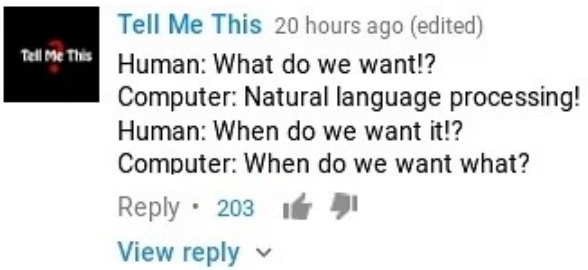
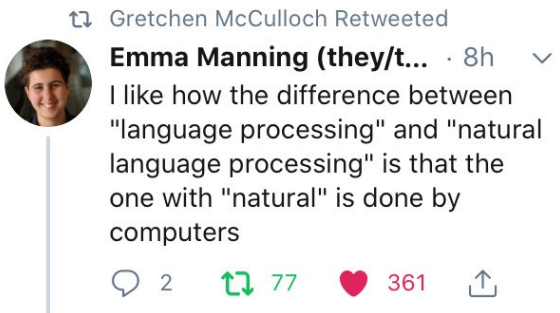
Session 3

Word2Vec





Agenda

1. Matrix Factorization
2. Word2Vec model
3. CBOW, Skip-gram V1/V2

Problem statement

v_1 v_2

good, great, nice → D dimensional vector

$\cos(v_1, v_2) = 1$

v_1 v_3

good, terrible → $\cos(v_1, v_3) \rightarrow$ v. low

TASK: $w_i \rightarrow v_i \in R^d$ using my corpus

PF →

Data matrix → U T V

SVD!

BOW → # words → Sentence

↳ Doesn't care about content / sequence

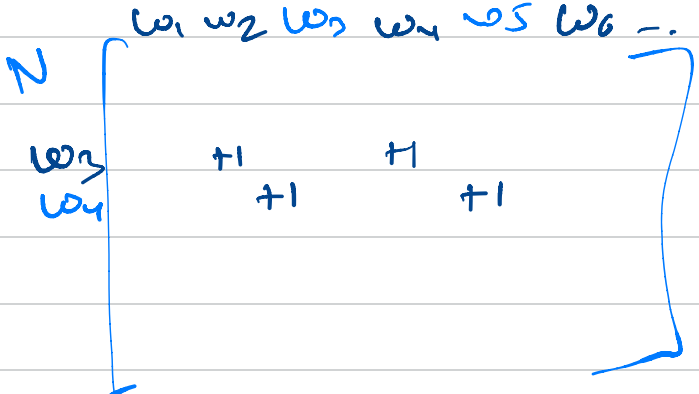
↳ U. Spark

→ Words Co-occurrence matrix

..... w_1, w_2, w_3, w_4, w_5, w_6 ...

