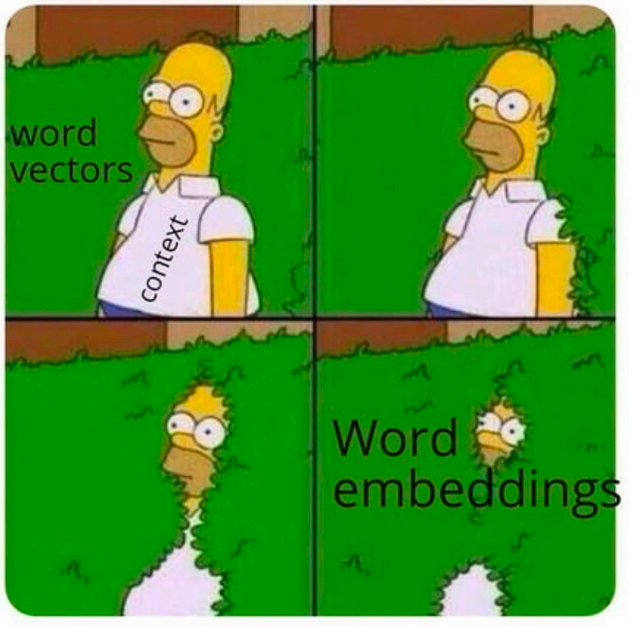
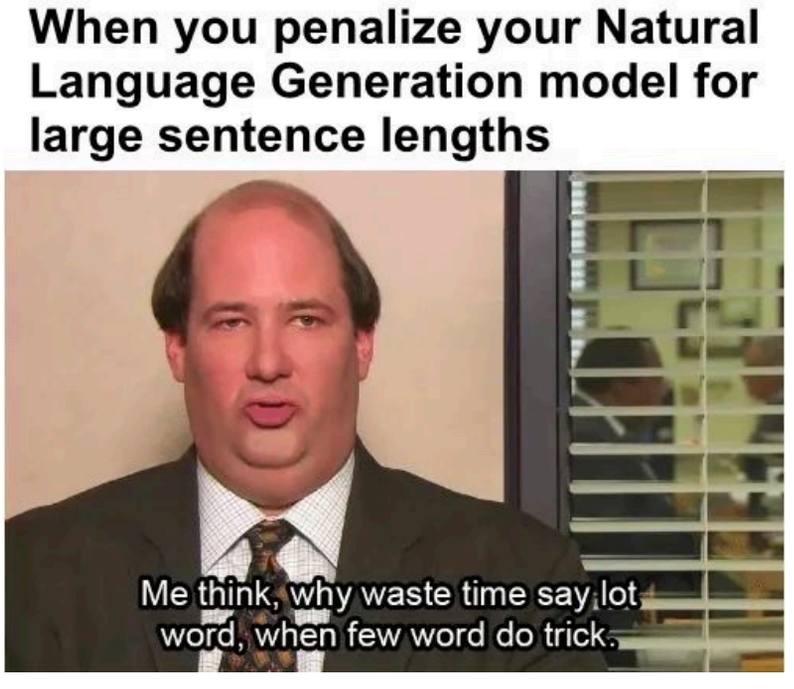
April 26, 2023.

DSML : NLP module.

Word Embedding : Word 2 Vec

Class starts @ 9:05





Recap:

* Corpus: Collection of documents.
* Documents: A collection of sentences.

→ Agenda (NLP): Find good vector representations for documents.

→ Bag of Words: Document → Vector.

→ TF-IDF: Document → Vector.

$= A \cdot B$

Agenda :

* Fundamental change : Words → Vectors.

* Why? We want vector representations which capture semantics (meaning).

* Plan: Let's build / invent techniques together !!
  (We have all pre-requisites to figure this out)

* But first... The business case !!

corpus. d_1 — [ ]

d_2 —

⋮

d_n —

↑ documents are paper abstracts.

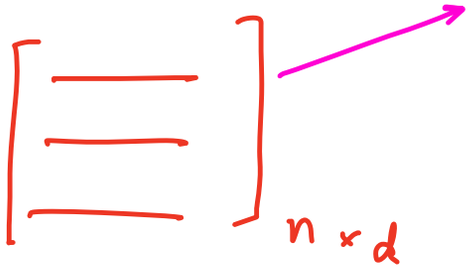
] d-dim vector.

↑ the size of the vocabulary.

Recap: Rec. Sys.

Movie recos.

