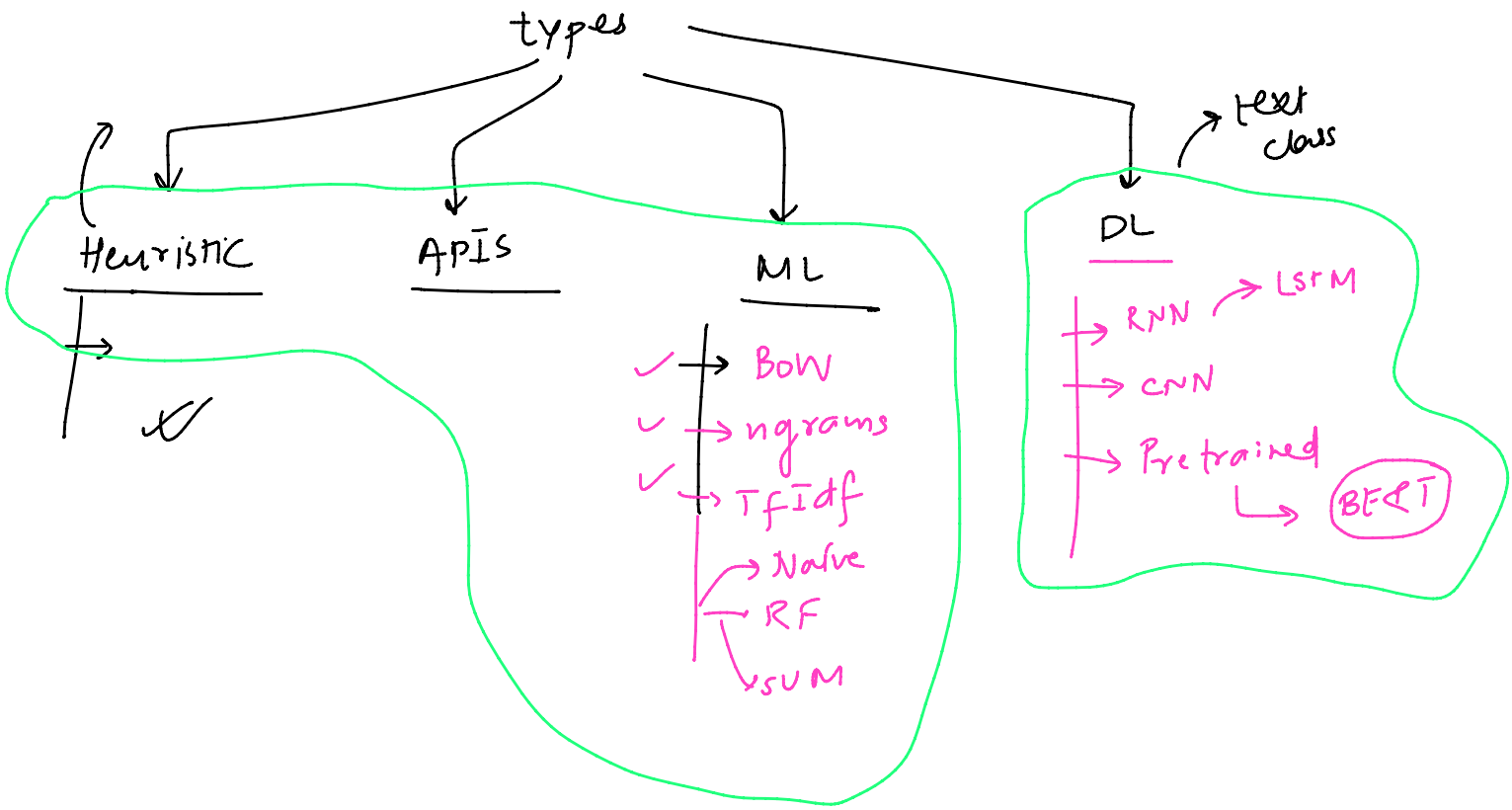
The Pipeline

Wednesday, January 5, 2022 2:45 PM



Different Approaches

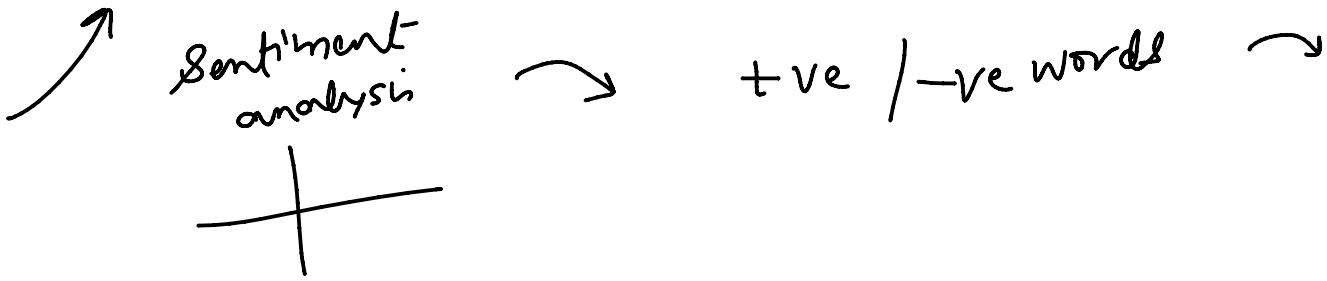
Wednesday, January 5, 2022 2:36 PM



Heuristic Approach

