10th April 2023

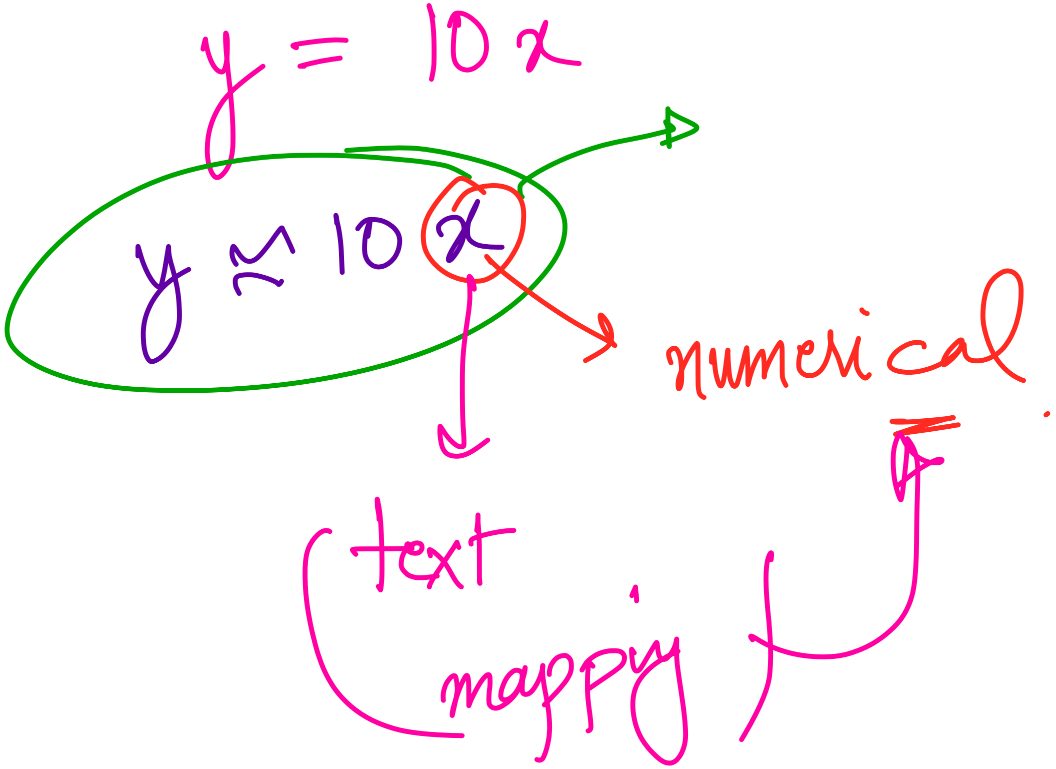
# Text Preprocessing Using NLTK

let's begin @ 9:05 pm

Images → num. → (0 − 255) RGB.

Text. → string

| Input $x$ | Output $y$ |
|---|---|
| 1 | 10 |
| 2 | 20 |
| 3 | 30 |
| 4 | 40 |
| 5 | 50 |
| . | . |
| . | . |
| . | . |



- Chat bots
- SA
- Speech to text
- Language models
- text translation
- text summarization

GPT-4

* High value in stocks creats high return .

high

* Cricket

'kohli'

'dhoni'

'catch'

'win'

'bowled'

Politics

'Modi'

'Putin'

'UK'

.

.

.

freq mapping

bank / bank

$2

riverside

mouse / mouse

computer

animal

bank: 2

* King Kohli is a good batsman. Kohli is great.

* Modi is PM of India.

Stopword ✓ ⟺ Keep Keywords *

↓

remove

* Grammerly* * → Stopwords may be imp.

