NLP class 3

Tuesday, 19 January 2021 6:26 PM

Problems :

→ Classification ↳ Spam filter, Sentiment Analysis

→ Regression

→ Ranking

→ Machine Translation (Source Language → Target L)

→ Text Summarization ↳ Mono-lingual Sum. ↳ Multi "

→ NER ↳ Entities, ↳ Code Mixed

Name, Organisation/Place, location etc.

→ Part-of speech Tagging (POS)

→ QA (Question <Who, What, Where> <Why, How>)

→ Image Captioning. <Image, Text> <Text, Image>

→ Image Summarization

→ Text to Image → Text → Image. (Art)

→ Text to speech Recogni.

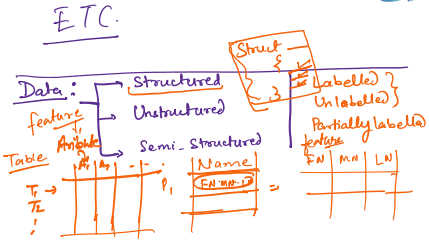
→ Speech to Text Translat.

→ Speech Data/Audio data → Methods of Text you can apply

→ Video Summarization → 25 fps, 1 to 5 sec, Seq of Img → Summ in Text, Video

→ Embedding Audio to Videos (Silent)

ETC.



Data

Sam $\langle f_1^1, f_2^1, \dots f_d^1 \rangle$ rows → data points | samples

Columns → Features | Dimensions

Sample N $\langle f_1^N, f_2^N \dots f_d^N \rangle$

2-features

$S_1 \to \langle f_1^1, f_2^1 \rangle$

$S_2 \to \langle f_1^2, f_2^2 \rangle$

$S_3 \to \langle f_1^3, f_2^3 \rangle$

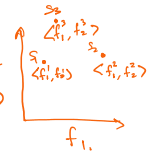


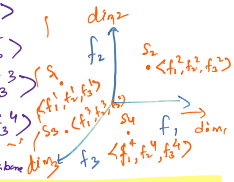
Table 3-features

$S_1 \to \langle f_1^1, f_2^1, f_3^1 \rangle$

$S_2 \to \langle f_1^2, f_2^2, f_3^2 \rangle$

$S_3 \to \langle f_1^3, f_2^3, f_3^3 \rangle$

$S_4 \to \langle f_1^4, f_2^4, f_3^4 \rangle$



2. i) pre processed/Database

→ Unstructured Data 96%

↳ Text, Audio, Image, Videos

→ Don't have the track of feature locations

S1 → Mohan is reading a book ⟩ [ ]

S2 → Reading a book is very hard ⟩ [ ]

→ Semi-Structures → Structured + Unstructured

Supervised	Un-Supervised	Semi-Supervised
Labelled $\langle S_i, y_i \rangle$ $y_i \in \{0,1,2\}$	Unlabelled $\langle S_i \rangle$ Infer y_i Problem	Partially Labelled (Labelled (2%), Un-labelled (98%)) 2% Tagging

Problem $\langle S_6, \hat{y}_6 \rangle$

$y_i \to R$ target $\langle S_i \rangle$ $\langle y_i$: Summary$\rangle$

$\langle S_i \rangle$ $\langle y_i$: Translate in Target L$\rangle$

Assignment 1: find out the publicly Available data in all the categories.

21/01/2021

↳ Source for the datasets:

↳ Kaggle web page ✓

↳ UCI Repository ✓

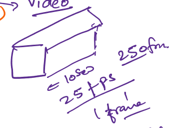
↳ SVM datasets ✓

↳ Twitter API ✓ Raw

- Kaggle web page ✓
- UCI Repository ✓
- SVM datasets ✓
- Twitter API ✓ Raw
- Facebook pages → API Raw
- Blogs → Product/Service Reviews. Raw
- Synthetic Datasets, Raw
  - Generate the Data.
    - Raw
    - Structured

→ Sequence Data:
- Text Data
- Signal Audio/Speech data
- Gene/RNA/DNA sequences
- Video

No sequence Preserved (Data)
- Images } Problem specific



t = 0 t = 1 t = 2

Step . step . . . . . .

Order .

$\frac{1}{25}$ sec

HAPPY

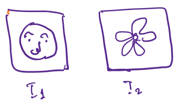
HAPPY

HPAPY ✗

I. am teaching NLP
Course.

I_1 → HF (Yes/No)

I_2 → HF (Yes/No)



M (I) → HF (yes/No)

eyes, nose

22/01/2021 Lab session

→ Get the Data Set
- Crawl (API)
- Download the News → Articles
- Audio/Speech

Unstructured Data — Raw form

→ Folder NLP_Lab → LabAssignment1
  ↳ raw_Data → raw Data

Browser / write some script in any of favourite programming lang.

