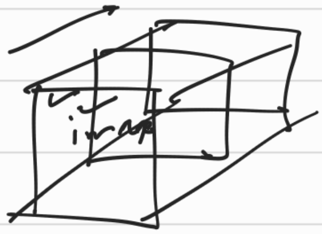
Natural Language Processing

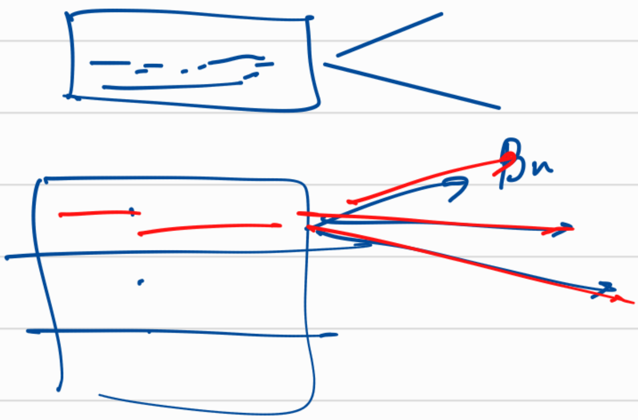
to enable machines → human language.



→ doesn't have a universal representation in numbers.

→ modeling.

→ Classification

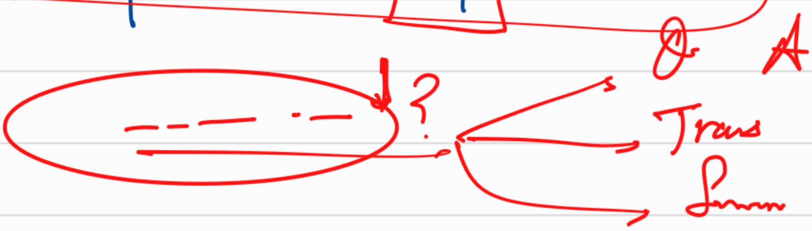


Timing, Topic

→ Named Entity

→ I have a ~~meeting~~ class at 11:00 am on Saturday about NLP.

Large Language modeling



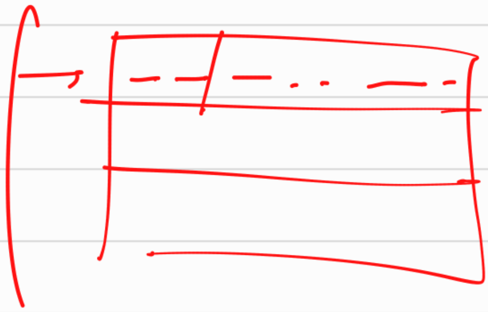
Definitions:

Corpus: ? → set / Union of document.

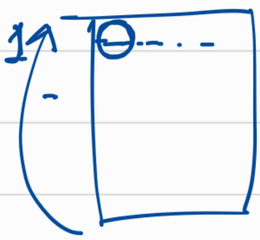
Document: ??

Token

Chatgpt → Corpus → all the internet data.



Token: ?? → 1 word



1 Token → Smaller

1 word (Running) → Run, ing

I am on hour . break

I am on having , each

Vocabulary:

Union / Collection of Tokens

NLP
① text → numbers
② Model.

text → numbers (Embedding)