M. F



$n \times d$

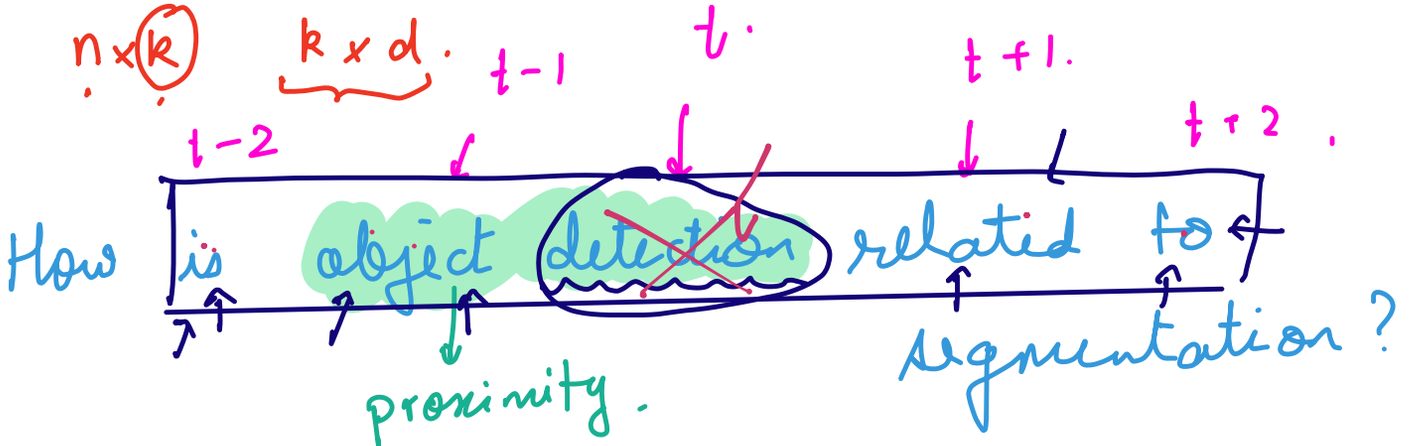
B.O.W representation.

$R = A \cdot B$

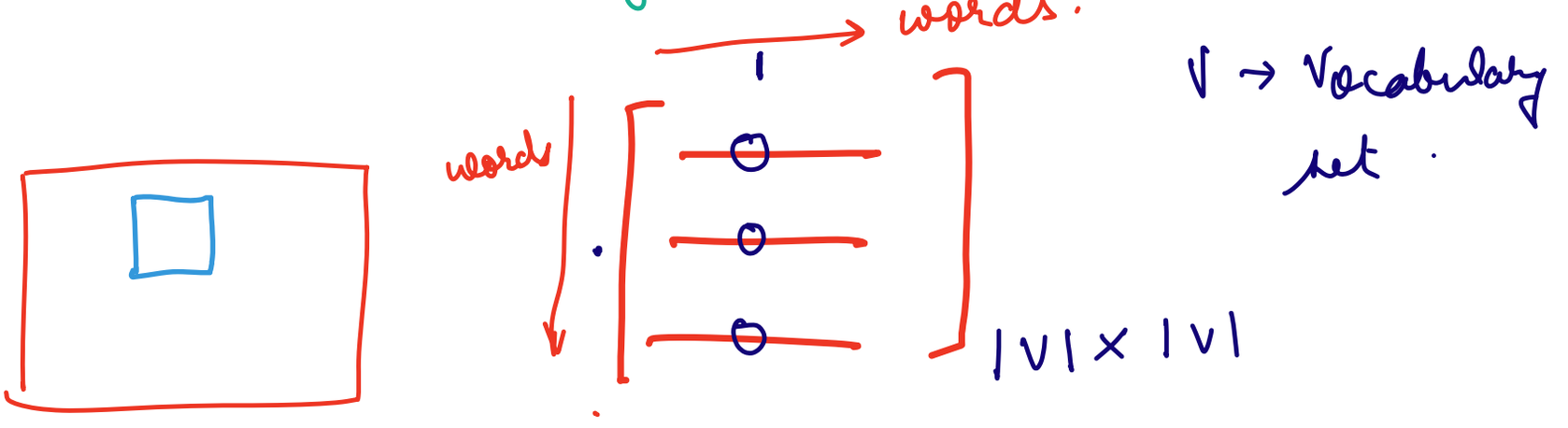
A: $n \times k$; B: $k \times d$.

$R = A \cdot B$

$n \times k$ $\quad$ $k \times d$



ngram $\Rightarrow$ $n = 2$.



V → Vocabulary set.