Window-size-1

N x N matrix: N → Vocab of my corpus



Today I got to buy a GPU

Yesterday I got Fired

content-window → 2

Today I got to

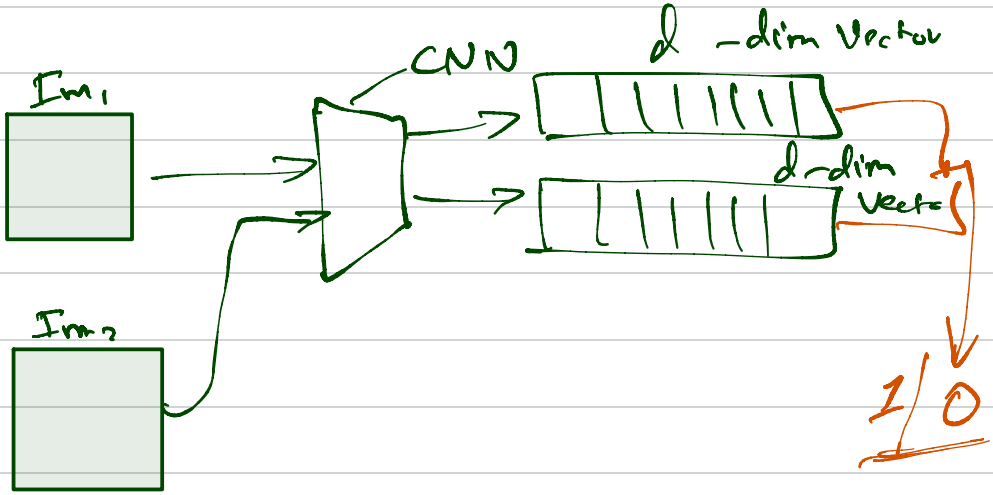
Today	+1	+1
I	+1	+1
got		+1

Co-occurrence → Pros → Partially Preserve context

→ Doesn't Preserve sequence

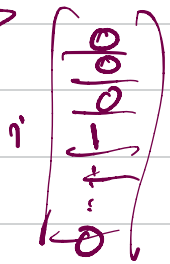
Word Vec

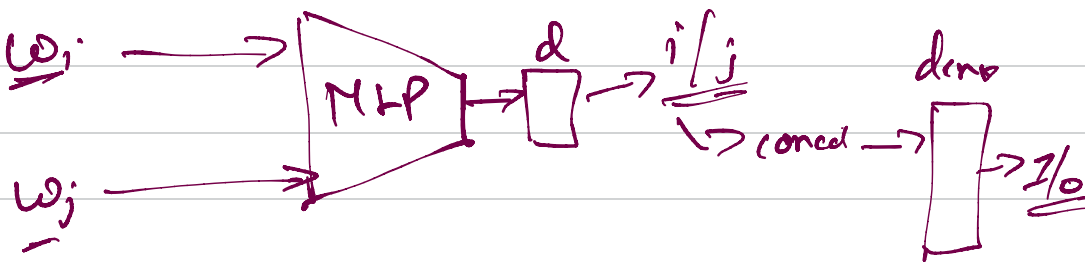
Siamese network



1. Input
2. Output
3. Loss
4. Architecture
5. Loss - differentiable

$w_i \rightarrow$ OHE $\rightarrow$





Vocab $\rightarrow$ [was, the, game, written, horses]

[0, 0, 1, 0, 0]

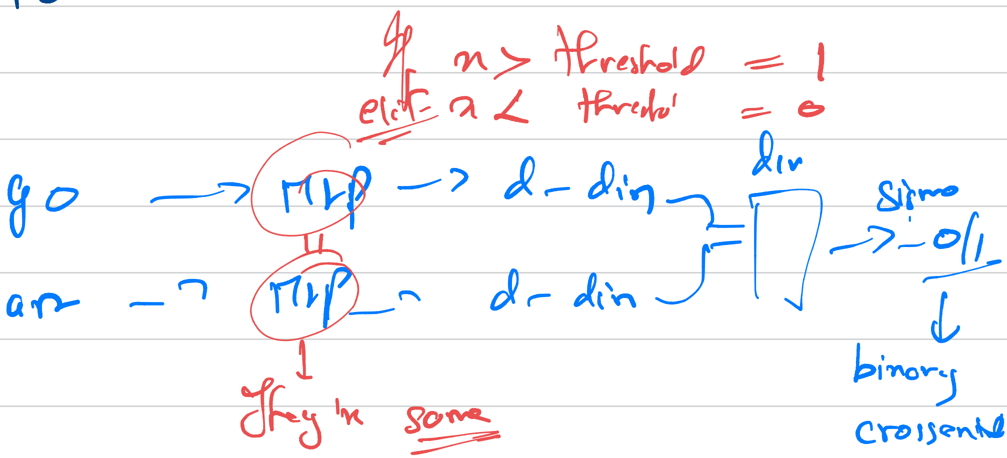
$\rightarrow$ In glomes $\rightarrow$ we measure similarity

$\rightarrow$ What we will do is train the model to learn whether w_i is in context of w_j or not.