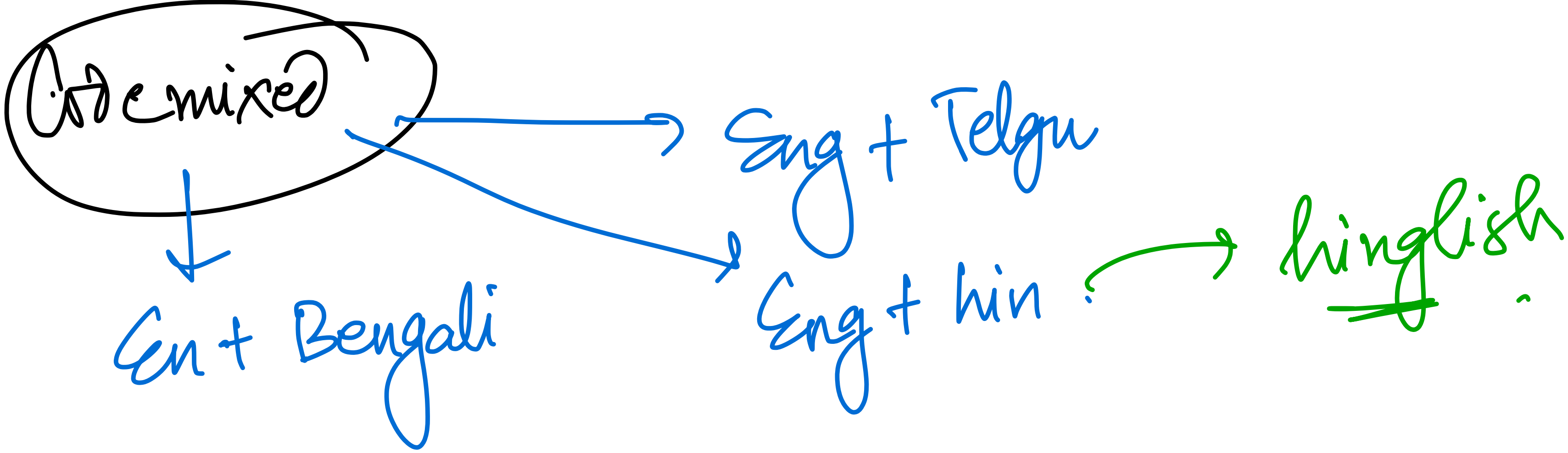
$\{w_1 : n_1,\ w_2 : n_2,\ \ldots\ \}$

Tokenization

strings → list of words.

↳ split on "spaces"

↳ tokenizer.