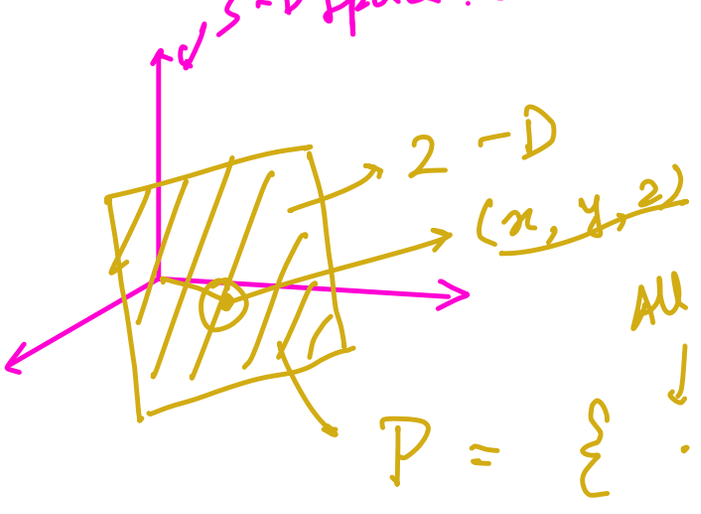
$R_{|V| \times |V|} = U \Sigma V$.

Singular Value Decomposition method.

SVD.

$$A_{m \times n} = U_{m \times k} \; \Sigma_{k \times k} \; V_{k \times n}$$

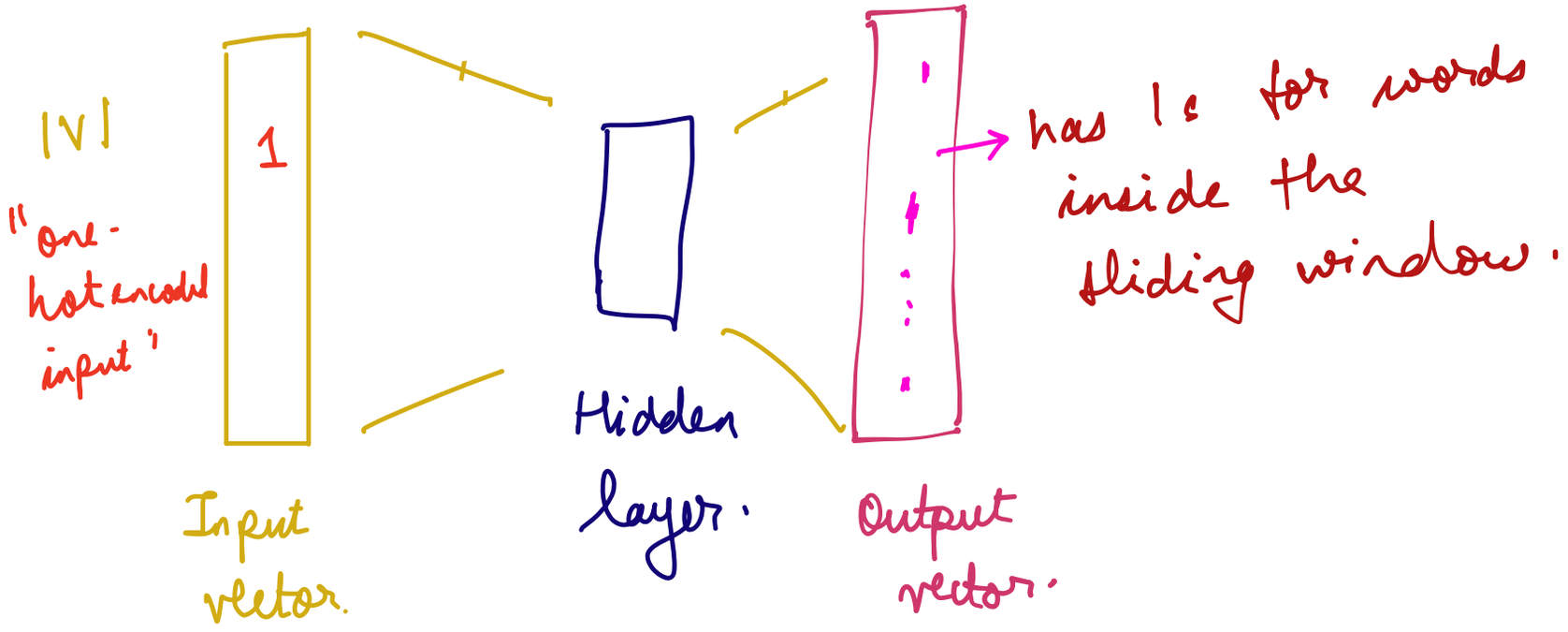
Inherent dimensionality of A.



Neural networks solution.

↳ Autoencoder solution for P.R. !!