Unstructured
⇓ Transforms
Structured Forms

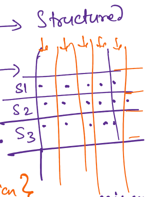
{Data preprocessing}

Unstructured ⟶ Structured

S1: Set of tokens →

S2: Set of tokens ⇒

S3: Set of tokens ✓



Tokens:

{Feature Extraction}

→ Noise Removal

→ Normalization
- range → 0-1 Min-max / mean-Variance
- {Stemming / Lemmatization}

X ⟶ [Algo model] ⟶ Y {goodness of the model}

[Group Name: Members]

→ Get the data < Set problems on this data
Lab1 Part (a)

→ Part (b) } Data cleaning + Data Preprocessing

25/01/2021:

Datasets → Forms

Label

X ⟶ [AI Mod] ⟶ Y O/P

Datasets → terms

Label

X → [AI model, learnable parameters θ] → Y (o/p)

Hyper Parameter

Structured as input — we have unstructured?

Unstructured Data → Structured

Basic Approaches

Text:
- S1: He is reading a book. — Bigrams: <He is>, <is reading>, <reading a>, <a book> — 4
- S2: Can you come here? — 3
- S3: Ram is playing a Music. — 4
- ⋮
- S4 S_m: Shyam is Happy. — 2

Data (Corpus)

= 13 bigrams ($^4P_2$)

STATS: # Samples, # Unique Words
- one pass with Data
- Sequential pass | parallel
- Exploit the H/W

Bag of Words representation

Dictionary: 14 — Word = token

	0 He	1 is	2 reading	3 a	4 book	5 Can	6 you	7 Come	8 here	9 Ram	10 Playing	11 Music	12 Shyam	13 Happy
S1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
S2	0	0	0	0	0	1	1	1	1	0	0	0	0	0
S3	0	1	0	1	0	0	0	0	0	1	1	1	0	0
S4	0	1	0	0	0	0	0	0	0	0	0	0	1	1

Token. Word ≡ Feature/dim

Lost: Context, ordering

S1: < (0,1) (1, 2), (2, 3) . . . . >

S2: < >

Uni-gram, bi-gram model,

2 words: $\langle (0, w_1), (1, w_2) \rangle$

$\langle w_1\, w_2 \rangle$, $\langle w_2\, w_1 \rangle$

$\langle w_1\, w_2 \rangle$ $\langle w_2\, w_3 \rangle$ $\langle w_1\, w_3 \rangle$ $\langle w_3\, w_2 \rangle$

$\langle w_2, w_1 \rangle$

n - words: $^nP_2 \Rightarrow n(n-1)$ $O(n^2)$

$n + n(n-1)$ ($^nP_2$) $O(n^2)$, $O(n^3)$

≃ 1m + 1m × 1m

$^nP_1 + {}^nP_2 + {}^nP_3 + \cdots + {}^nP_k = ({}^n_2 ?)$?

Theory Assignment 2

Worst Case Size of Dictionary
- Large No. of columns in the data
- High dimensions & Large # features
- Highly sparse Data

Words — base form
- (Prefixes, Suffixes, rules) — Fast — Stem
- Stemming (Rule Based)
- Lemmatization
  - Database
  - root, - - -
  - WordNet . . ?

Morphological variations

Mohan went School

Morphological Analysis

Mohan Went School

Root, - - -

go

Reading a book is a good habit

Wordnet

Time complex process

root + POS

Semantic Enhanced meaning

Went (Verb, Past)

⇒ go + Verb + Past

Prefix + Base + Suffix

Parsers

Lab Assignment: $X = \{x_i\}_{i=1}^{n}$

56

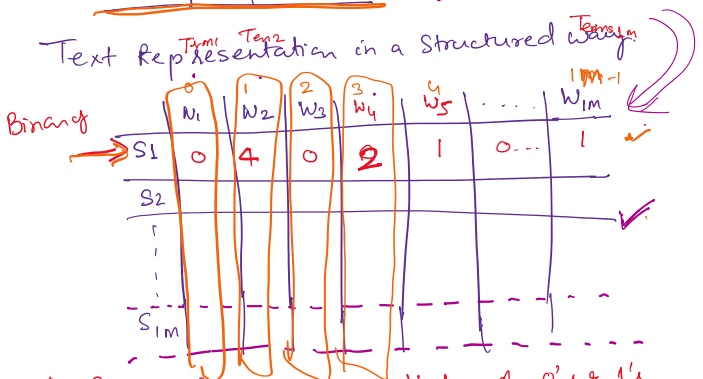
five Six

number

28/01/2021

Unstructured

Text Representation in a Structured way



1. Binary Representation → Vector of 0's & 1's
2. Term Frequency Rep. → Vector of Integer including zeros.

↳ Uni-gram {each word is a term}
↳ Bi - - - {each pair of a words is a token}
⋮
↳ K-gram {group K-words will form a token}

Highly sparse !!