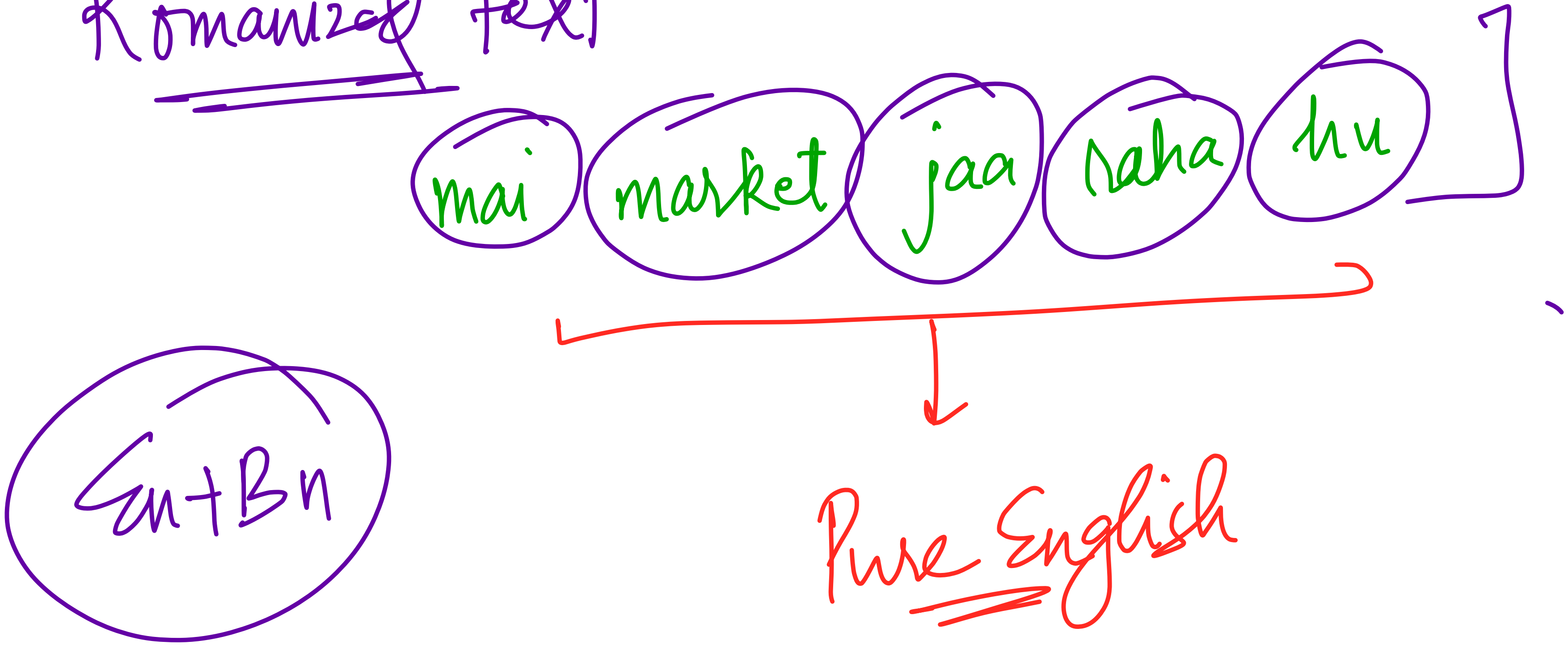
Codemixed

Eng + Telgu

Eng + hin → hinglish

En + Bengali

Romanized text



mai market jaa raha hu

Pure English

En+Bn

* king kohli good batsman. Kohli great.

* Modi pm india.

lower → r SW → Keywords → unique.

↓

Vocabulary $V$

Bag of Words *

| | india | pm | kohli | good | king | batsman | great | modi |
|---|---|---|---|---|---|---|---|---|
| $d_1$ | 0 | 0 | 2 | 1 | 1 | 1 | 1 | 0 |
| $d_2$ | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

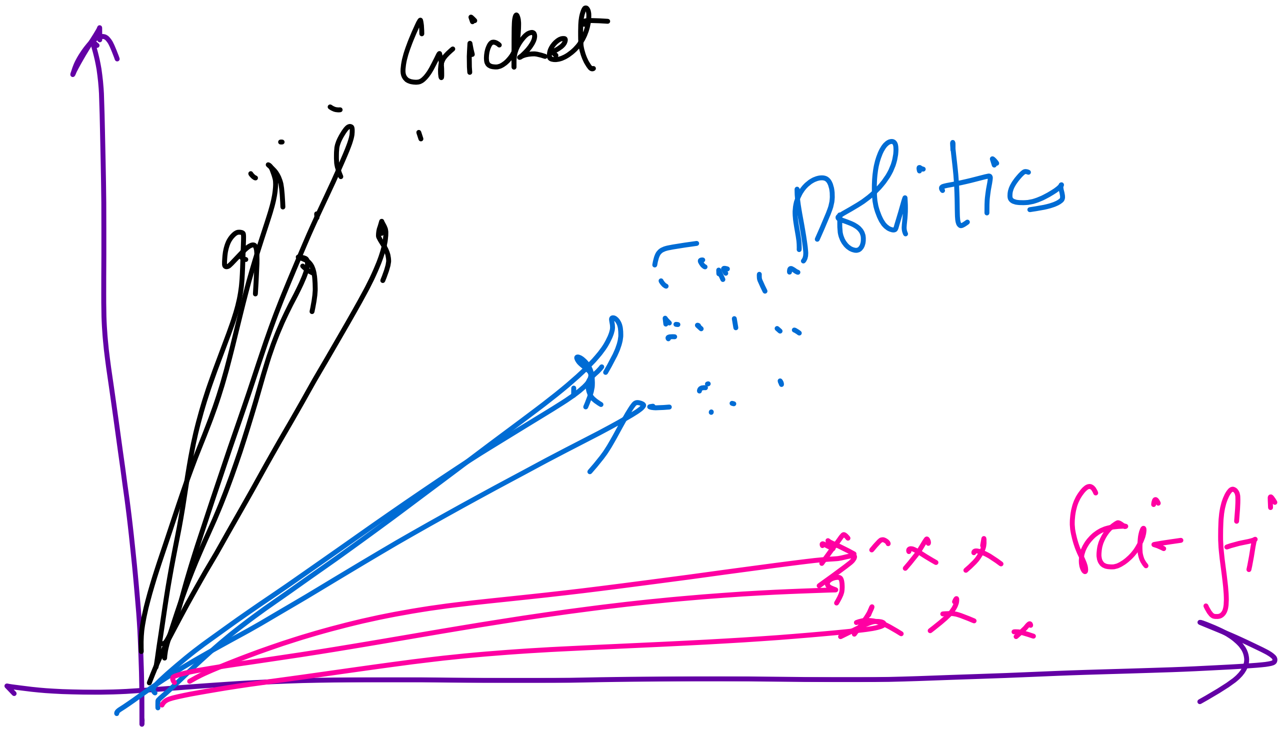
size = 10k × $V$

Only 3 type of articles.