Prediction — Frequency

text → numbers

Frequency

Word

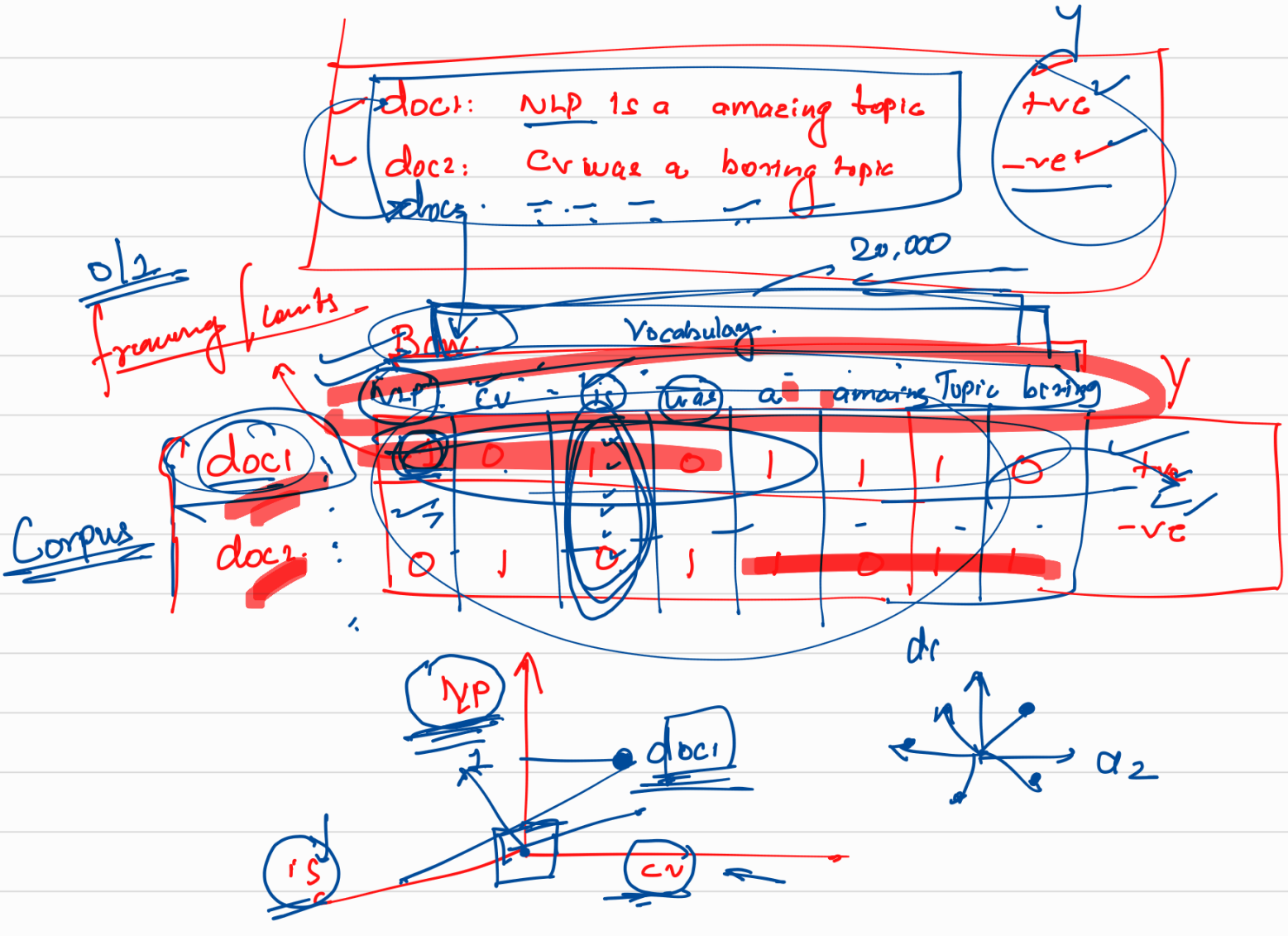
Word →

1 dimensional

dense

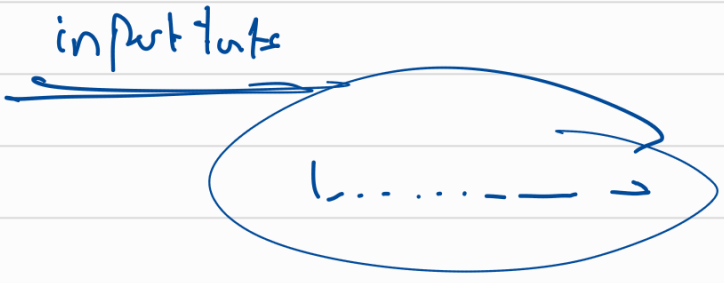
Embedding

Video



High dimensions → Sparsity

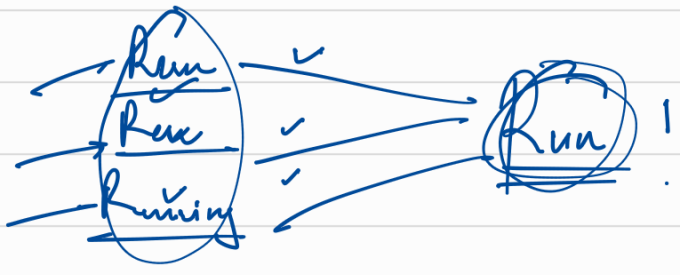
Ignores → order

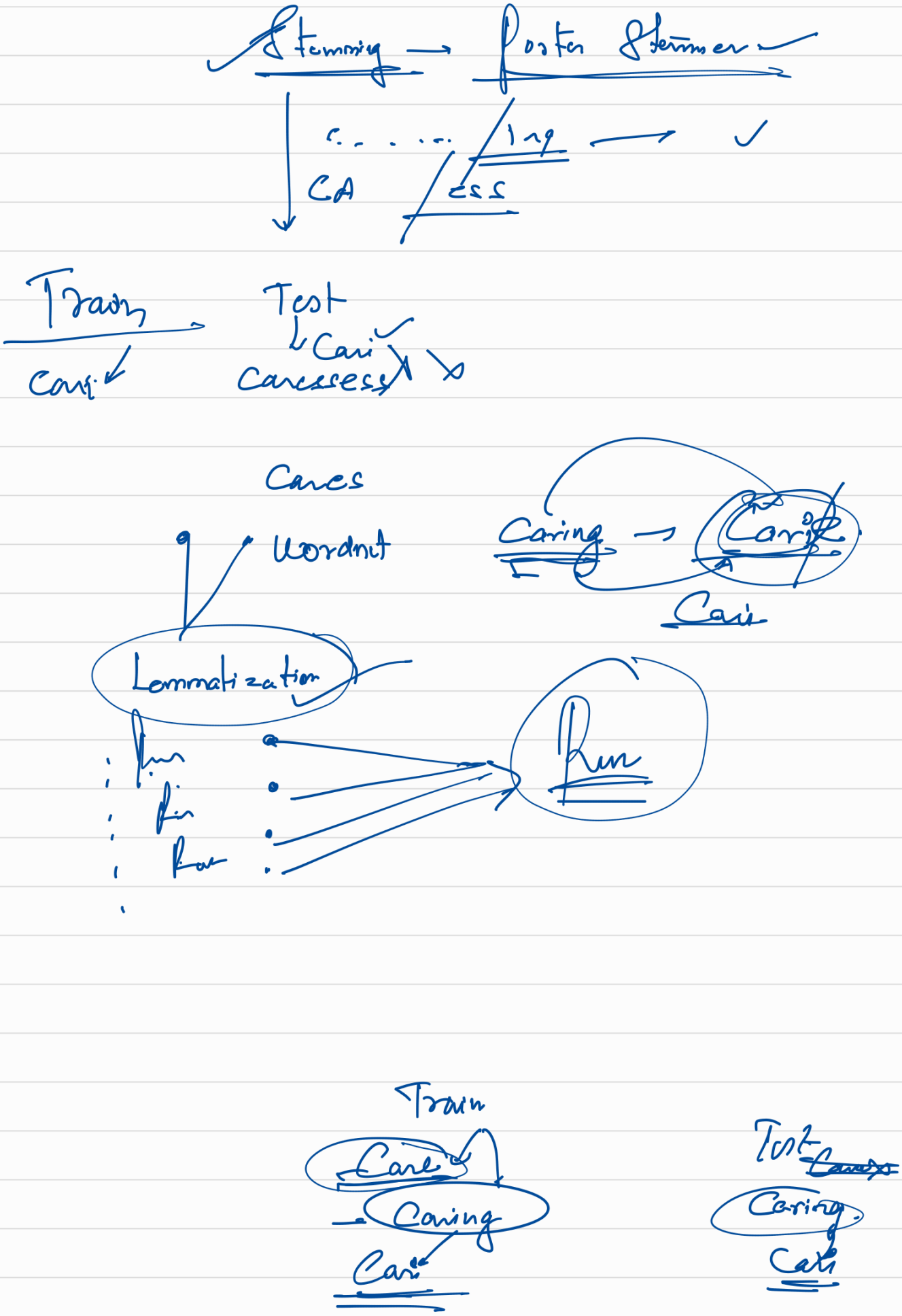


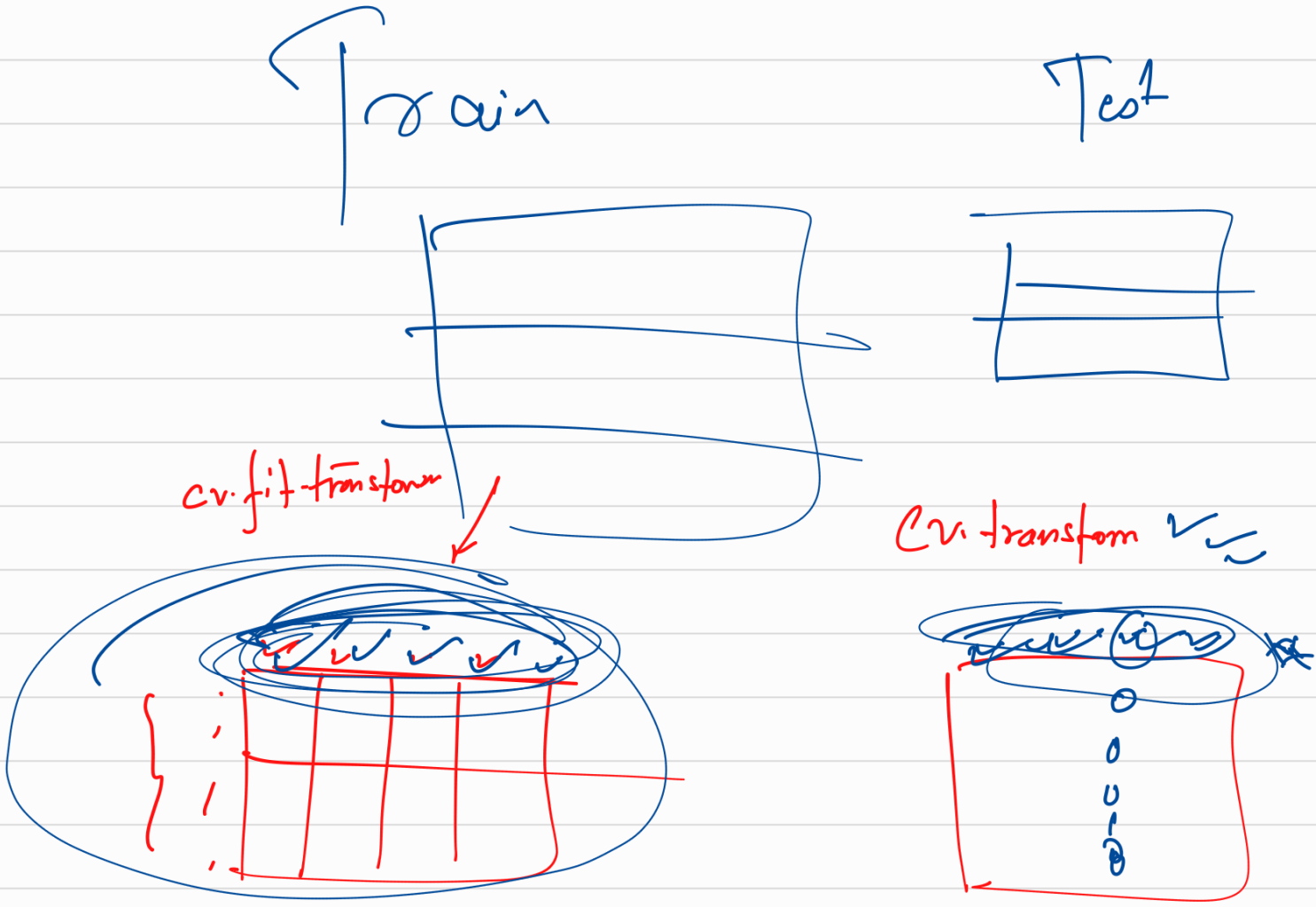
12: IDF