L K-gram → group K-words ...

Highly sparse !!

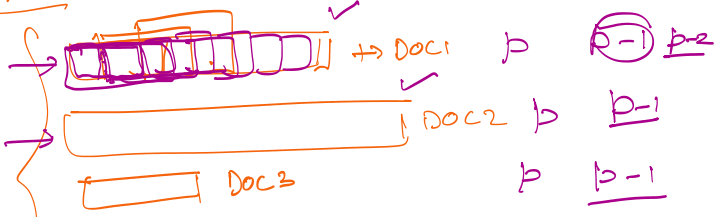
S1: <1, 47> <3, 2> <4, 1> <|m-1|, 1>

S2: <3, 2> <4, 47> <2, 3>

SVM format: <fid_i : value> <fid_{i+1} — — —>

3-gram

window = 4



DOC1 → p → p-1 → p-2

DOC2 → p → p-1

DOC3 → p → p-1

"N"

Average length of a document is 'p'. words

Dictionary Size w.r.t Uni-gram is 'n'

tokens = 'n'

Docs = N

Unigram model

Expected # token = Np if each sentence has unique words.

$NP = n$

$n \leq NP$

$n_b \leq N(p-1)$

$n_t \leq N(p-2)$

$n_k \leq N(p-k+1)$

→ Raw words/terms

→ Preprocess it → Reading → Read

Went → go

Playing → Play

Affixes, base word/root word + POS

Form

Reading → Read + Present Continuous

VERB

Stemming → Role defined

Stemming → Rule defined → VERB

Lossy

defin + ed

Pre + stem + suffix

past

Time complexity

Lemmatization:

WordNet

base	Predefined		
define	defining	defined	Pre...

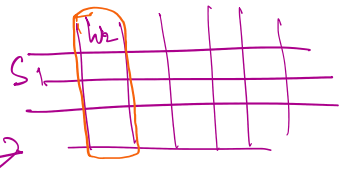
→ Read Data

→ Tokenize the Data

→ Remove stop words.

→ Stemming / Lemmatization

→ Vector Representation → Dictionary Create the Data

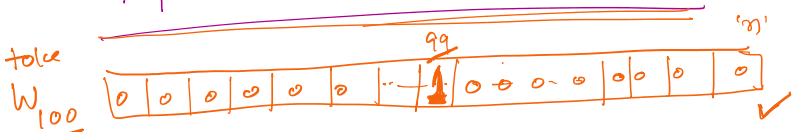


Lab session

→ ML - Sklearn

Dictionary of 'n' tokens

→ Represent a token in Vector form

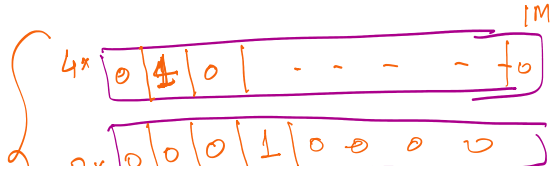


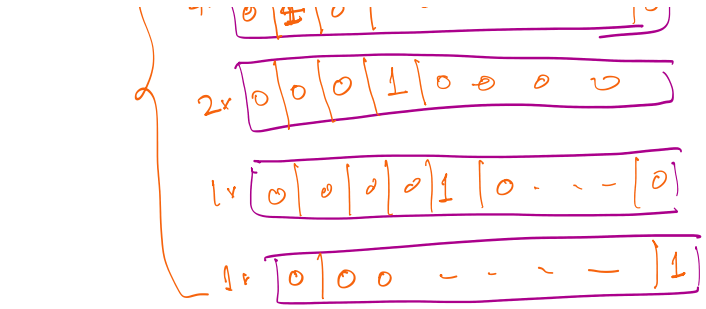
one - hot - vector Representation of a word

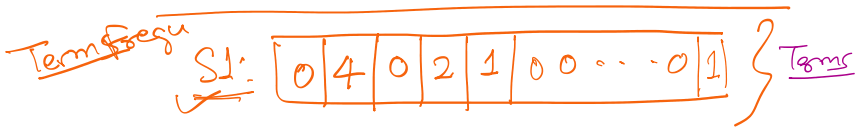
{Similar Words / Group Words} How will you do that? ??