Cricket

politics

Sci-fi



|  | 0th. | 1 | 2 |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | india | pm | kohli | good | king | batsman | great | modi |
| $d_0$ | 0 | 0 | 2 | 1 | 1 | 1 | 1 | 0 |
| $d_1$ | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

⋮

$d_n$

↓

1M

$V = 10,000$

very sparse data *

|  |  |  |
|---|---|---|
| (0,2) | → | 2 |
| (0,3) | → | 1 |
| (0,4) | → | 1 |
| ⋮ |  | ⋮ |

- Stemming & Lemmatization: → (Lemma).

warm / warmer / warmest → 1 col$^{m}$

play / playing / played → 1 col$^{n}$.

- Stemming → Cutting.

↓

Algo. → rules

- Poster → English. ✓

(diff. rules)

✓

- Snowball.

↓

more robust.

International

✓

* caresses → Caress *

rule. 'sses' → 'ss' → Caress

rule. 'ies' → 'i'

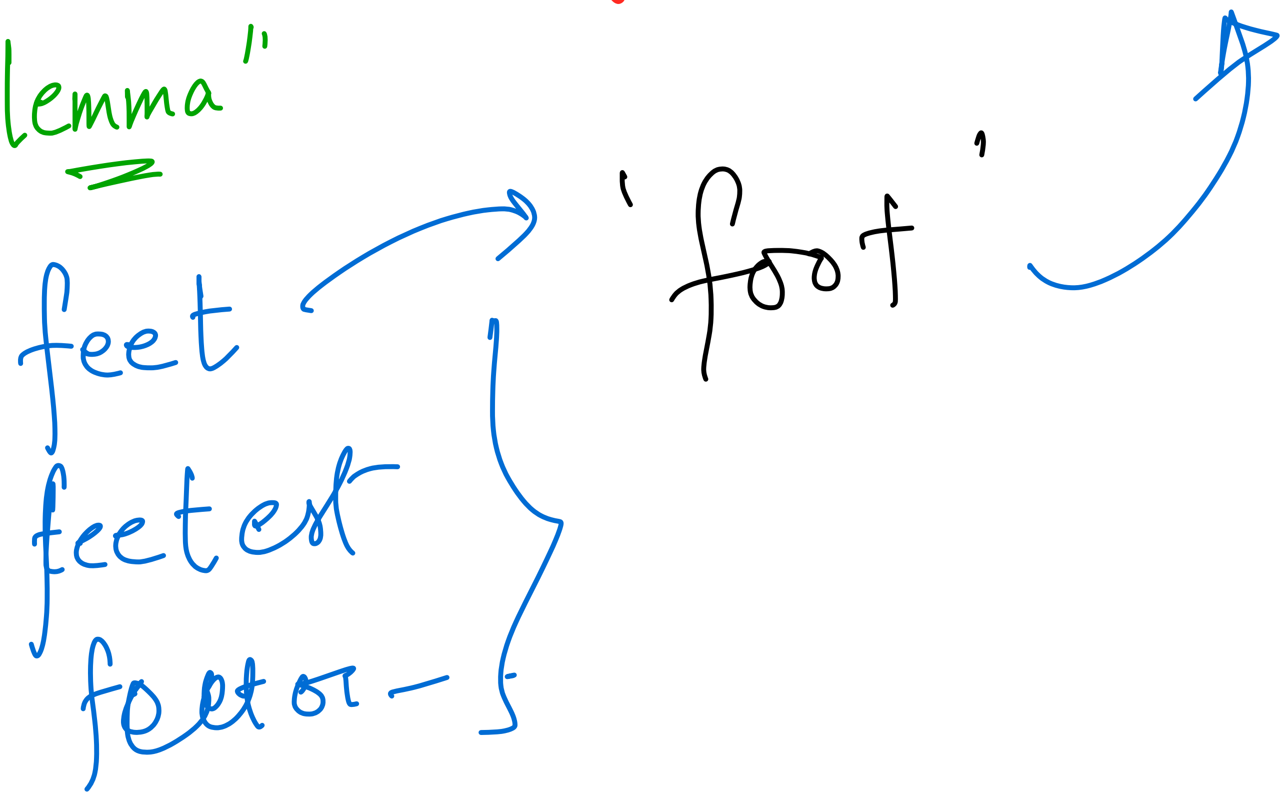
. ties → 'ti'