N → vocab → entire corp

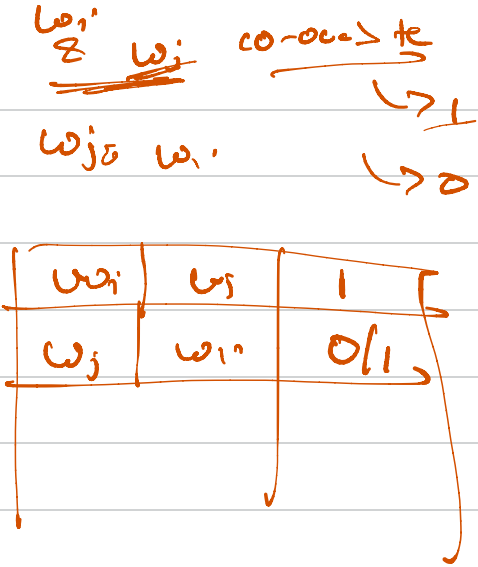
	the	am	go	ride	bb
the					
am					
go		2			
ride					
bb					

N

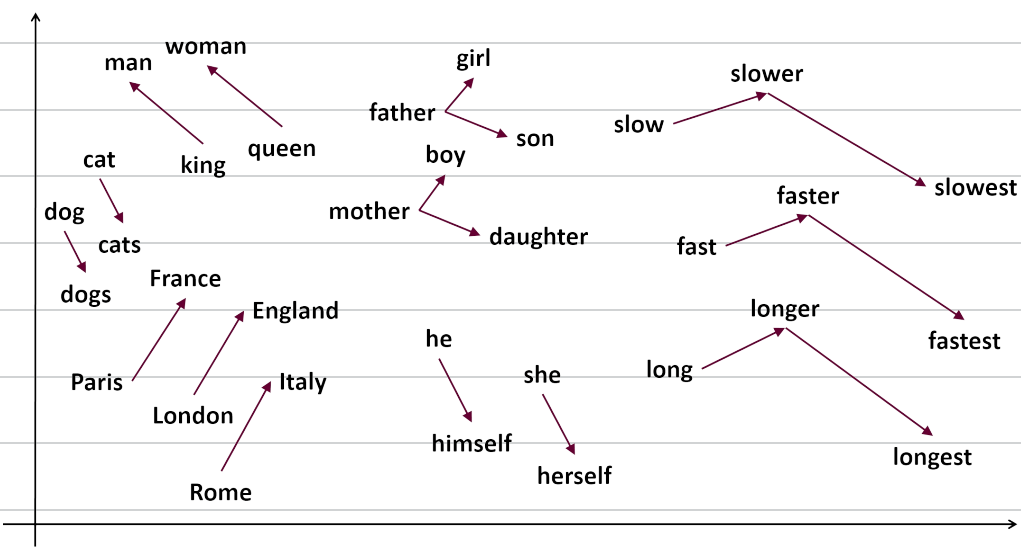


⇒ When training the embedding → to learn the context in which the word is present

only 1 MLP (Neural network)



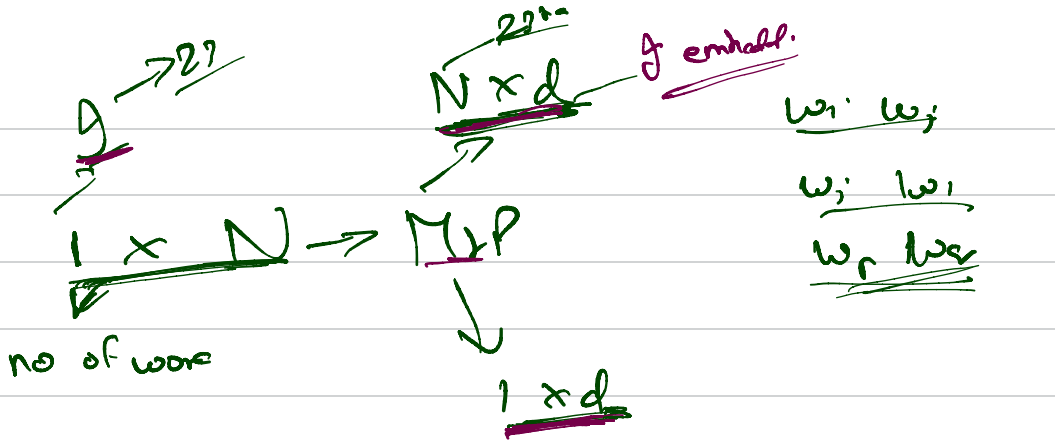
Word2vec in 2d representation



man
King
woman
Queen

(King - man) + Women
= Queen

Royalty



$w_1 \; w_2$

$w_2 \; w_1$

$w_1 \; w_2$

N = 1

$[0\ 0\ 1]$ $\quad 1 \times 3$ $\quad \rightarrow$ we

$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$ $\quad 3 \times 3$

$= [7 \times 9]$

all the words in which we is context wise

$\downarrow$

hit Cosine Simi:

d = 100

512

1024

Input

Cat fridays are the best
Cats in boxes
Cats on heads
In every small compartment
Entangled in threads

WordToVec

Output

Cats	0.434	0.239	0.123	0.934
Fridays	0.126	0.996	0.453	0.124
Boxes	0.924	0.534	0.195	0.845
.				
.				
.				