Higher dimensionality →

① I can remove stop words, remove punctuations, remove #, , remove numbers ✓

② 

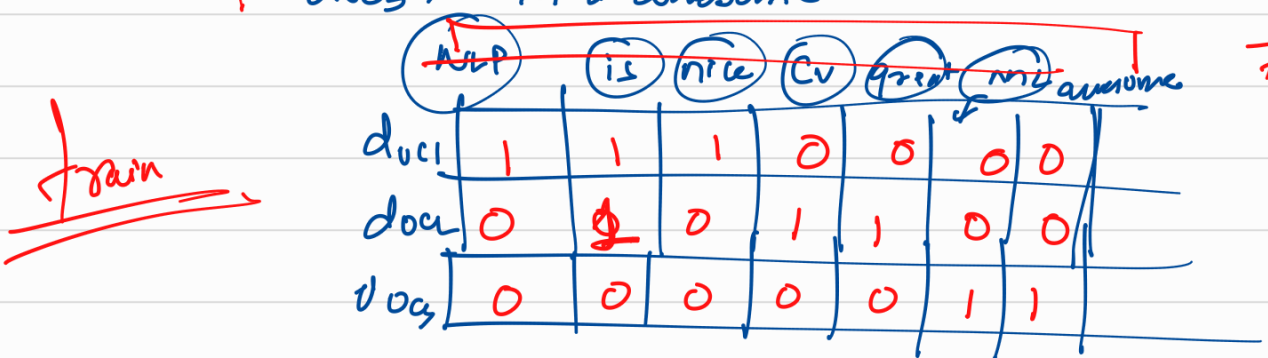

Stemming → Porter Stemmer
ing → ✓
CA
ess
Train
Test
Caring → Caring
Caring
Lemmatization
Run
Train
Caring
Caring
Test
Caring
Caring



doc1: NLP is nice

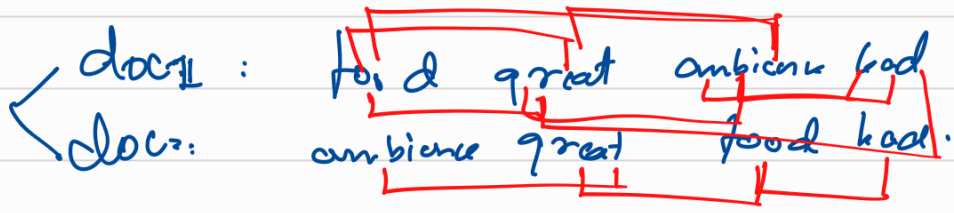
doc2: CV is great.