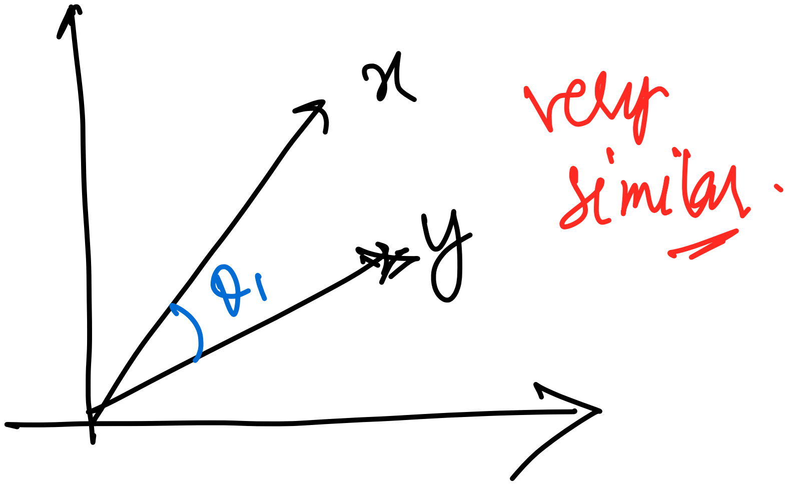
Caress ↕ 'ti' By stemming.

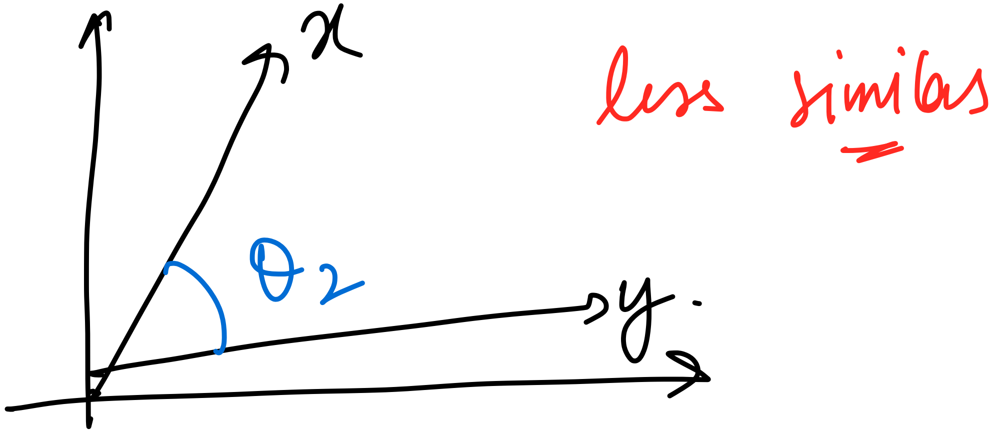
not taking grammer into account

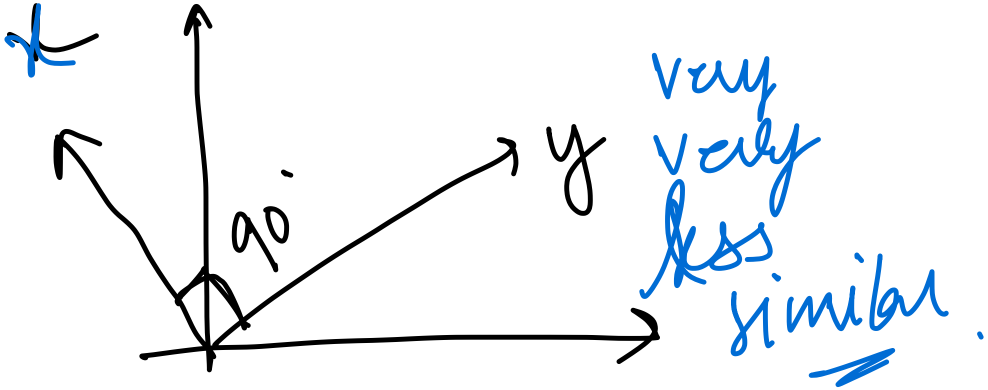
② Lemmatization : ( grammer into account )