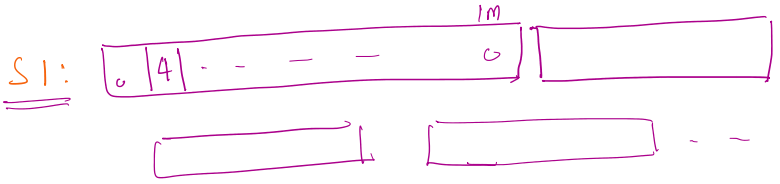
S1.. W_2 W_4 W_5 W_{1M}

S1: Sum of Term Frequency

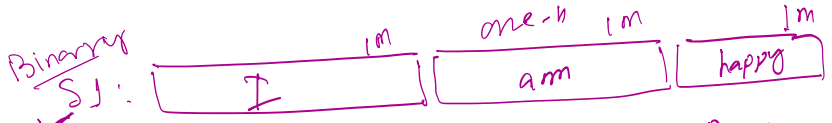




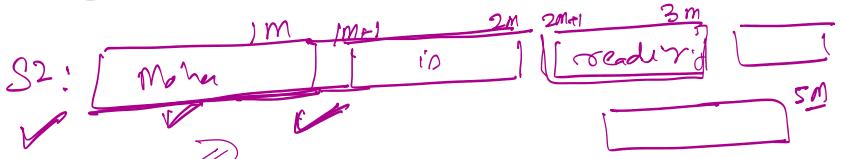




S1: I am happy

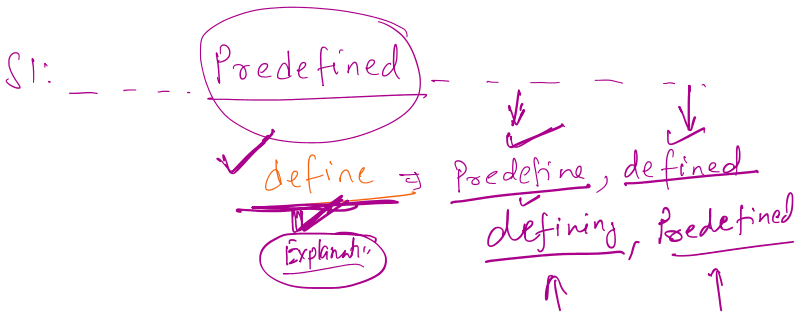


S2: Mohan is reading NLP Book



#tokens > K → Discards

< K → Pad it with Zero vectors



Feb 01, 2021

Raw Data (Text)

↓

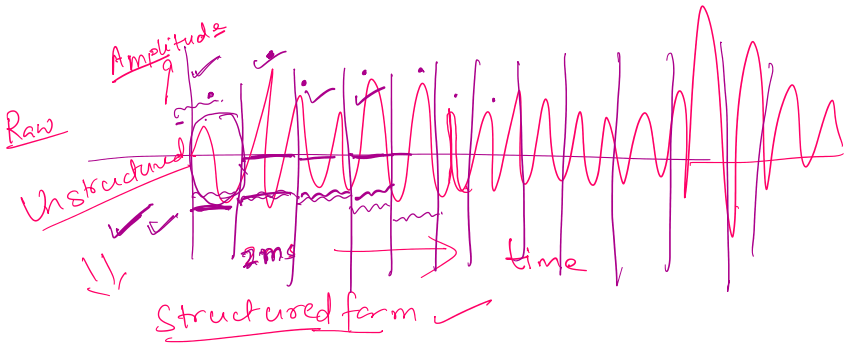
Term Frequency

	w_1	w_2	...	w_d	label
S1					+1
S2					+1
					-1
					-1

{Train — Pr / Valid}

Test

Audio/Speech/Signal



→ Sampling rate
→ Samples
basic ingredients in Signal

Speaker recognitions

	2ms f_1	2ms f_2	2ms . . .	2ms f_d	label (integer)
S1					$a_1 \leftarrow 1$
S2					$a_1 \leftarrow 1$
S3					$a_2 \leftarrow 2$
S4					a_2 2
S5					a_2 2
					✓
Sm					a_{1k} 1000

Cropping the Signal

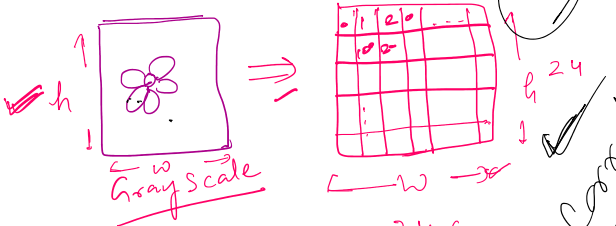
Stripping off the signal

Padding

Model(S_{test}) → 596

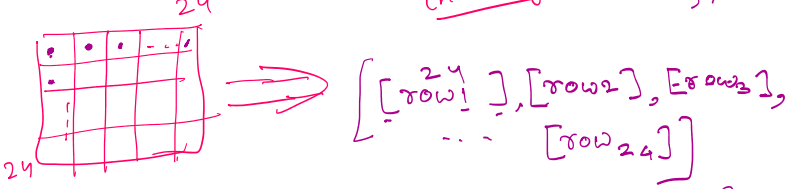
Image Data:

8-bit $2^8 - 1$ → 0 (Black)



0 - 255
intensity

576



[0.1 200. 243. . . . .]