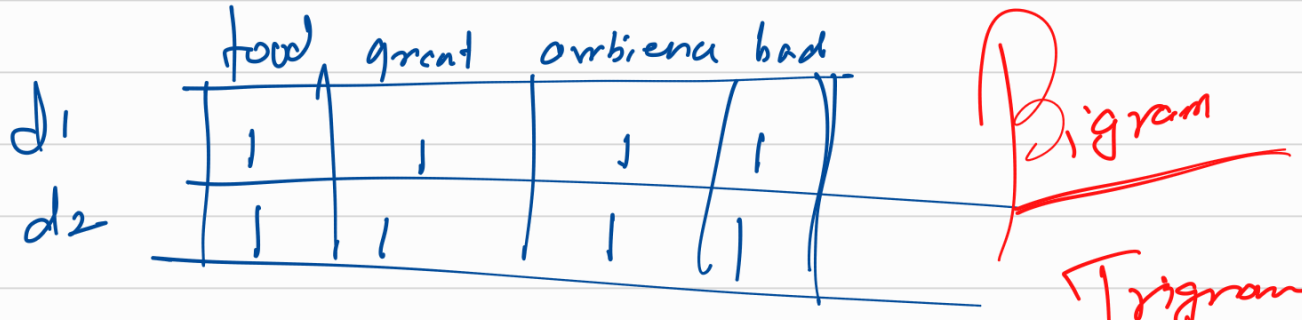
doc3: ML is awesome

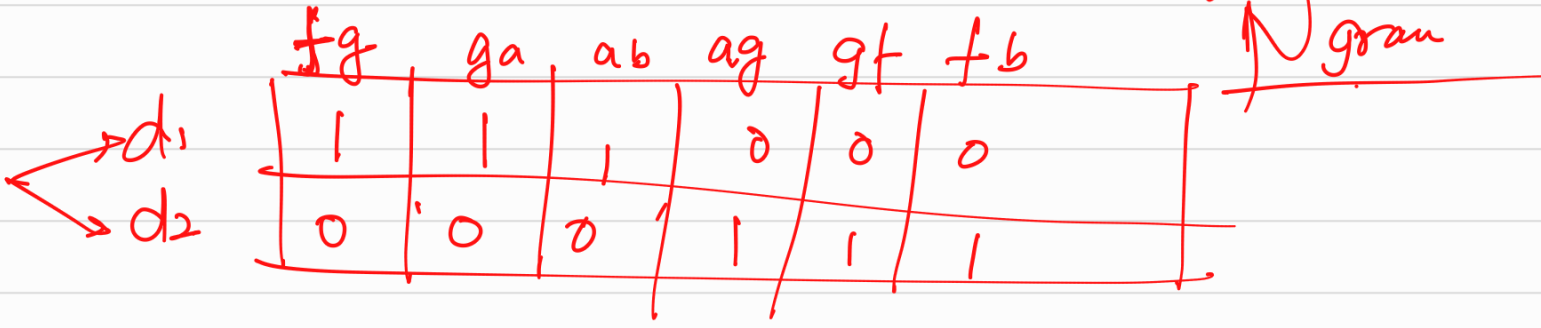


Count

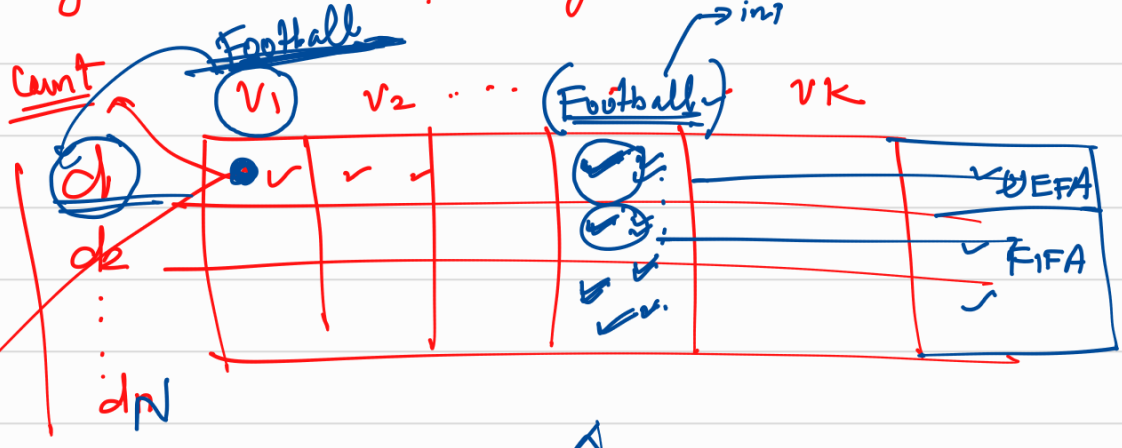
order of the words

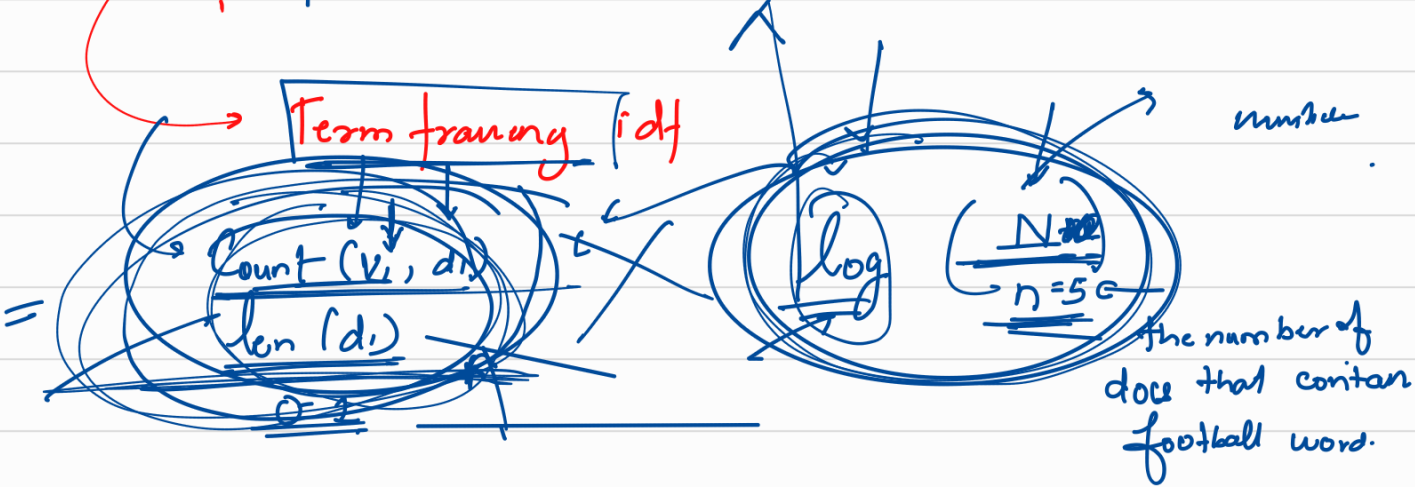




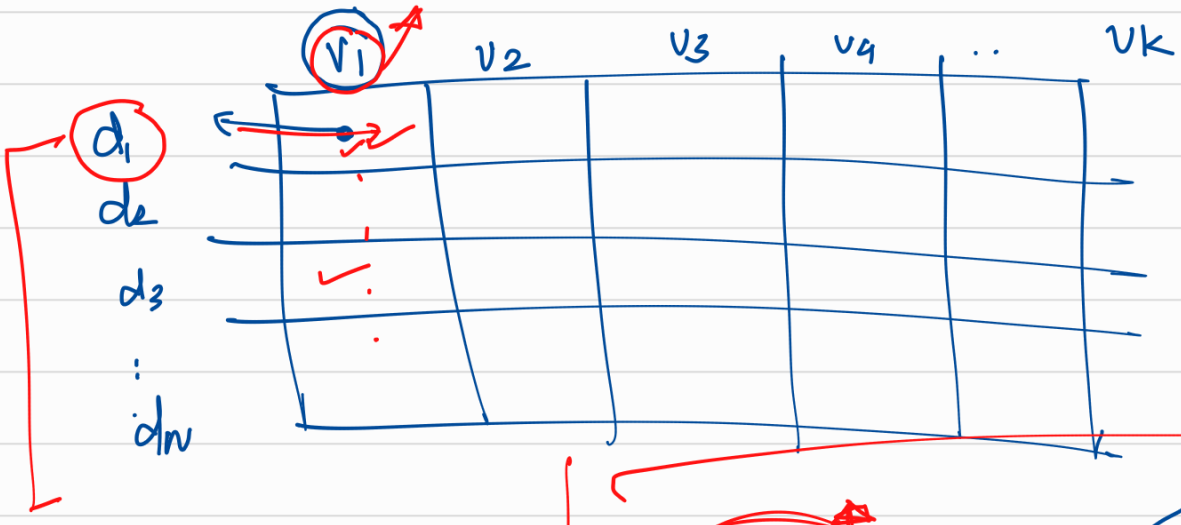


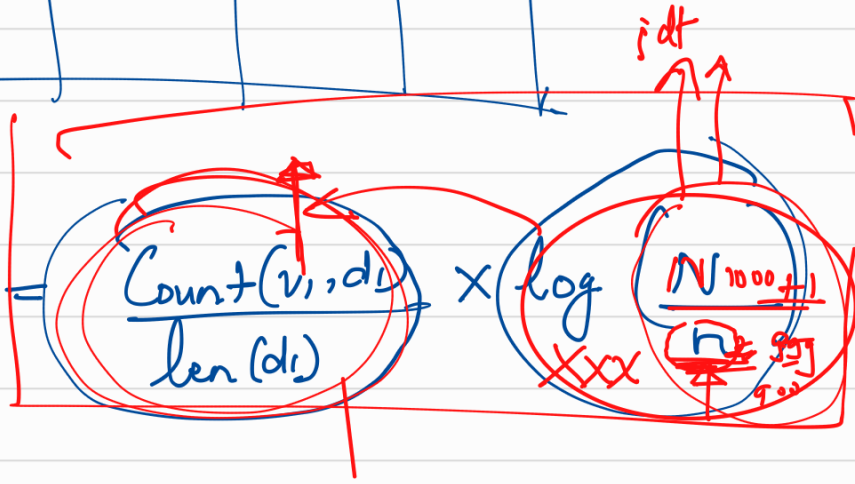
Term frequency inverse document frequency (Tf idf) →