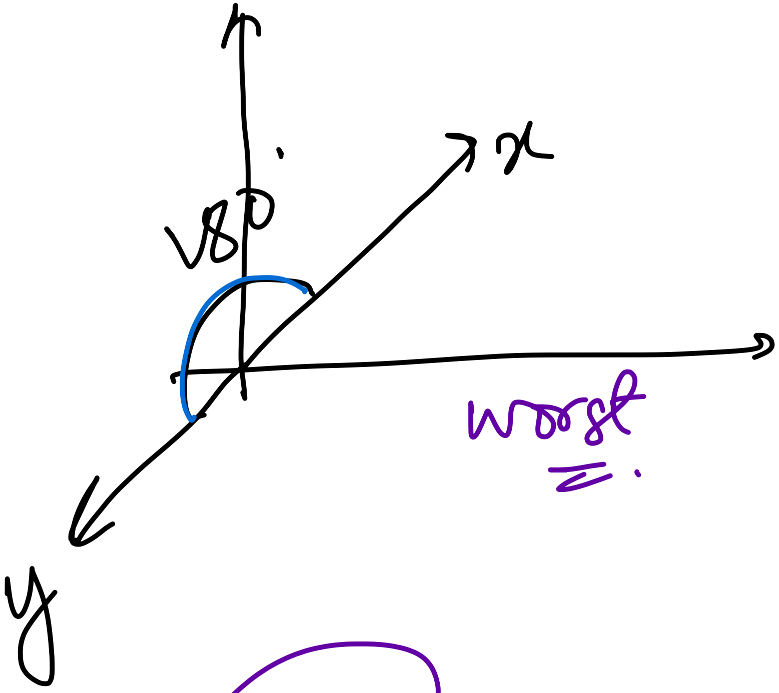
" lemma "