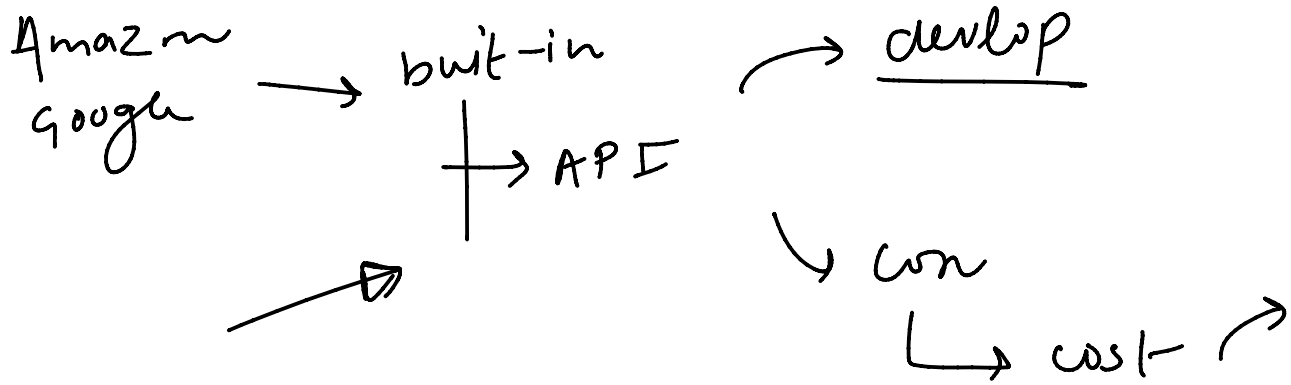
Wednesday, January 5, 2022 5:46 PM

1980's



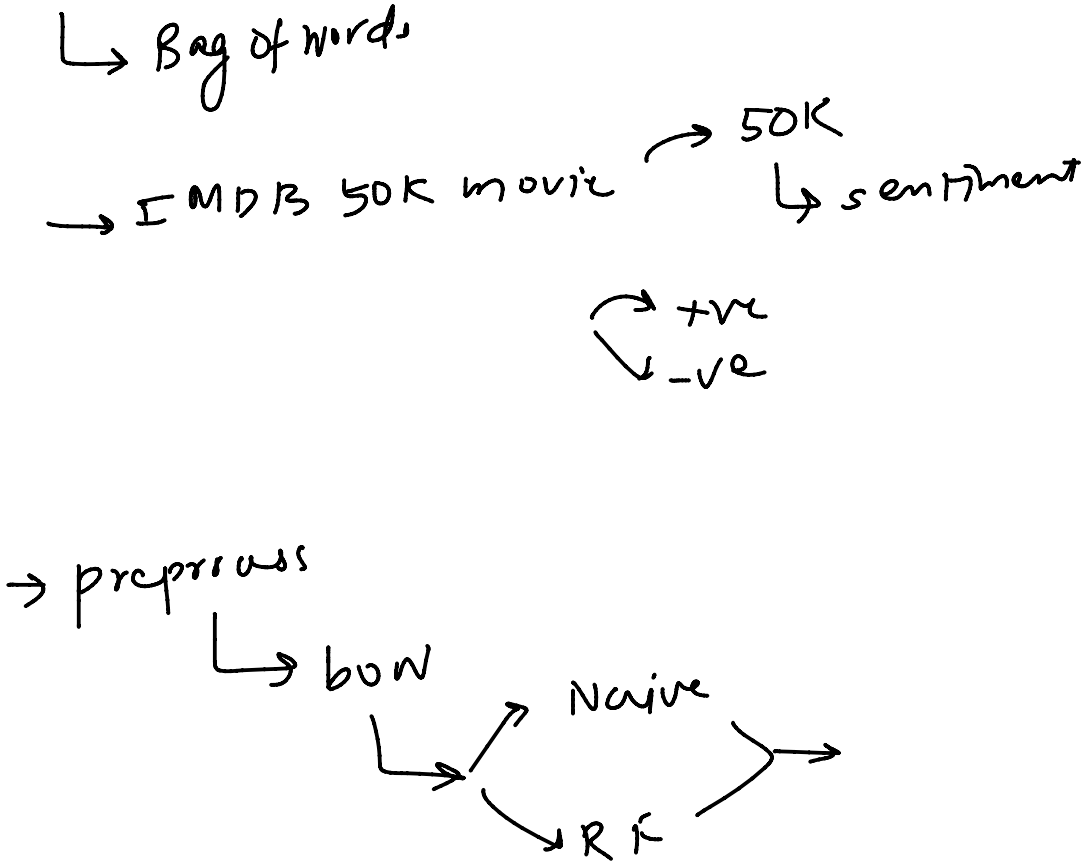
Using API

Wednesday, January 5, 2022 5:46 PM



Using BoW and n-grams

Wednesday, January 5, 2022 3:24 PM