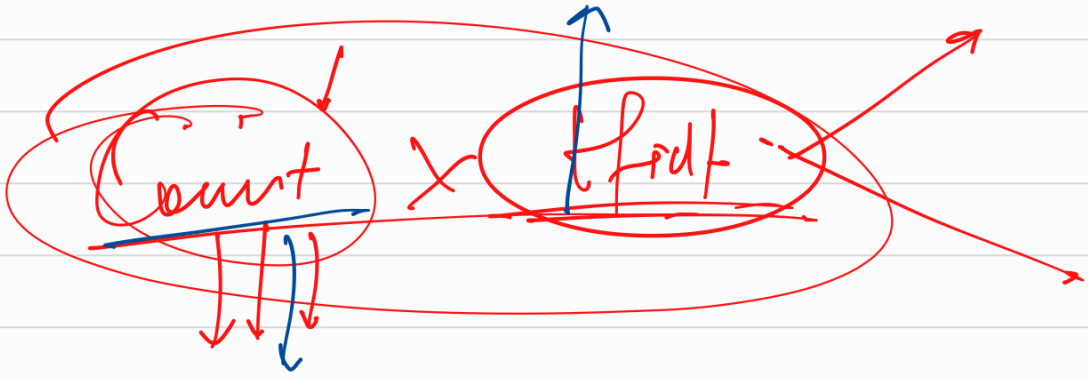


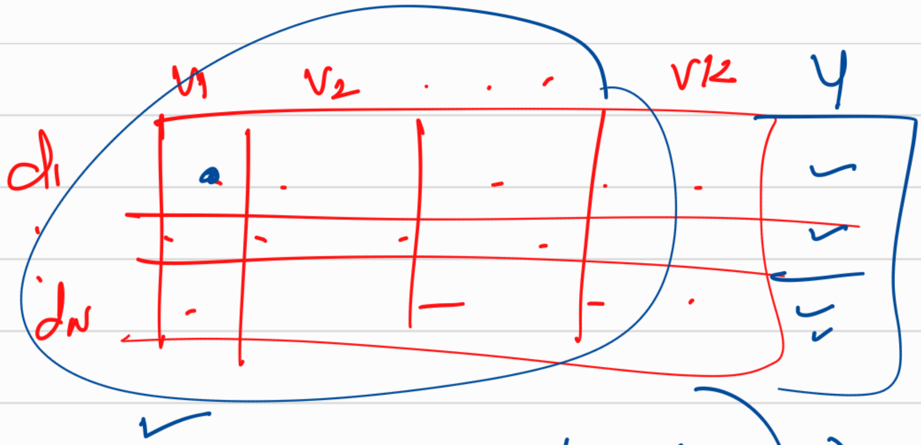


$$\underline{tfidf} = imp(w_i, d_i) \times \log\left(\frac{N}{n}\right).$$





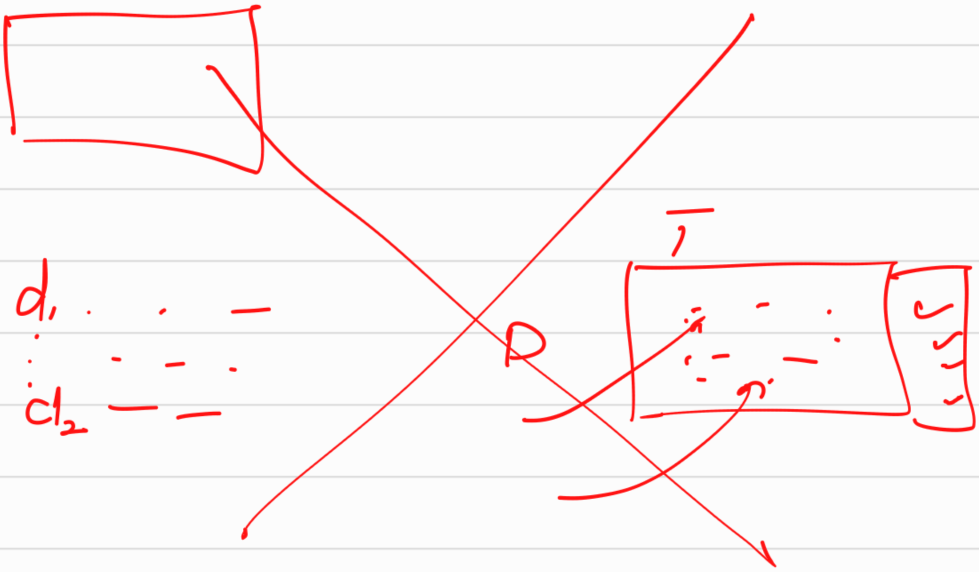


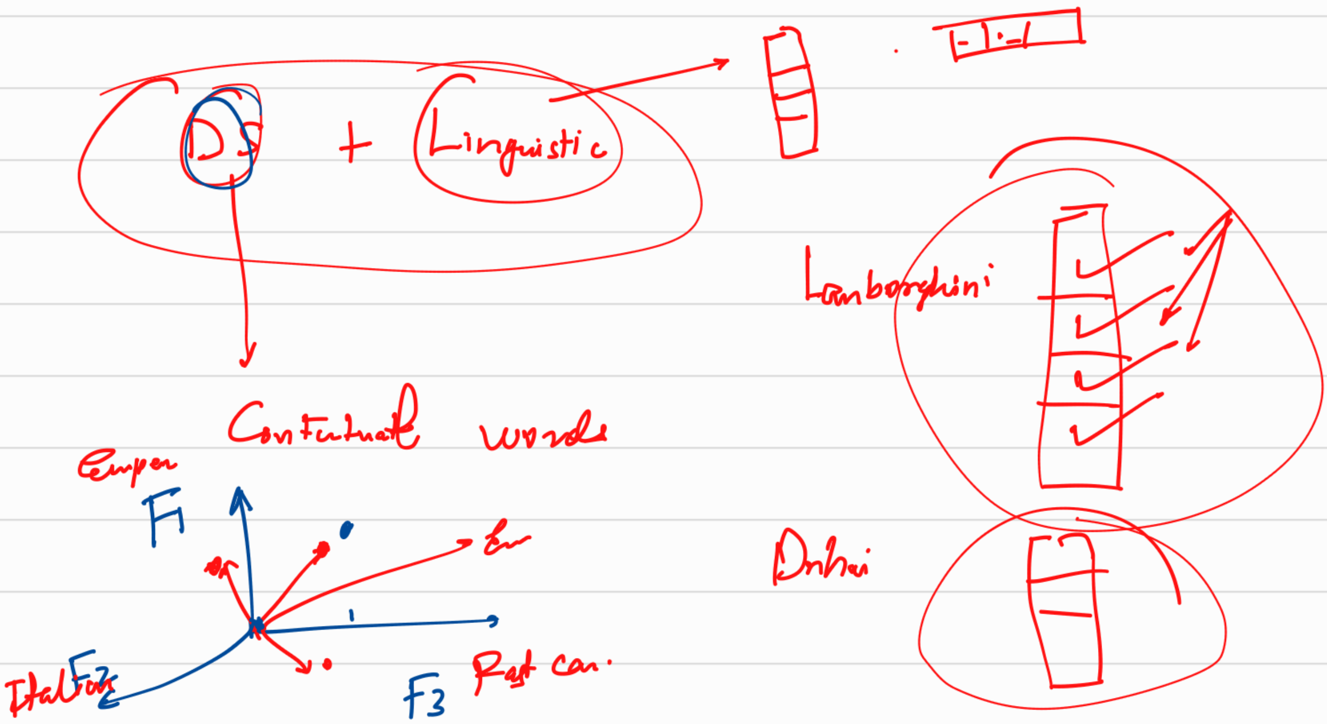

v1
v2
v12
y
d1
dN

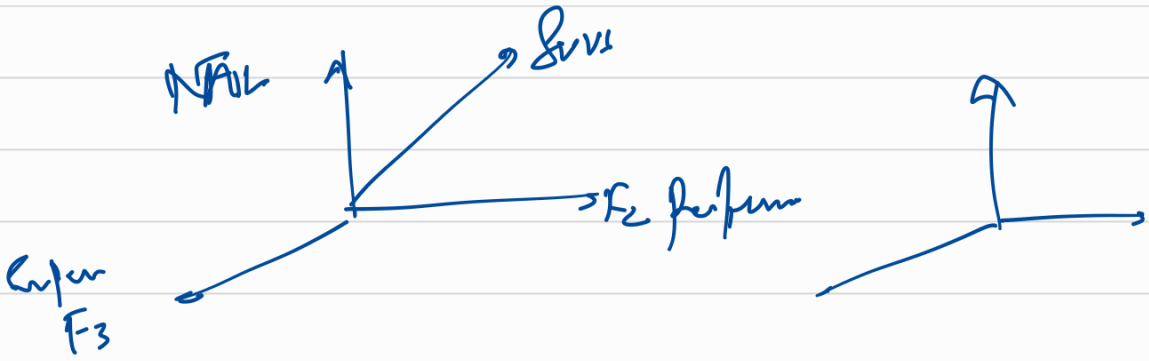
$y_2 + ( )$

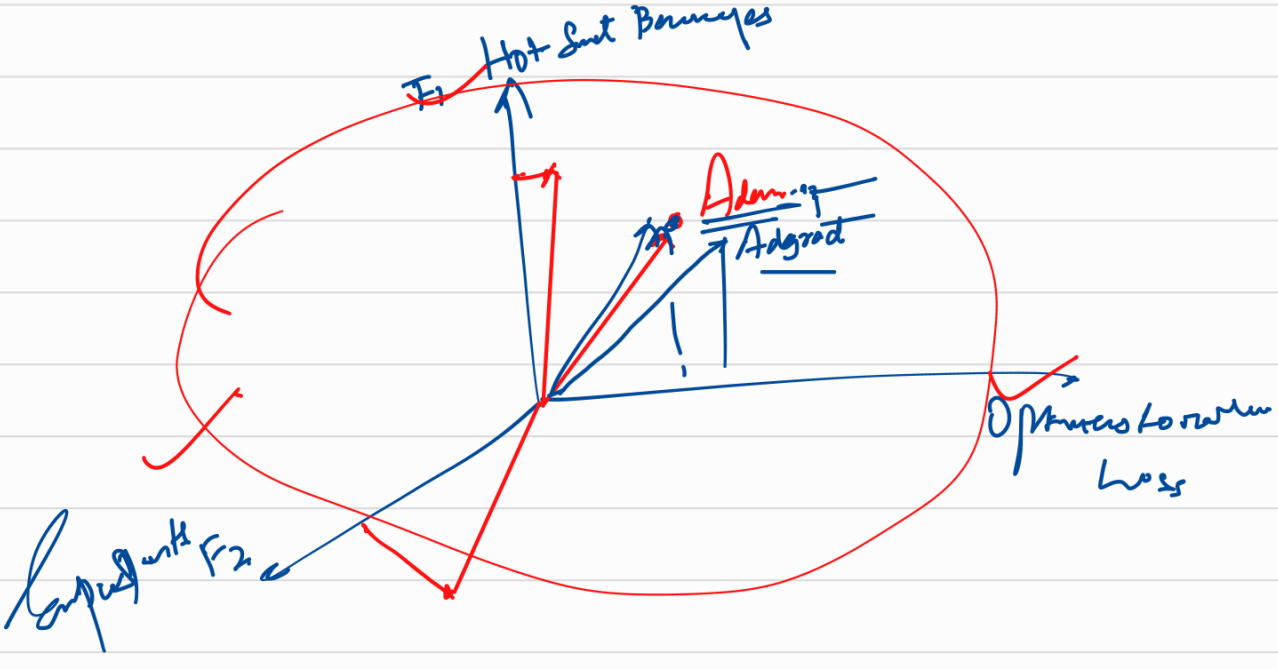
[illegible]

[illegible]