Word 2 Vec. → Name given to autoencoders which gives word-embeddings.



1] Fill-in-the blanks?

ResNet is a type of Convolutional Neural Network.

Continuous Bag of words strategy for Word2Vec.

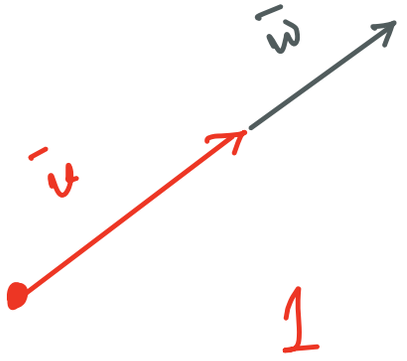
2] Write an essay on ... → skip-gram strategy for Word2Vec.

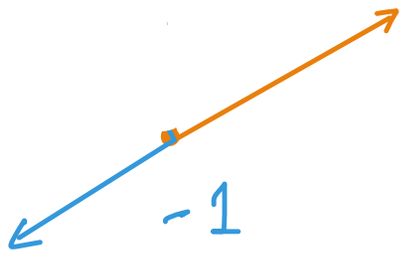
Convolutional.

$|V| \approx 10^5$

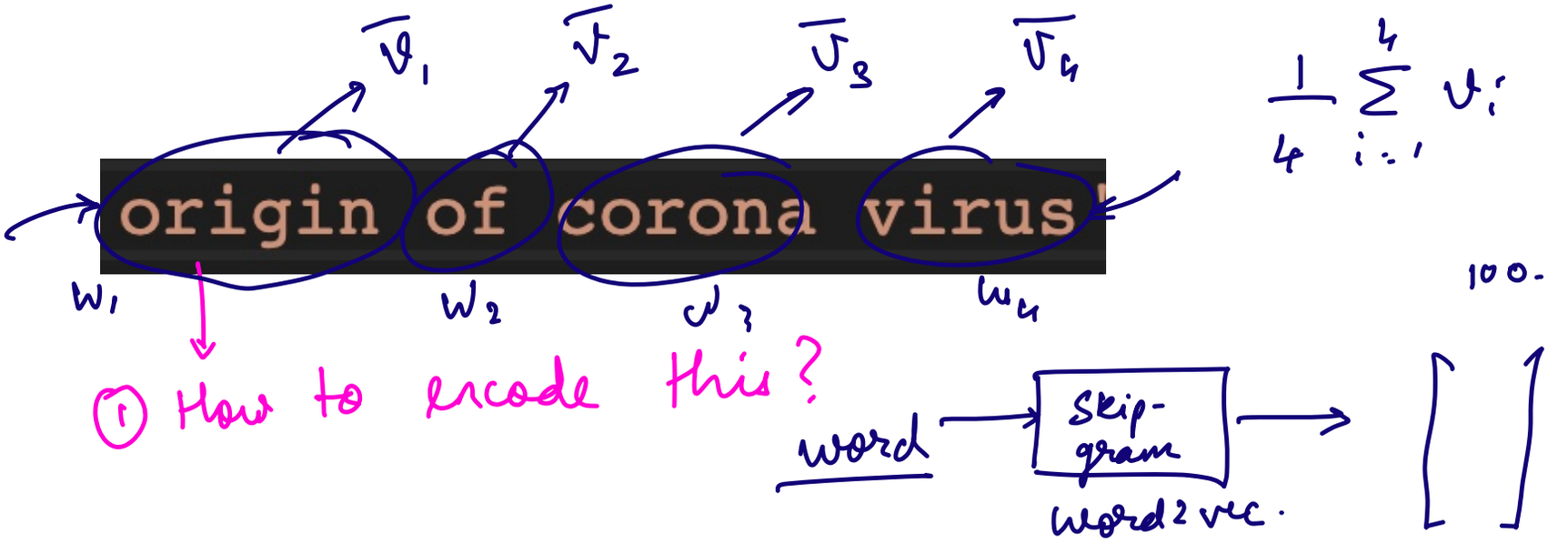
$$\text{Softmax} = \frac{e^{z_i}}{\sum_{k=1}^{|V|} e^{z_k}}$$

$O(|V|)$ → This will take too long!!





$$\arg\max_{\theta} \sum_{(w,c)\in D} \log \sigma(v_c.v_w) +$$


origin of corona virus
w_1 w_2 w_3 w_4
v_1 v_2 v_3 v_4
$\frac{1}{4}\sum_{i=1}^{4} v_i$
100.
① How to encode this?
word
Skip-gram
word2vec.
② How to encode these?
d_1
d_2
d_{n-1}
d_n