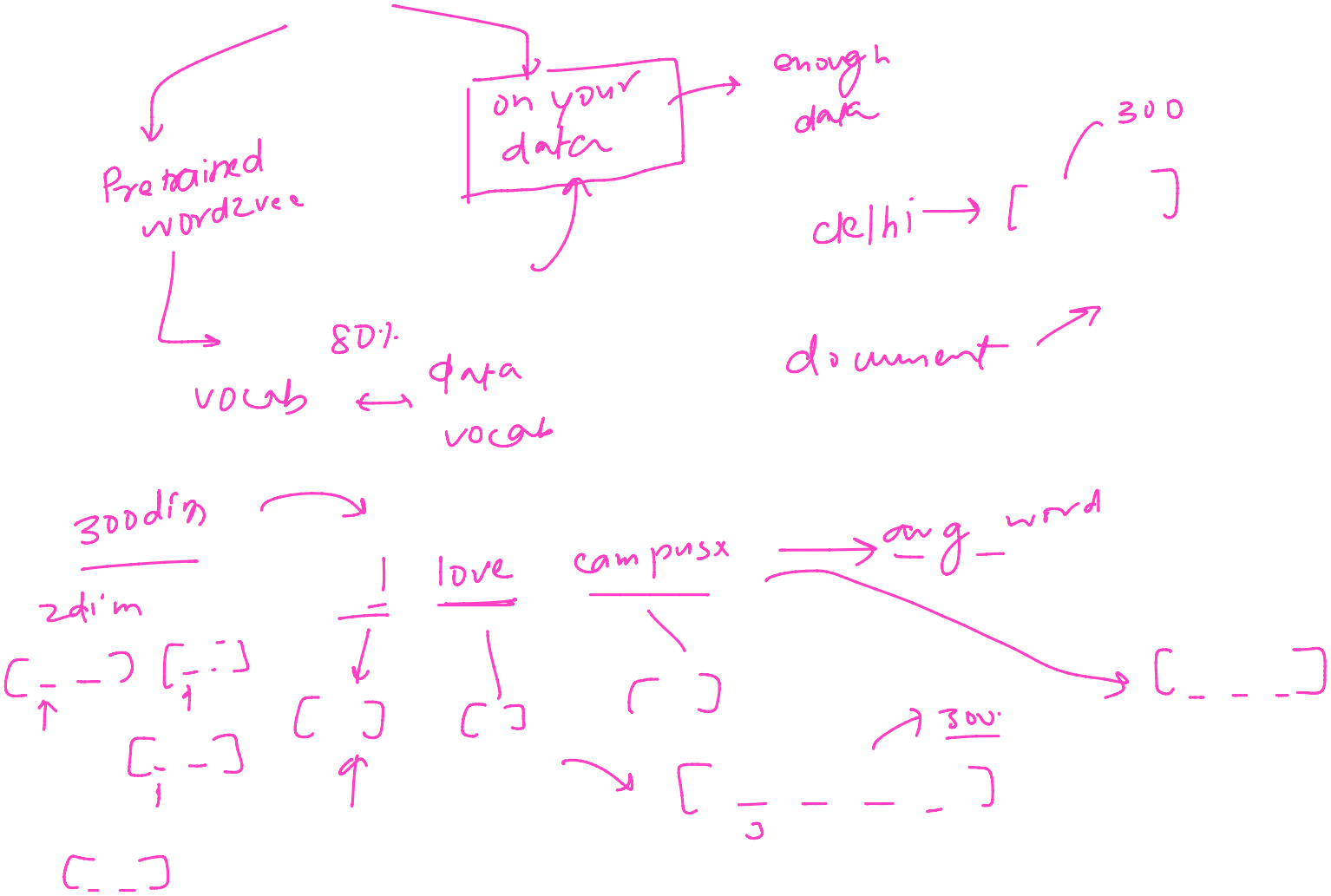


Using Tf-Idf

Wednesday, January 5, 2022 3:24 PM

Using Word2Vec

Wednesday, January 5, 2022 3:24 PM

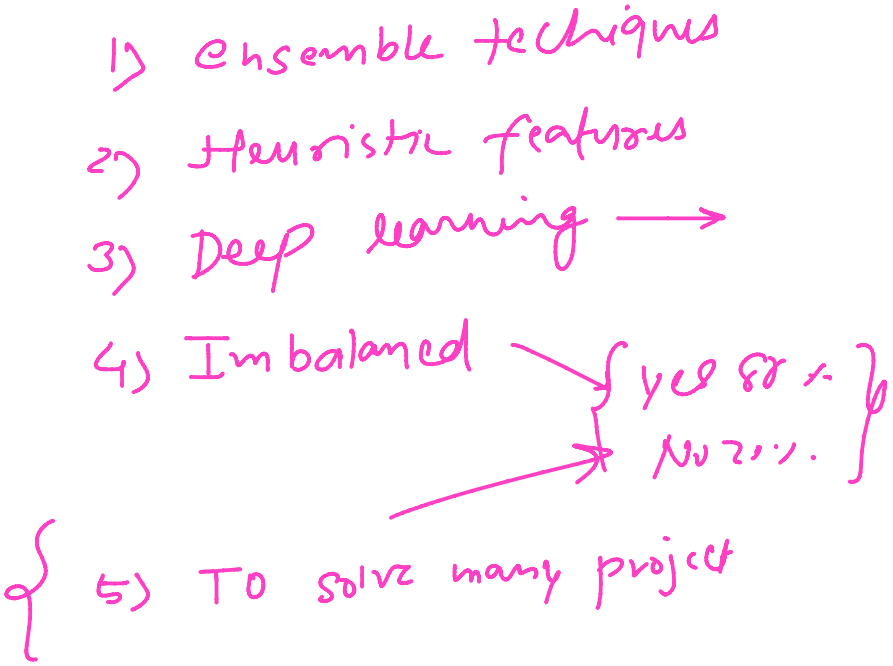


Interpreting models

Wednesday, January 5, 2022 3:45 PM

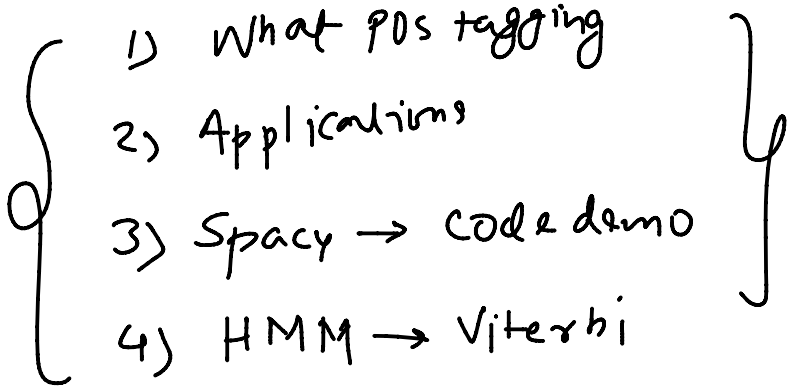
Practical Advise

Wednesday, January 5, 2022 3:45 PM



Plan of Attack

Thursday, January 20, 2022 1:44 PM



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.