[row 24]

: [0, 1, 200, 243, ...] 576

$I_{24\times24}$ → [ ] 576

$I^2_{24\times24}$ → [ ] 576

	P_1	P_2	. . .	P_{576}	label
I_1					
I_2					
.					
.					
I_m					

Crop it
Padding

	0	1	2 . . .	255	lab.
I_1					
I_2					
.					
.					
I_m					

ML model

Two Types → order of Pixel values does not matter location of object de
→ order matter (Sequence of pixel is important)

Language : $\Sigma = \{0, \ldots, 255\}$ W_1 W_{256} Dict size = 256

I_1 : [0, 1, 200, 143, ...] 576] 1000 Text

I_2 : [

.
.
.

T : [ ] 2000

one-hot-vector

$W_1 \rightarrow [1\ 0\ 0\ 0\ 0\ \ldots]$ 256

$W_2 \rightarrow [0\ 1\ 0\ 0\ \ldots]$ 256

2000

Assignment

Text Data → Image Data

Plot & View

Write some story about this

~~10×20~~

S1 — 200

S2 — 200

RGB → 8 + 8 + 8 = 24 bit

2^{24}

Pixel Value

10×20

10 mins:

Video:

$h \times w \times 3$

channel

frame rate 25 FPS

h, w, C channels

$h \times w \times c$

CF: $h \times w \times 3$

CF_1

100 frames

300 channels

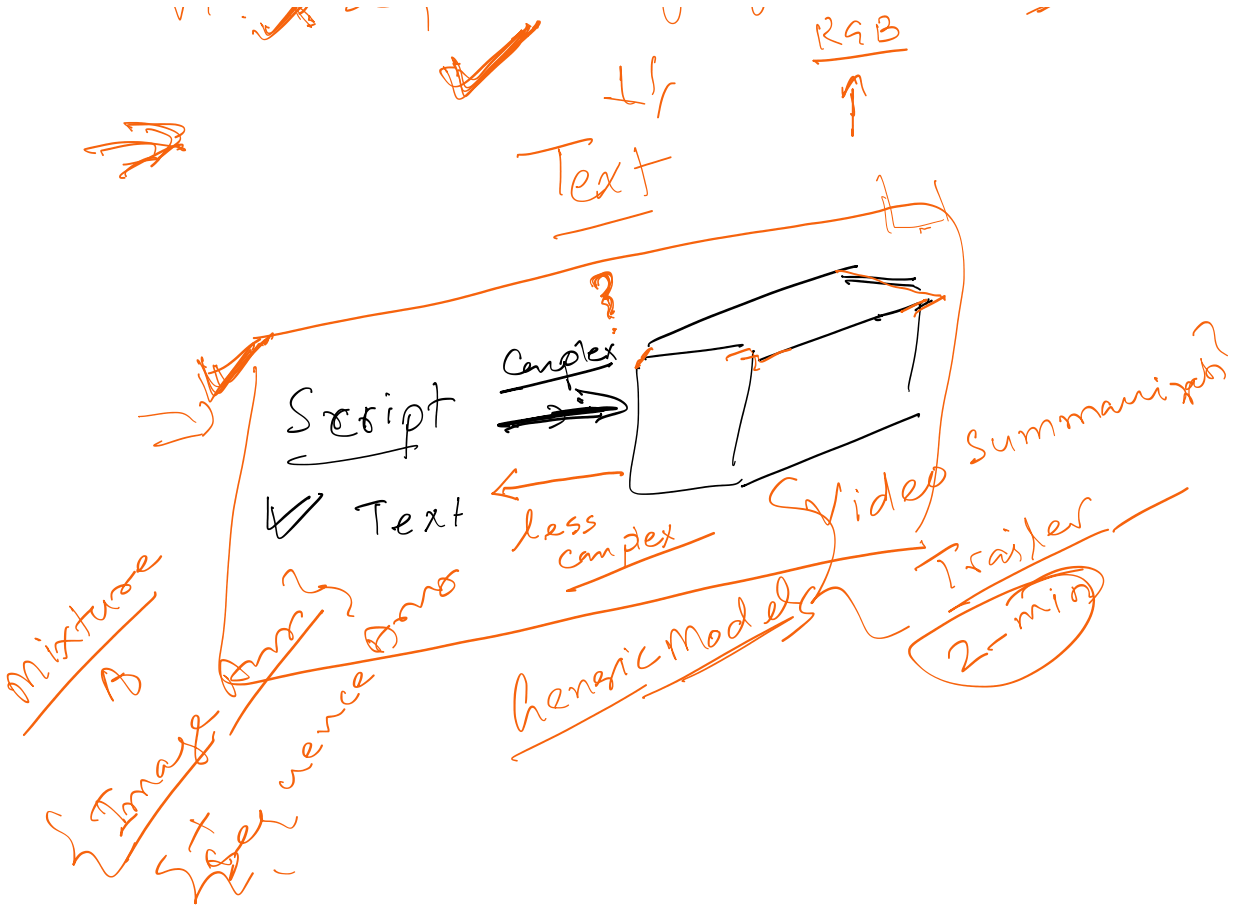
4 second

10 min Vide

(10×60×25) frames

V1: Sequence of frames

RGB



Text Data

Unstructured → TF matrix

Unigram	w_1	w_2	...	w_d	Lab. Senti.
S_1					+1
S_2					+1
					−1
					−1
S_{1m}					−1

Sentences / Documents

Information