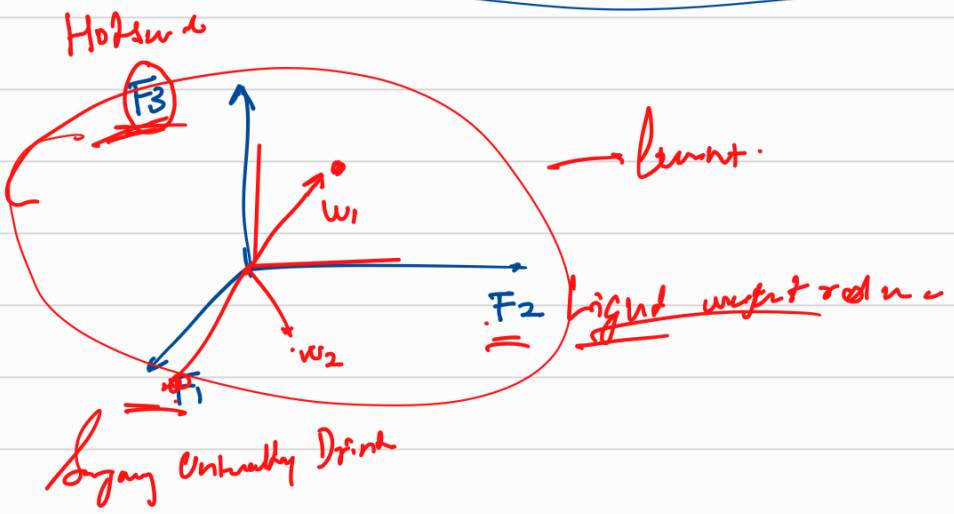

d1
d2
D


DS + Linguistic
Contextual words
F1
F3
Fast car.
Lamborghini
Dubai

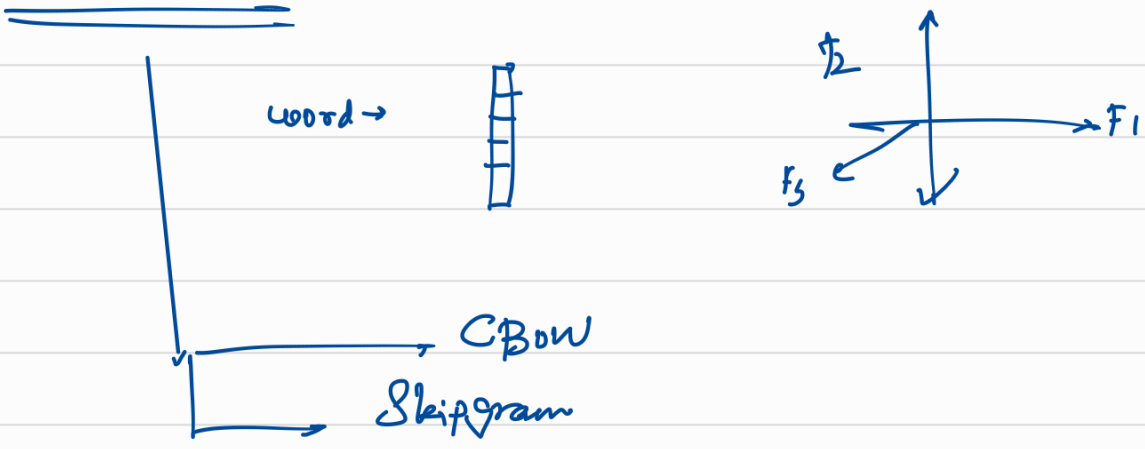



Hot Start Bounces
Adam
Adgrad

Words Similar meaning → high cosine similarity

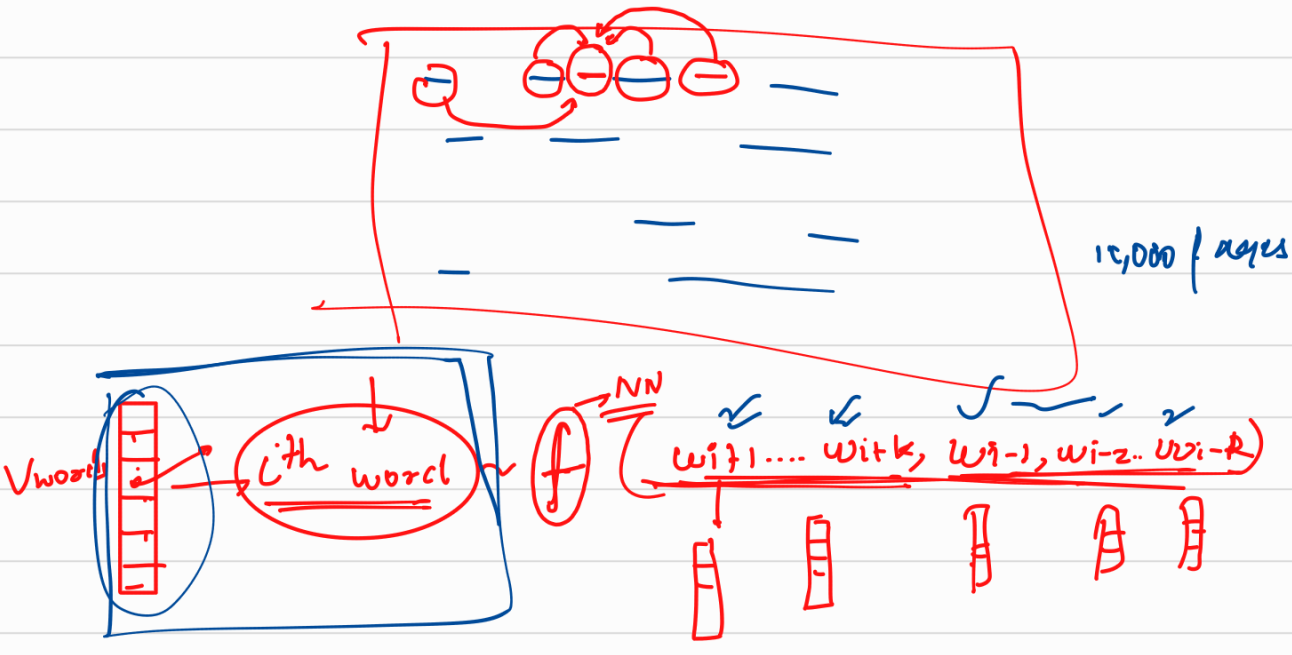