Word2 Vec → 2013

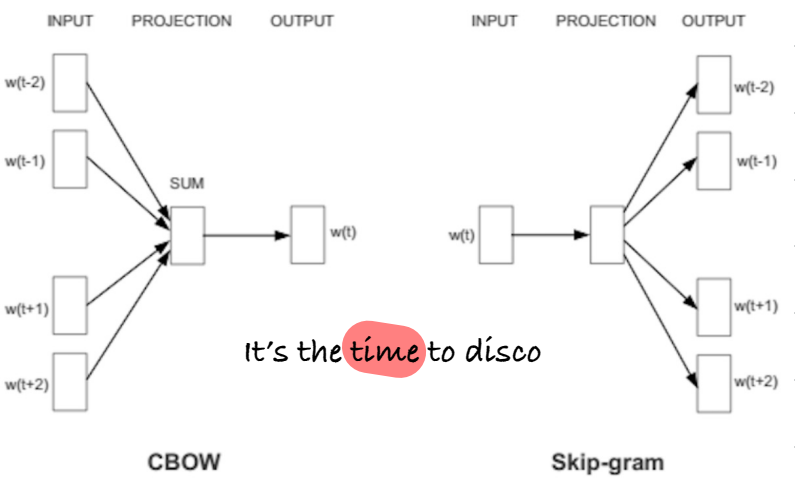
Glove → 2014 ← Co-occurrence matrix

MF

Fasttext → 2016-17 → Phonem level

Skip-gram with neg samplin

Voldemort



w_1 — d

w_{10} — d

100

S_1 → 10 word → Entire Sentence

↳ (10 × 100)

S_2 → 20 word → (20 × 100)

1. Average / centroid
2. Average with IDF → S_1, S_2
   ↳ rare

multiply embedding of word with rarity factor before averaging

5 - Rc

df-covid

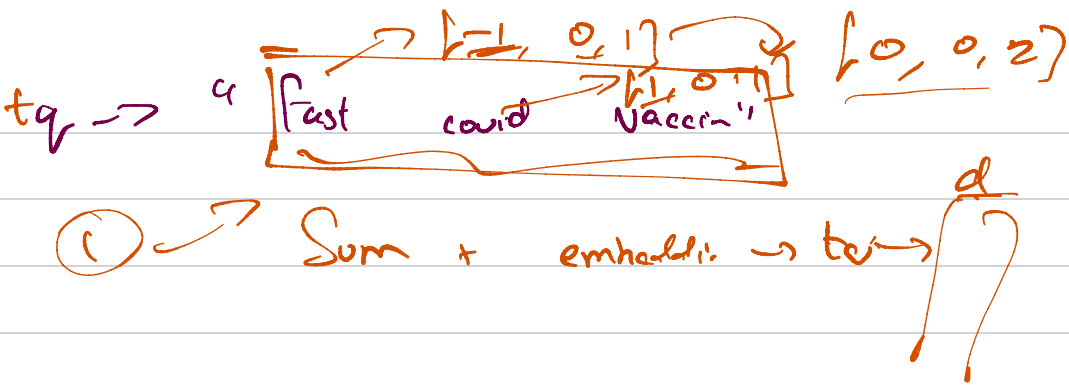
0 → Sum of all embedding of word

1 → " " " a_2

2 → a_1

3 → a_4

4 → c



(1) → Sum + embedding → tq → d

(2) Fast → d with all abstract

covid → d / cos → with all abstra

Vaccin → d / cos → with all ds

Answer it out

Fast → a_1 (0.3), a_2 (0.2), a_3 (0.4)

covid → a_1, a_2, a_3

0.2 → 0.4 | 0.1 → 0.15 | 0.1 → 0.25

$a_1 > a_3 > a_2$