$w_i \rightarrow S_{w_i}$ 80%

$w_j \rightarrow S_{w_j}$ 10%

IV - word

$K \Rightarrow$ unique

$Doc = -\sum p_i \log p_i$

$\sum_{i=1}^{K} p_i = 0$

$p_i \Rightarrow$ Unique i^{th} word

$\log$

$w_i \rightarrow S_{w_i}$

Document/Sentences : N

$w_i \rightarrow S_{w_i}$ $\quad \frac{S_{w_i}}{N}$

$$IDF = \log\left(\frac{N}{S_{w_i}}\right)$$

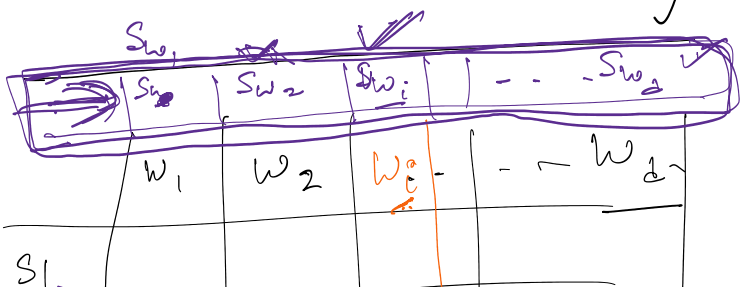
S_{w_i} : { How many documents/sent Contains the word w_i }

$S = \{S_1 \ldots S_N\}$

K^{th} sentence : $w_i \Rightarrow TF_{w_i} \times \log\left(\frac{N}{S_{w_i}}\right)$

Term Freq matrix

S_{w_i}



$N \times d$

N x d

TF-IDF

IDF

$T(k, i) \times \log\left(\frac{N}{s_{w_i}}\right)$

S1

S2

Sk

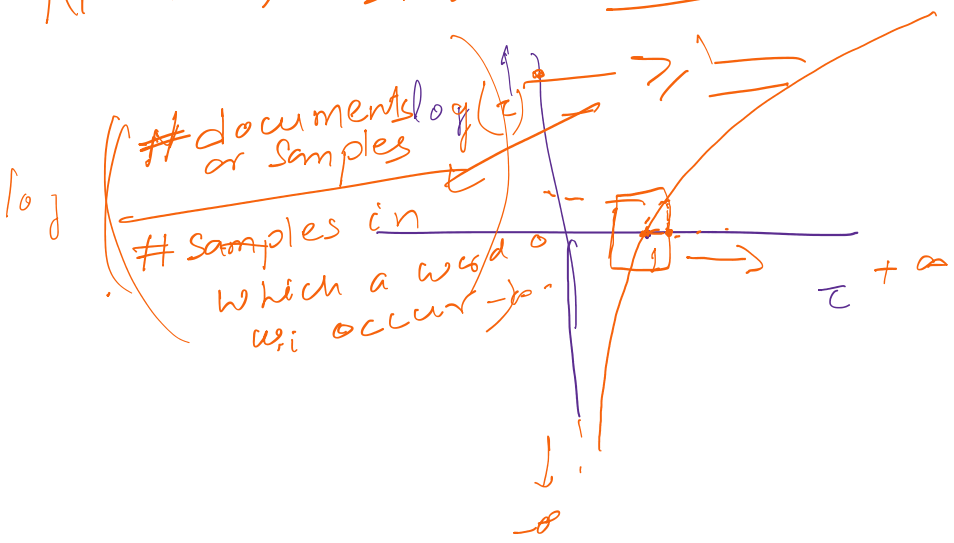
$TF_i \times \log\left(\frac{N}{\#\text{number of samples containing } i^{th} \text{ word}}\right)$

"The" 95%

$TF(\text{The}) = 2 \times \log\left(\frac{100}{95}\right) \to 1$ small (+ve)

N = 1M, 95% of 1M = 95x10^4

$\log\left(\frac{\#\text{documents or samples}}{\#\text{samples in which a word } w_i \text{ occur}}\right)$



K-gram model

K = 2 bi-gram model

S2: He is reading a book

S3: Ramayana is a good book.

bi Windows Size = 1

(This, is)
(is, NLP)
(is, This)

S2: He is reading a book.
S3: Ramayana is a good book.