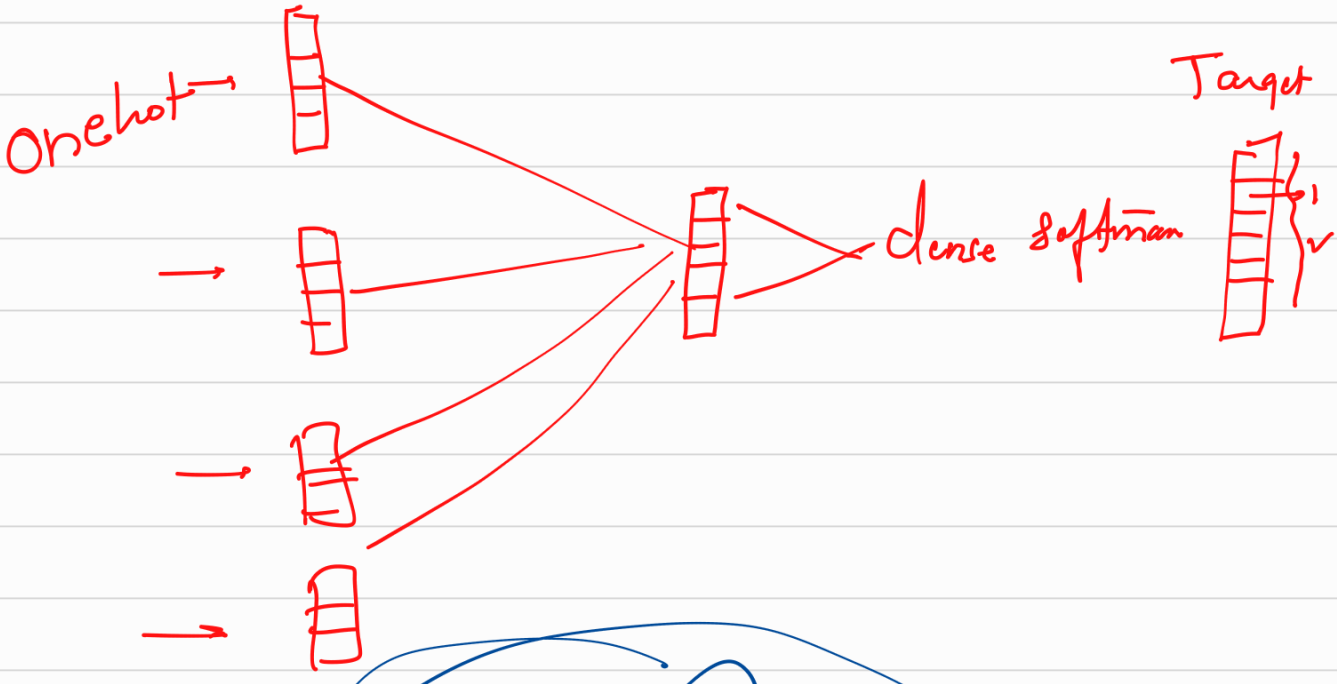
Word2vec

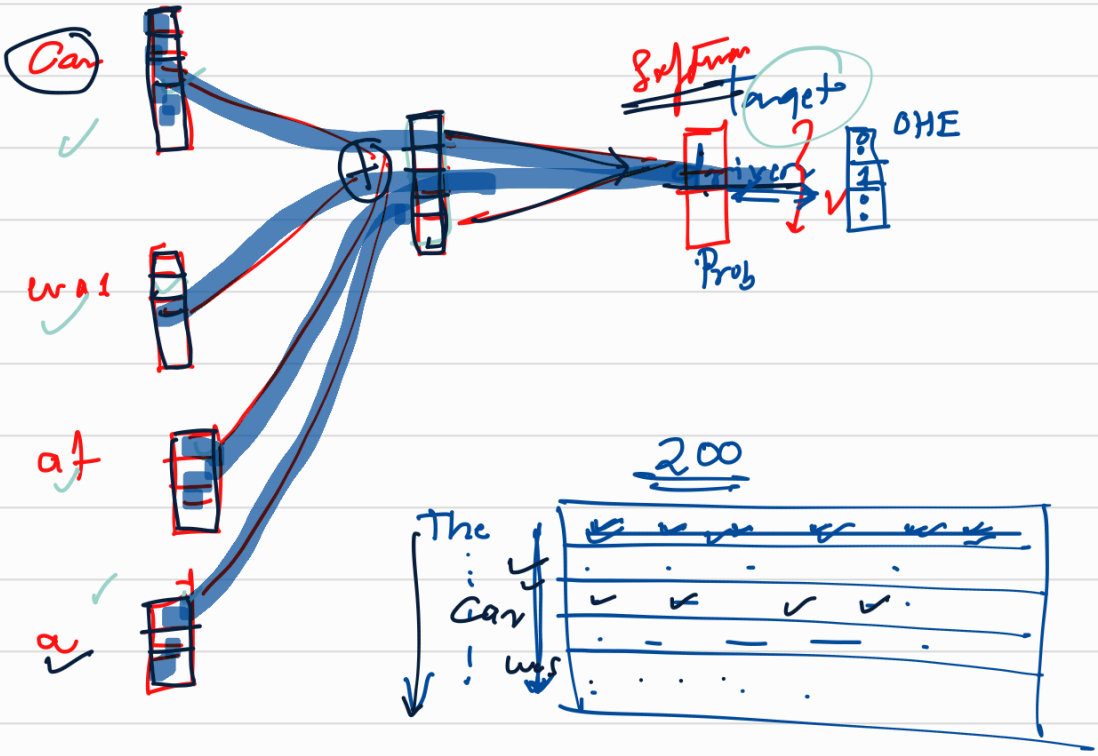


CBOW

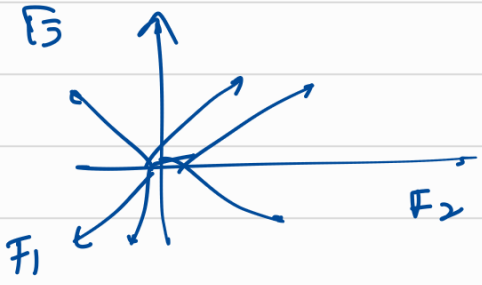
Corpus — Wikipedia / Google news

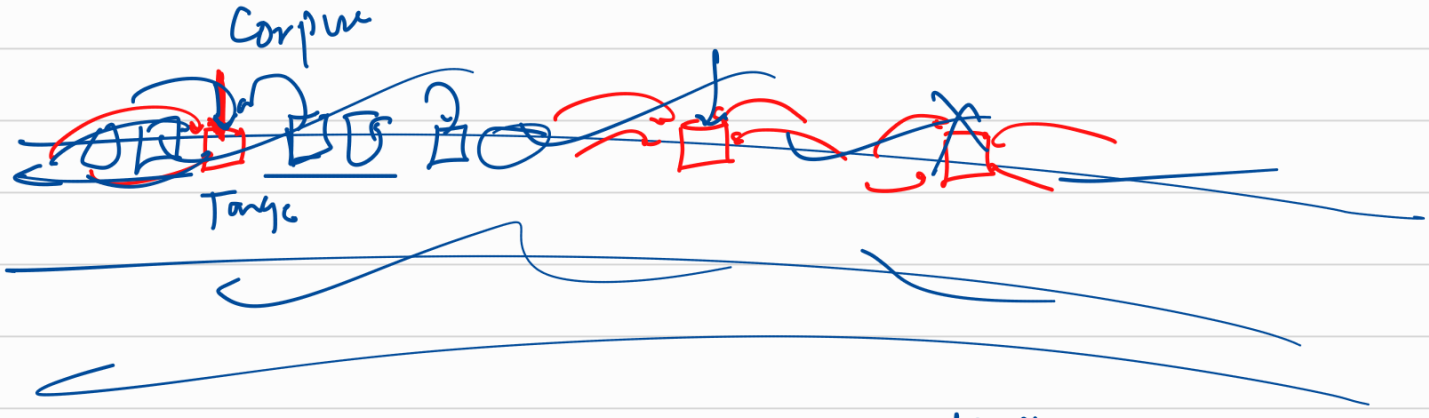




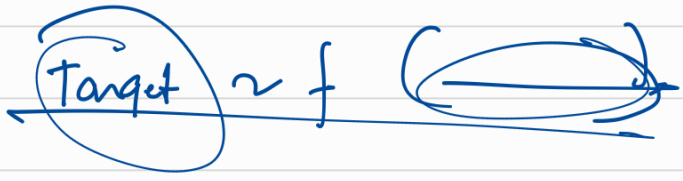