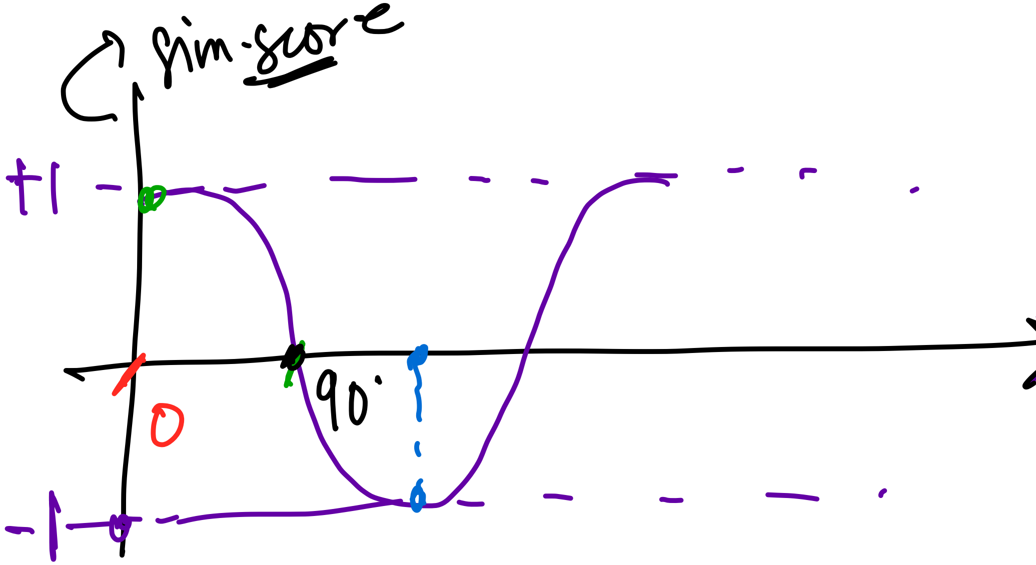






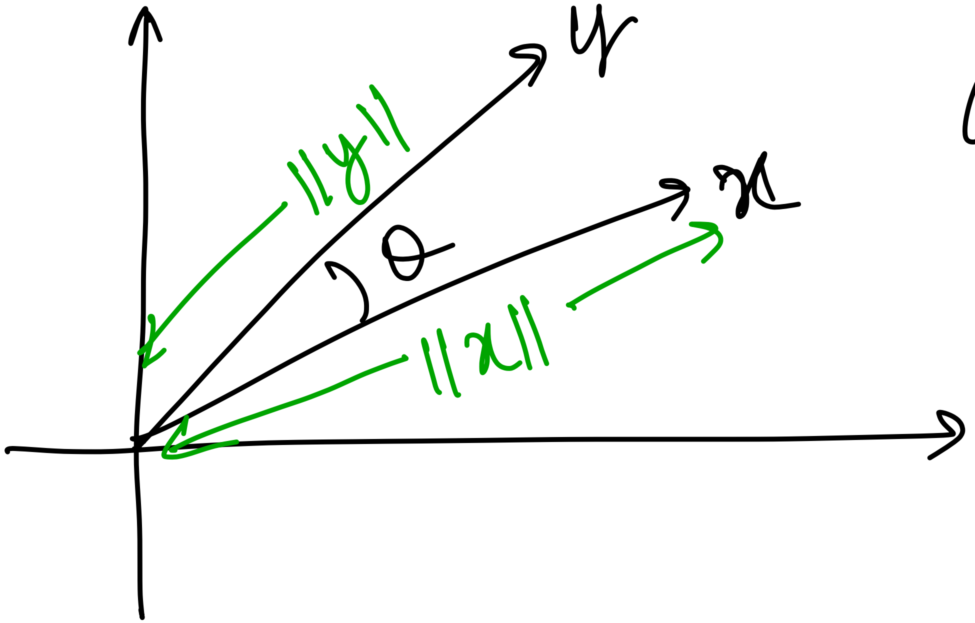






$\cos\theta$

[-1 to 1]



$$\cos\theta = \frac{x \cdot y}{\|x\| \, \|y\|}$$

dot product

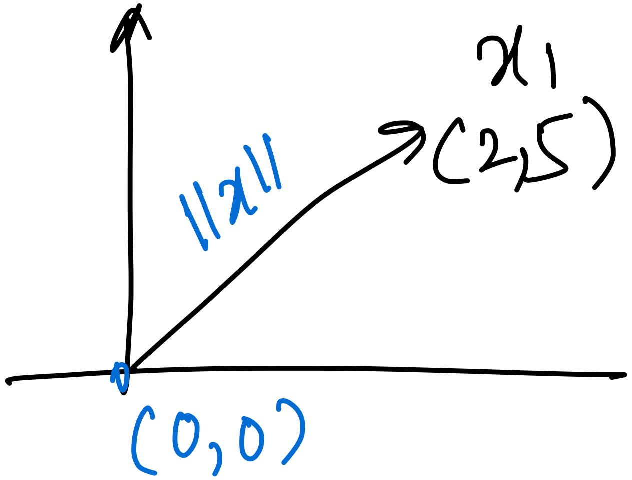
· elementwise multiplication & addition.

$x_1 \rightarrow (2, 5)$

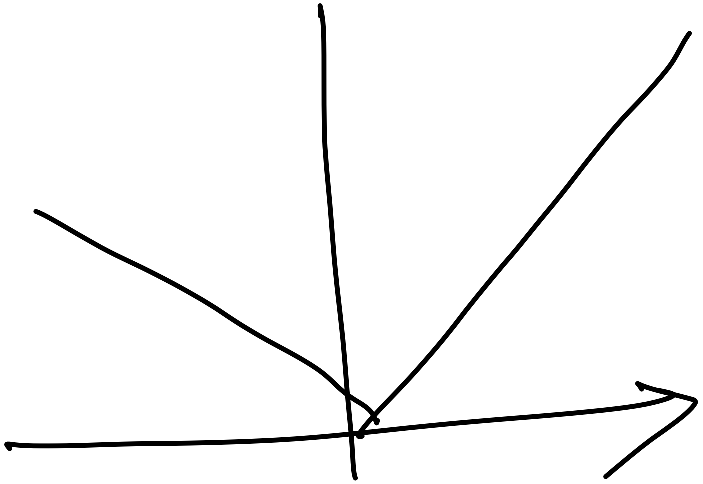
$x_2 \rightarrow (-1, 3)$

$$x_1 . x_2 = 2 \times (-1) + 5 \times (3) \checkmark$$

$$\|x\| = \sqrt{(2-0)^2 + (5-0)^2}$$



$|x| \longrightarrow$ mod $f^n$



$|-5| \longrightarrow +5$

$|5| \longrightarrow 5$

$\|x\| \longrightarrow$ vector length.

dist. formula

$\boxed{\text{norm}} \longrightarrow \boxed{L2}$ norm