(is, NLP) ✓
(is, This)
(NLP, class)
(NLP, is)
(class, NLP)

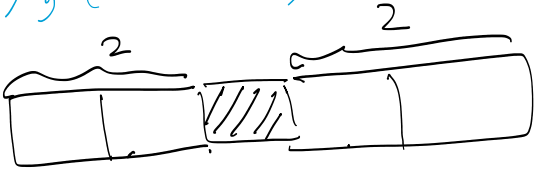
S1 {(This, is), (is, NLP), (NLP, class), (NLP, is), (class, NLP)}

1-skip (is, This), (NLP, is), (class, NLP)

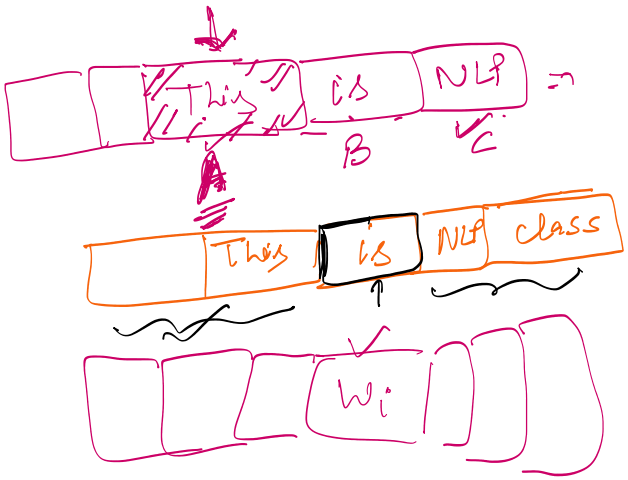
S2: {(He, is), (is, reading), (reading, a), (a, book)
(is, He), (reading, is), (a, reading), (book, a)}

S3: {(R, is), (is, a), (a, good), (good, book)
(is, R), (a, is), (good, a), (book, good)}

Window size = 2

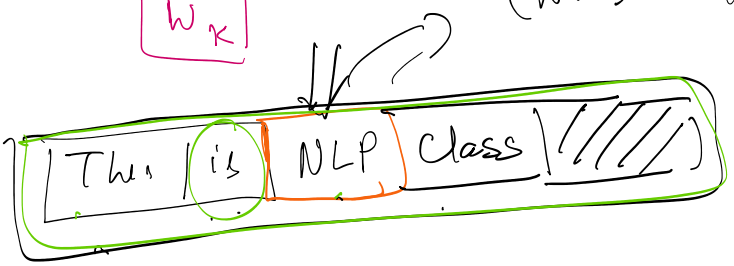


S1: This is NLP class. ✓



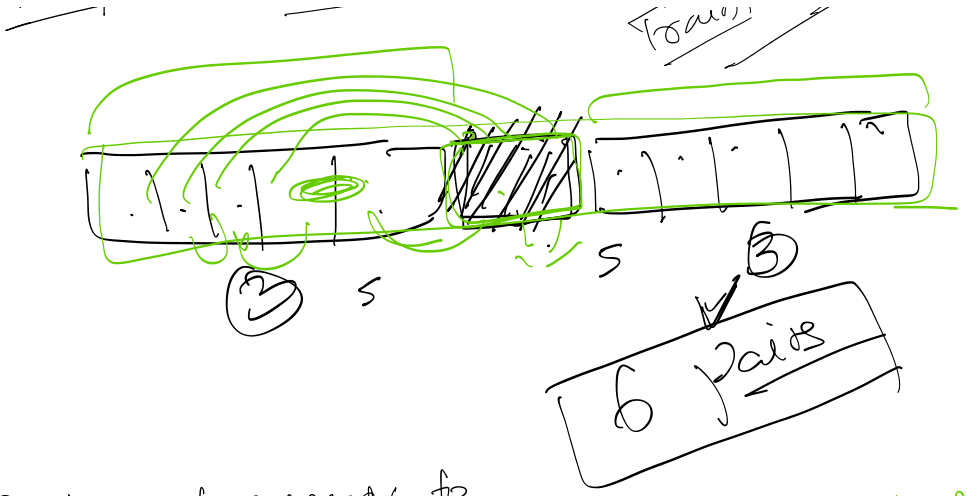
(This, is)
(This, NLP)

(is, "NLP")
(is, class)
(is, This)
(NLP, class)
(NLP, is)
(NLP, This)



Sample ~~bias~~ Random Sampling

Training complexity



→ Combine few words to a single word Token = each phrase must be a single token

↳ Phrases.