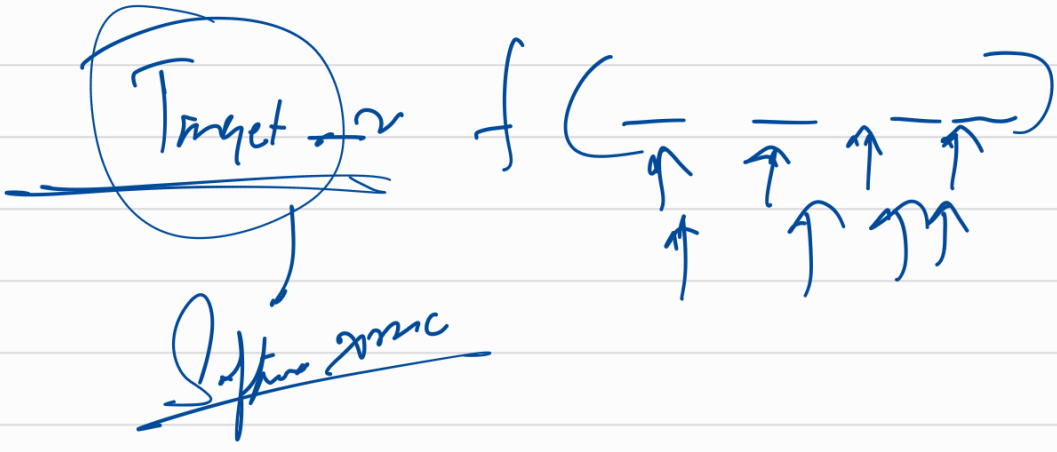
Embedding matrix

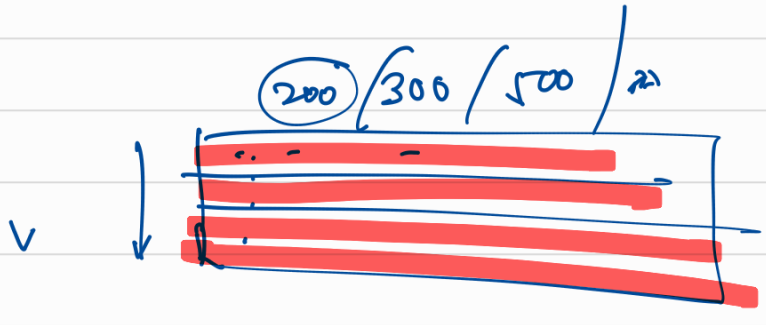


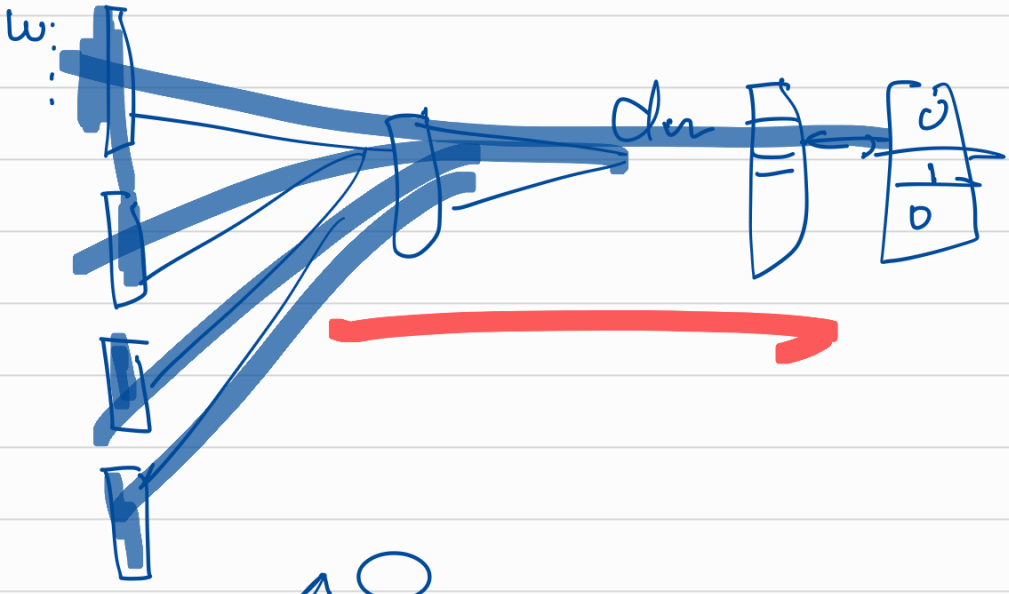


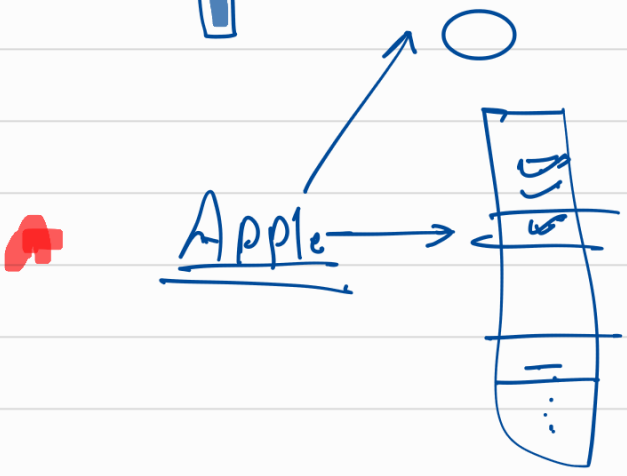
10.000

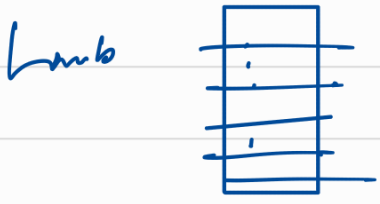







w
den
0
1
0


Apple


Lamb