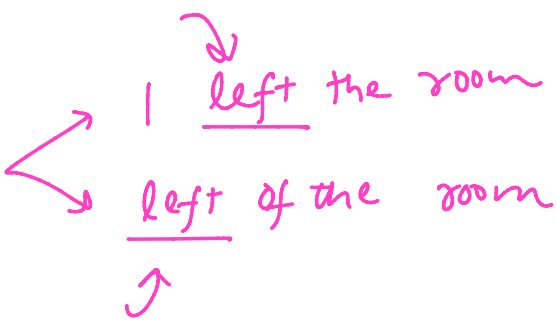
Why	not	tell	someone	?
adverb	adverb	verb	noun	punctuation mark, sentence closer

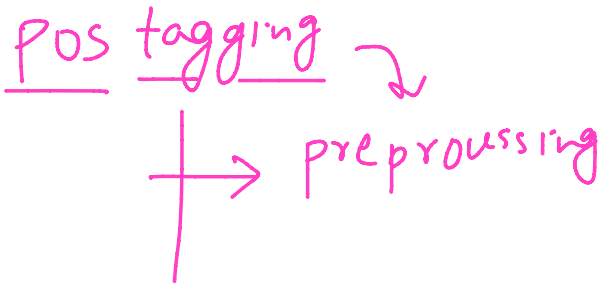
Application

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]

N V M

Emission Prob

Transition Prob

S (N) Nitish (V) loves (N) campusx E

S (M) Can (N) Nitish (V) google (N) campusx E

S (M) Will (N) Ankita (V) gougle (N) campusx E

S (N) AnKita (V) loves (N) will E

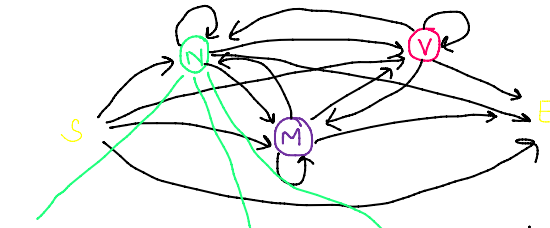
S (N) Will (V) loves (N) google E

→ Emission

	N	M	V
Nitish	2/10	0	0
loves	0	0	3/5
campusx	3/10	0	0
google	1/10	0	2/5
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