(phrase, ____)

SKIP-Gram.

Unigrams × 2. Window Size

Maximum possible Dictionary

one-hot-vector

"happy" → [ ] 10

"Laugh" → [ ] 10

Intelligent → [ ] 10

Smart

ant → [ ] 10 50, 100, 300

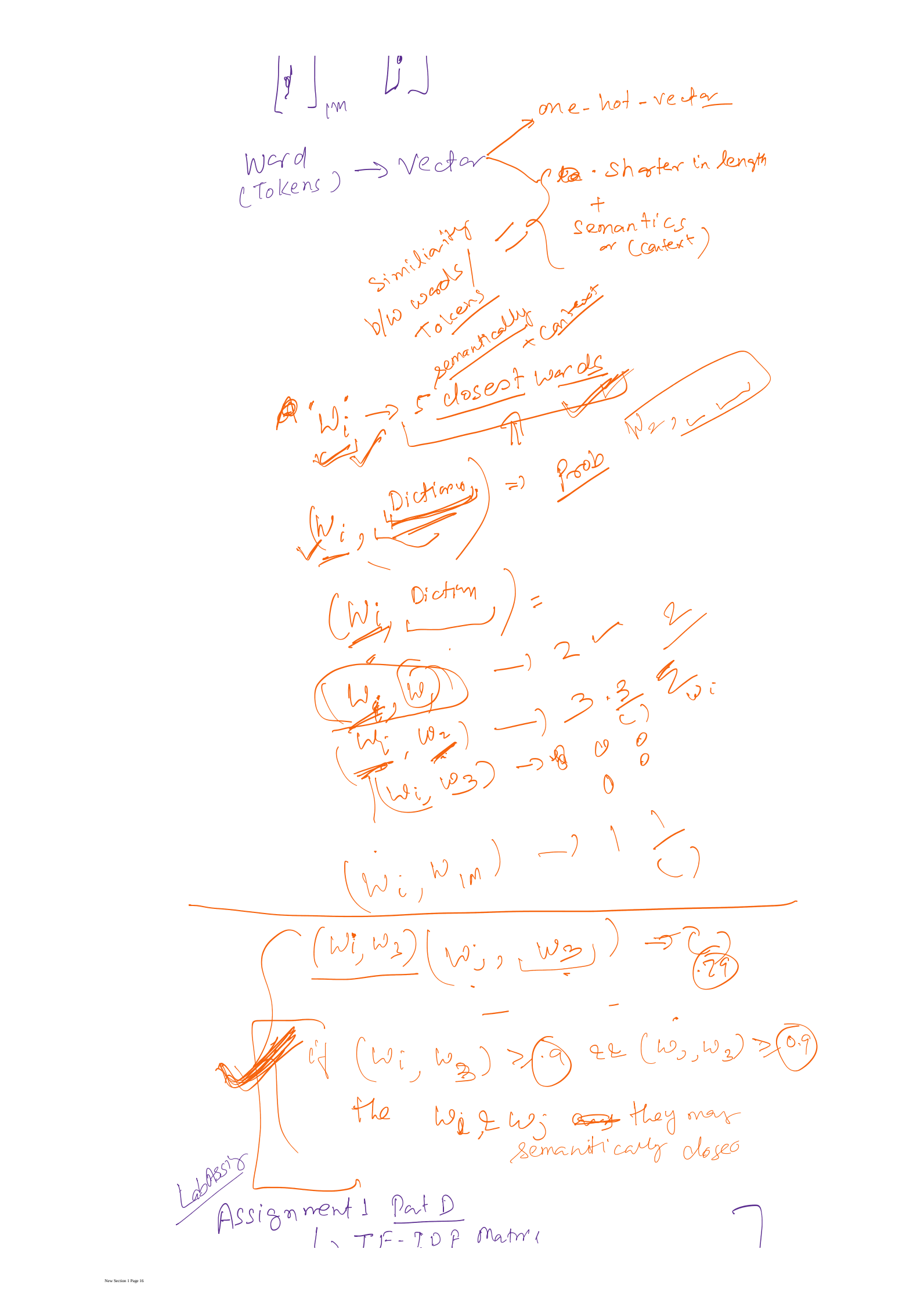
ants →

100 10,000 1M

happy heat

→

one-hot-vector


one-hot-vector
Word (Tokens) → Vector
shorter in length
+
semantics or (context)
similarity b/w words / Tokens
semantically + context
5 closest words
Dictionary
Prob
2
3
0
1
Assignment 1 Part D
1. TF-IDF Matrix

Lab

Assignment 1 Part D

- TF-IDF Matrix
- Generate Bigrams using skip-gram with window size 2-3.
- Generate & store in a file