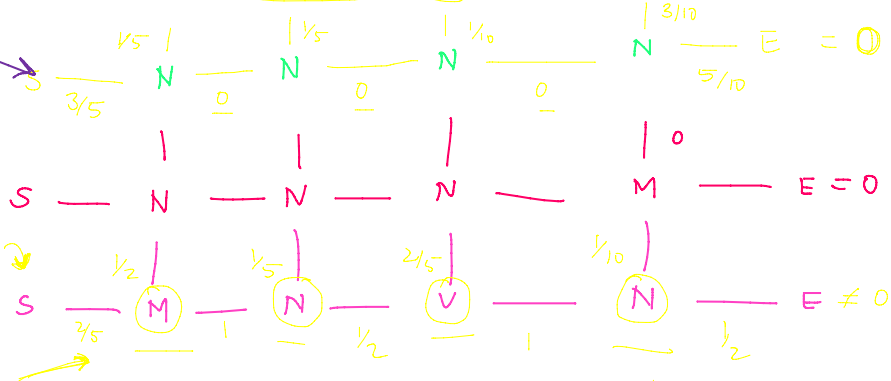
Nitish loves campusx will ankita can

$3^4 = 81$ → exponential

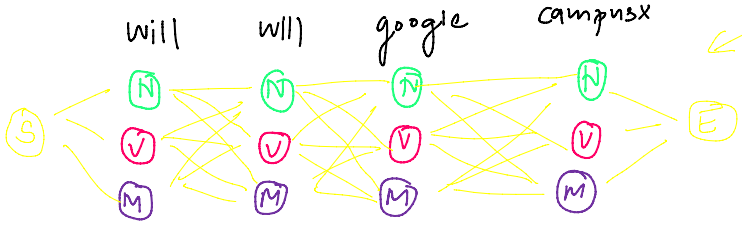
NVM

M N V N

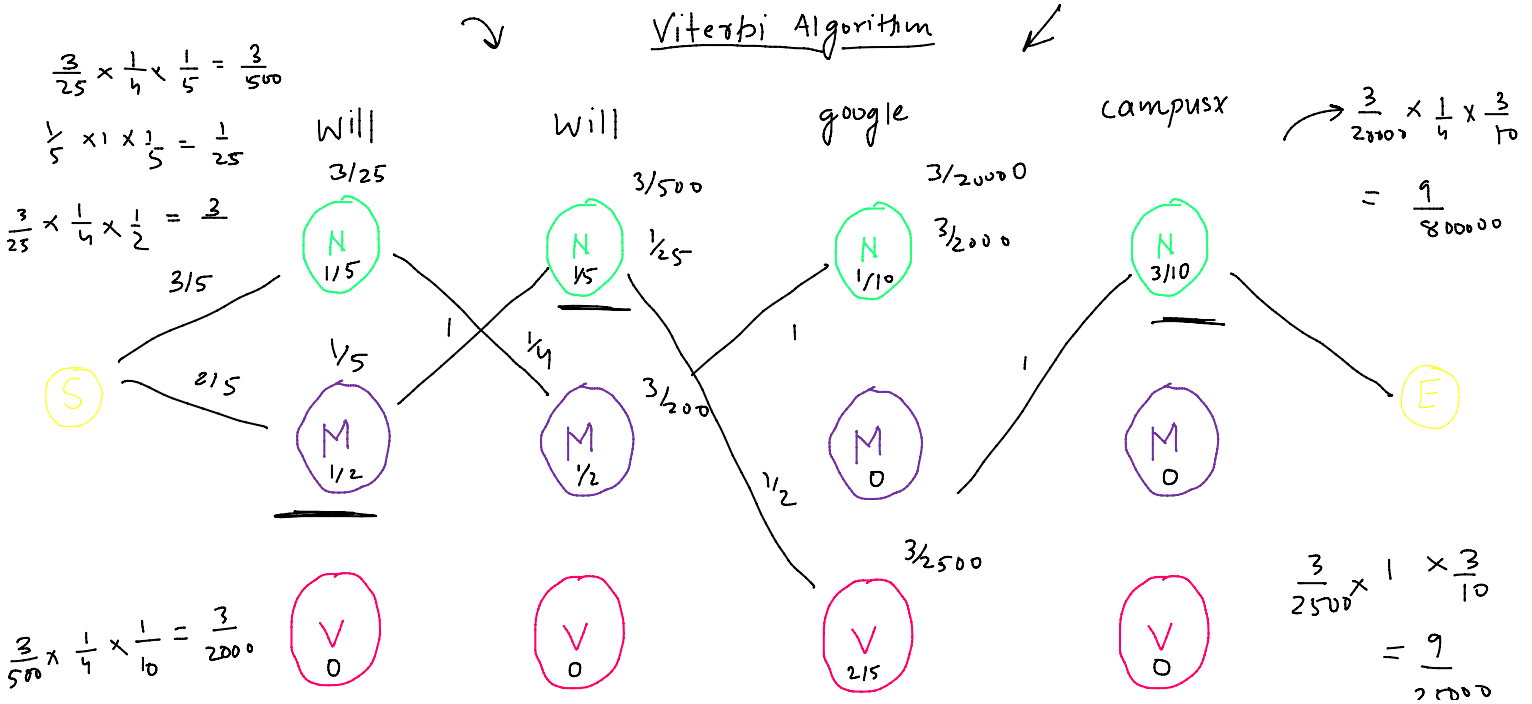
will will google campusx



Optimization → Viterbi Algo



Viterbi Algorithm



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

V
0

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

V
0

$\frac{3}{500} \times \frac{1}{x} \times \frac{x}{5} = \frac{3}{2500}$

V
215

V
0

2500

$= \frac{9}{25000}$

19 December 2023 14:43