11th April 2023

Text Preprocessing Using NLTK

let's begin @ 9:05 pm

Image: → numerical → (0 – 255) R·G·B

Text: string.

Input x	Output y.
1	10
2	20
3	30
⋮	⋮

$y = 10x$

$y \approx 9.98x$

num. text num.

map text → numerical.

- chatbots
- SA → 🙂
 → 😐
 → 🙁
- recommendations
- text to speech / vice versa.
- language models.
- text summarization.

Chat GPT-4

text → num

★ flow. text → numerical.

↓

raw — Lowercase the data.

- stopwords.
- tokenization
- Bag of words (BOW)
- Stemming / Lematization
- Case study.

↳ Similarity score.

eg.) High value in stocks create high return.

high

lowecasing

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

⇒ Modi is PM of India.

is/a/the/of → stopwords .. → remove X.

Keep keywords

Grammerly : → stopword may be dup.

d1 : King Kohli good batsman Kohli great.

:

d2 : Modi PM India.

:

dn :

lower → Remove SW → Keywords → unique → Vocabulary V

Bag of Words *

	King	kohli	good	batsman	great	modi	pm	india
d1	1	2	1	1	1	0	0	0
d2	0	0	0	0	0	1	1	1

:dn

8 # unique keywords : V

n :

(10000 × 8) → numerical

Cricket ✓

2 types of documents

Politics. ✓

	0	1	2	3				
	king	kohli	good	batsman	great	modi	pm	india
0: d1	1	2	1	1	1	0	0	0
1: d2	0	0	0	0	0	1	1	1

1000 x 500

90% are zero,

(0,0) →	1
(0,1) →	2
(0,2) →	1
(0,3) →	1
(0,4) →	1
(0,5)	
(0,6)	
(0,7)	

·only stores nonzero positions in matrix.

① Stemming : cut down.

warm / warmer / warmest

play / playig / played

→ only one word.

Root Word

Algorithm

Cuttig with some rules.

① PORTER

↓

(english) old.

② SNOW BALL

↓

(chinease, japanese, - - - - .)

* Caresses → Caress *

rule: 'sses' → 'ss'

rule: 'ies' → 'i' → ties → 'ti'

↓

making sense?

*

not taking GRAMMER into account

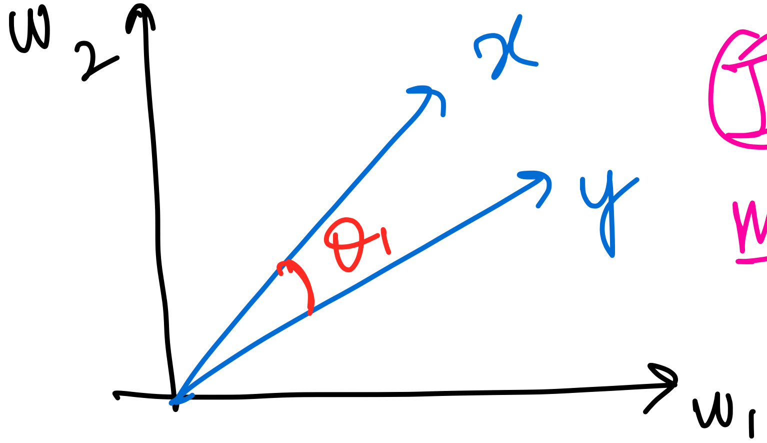
② Lemmatization ✓ → lemma Base word

↓

taking grammer into accout.

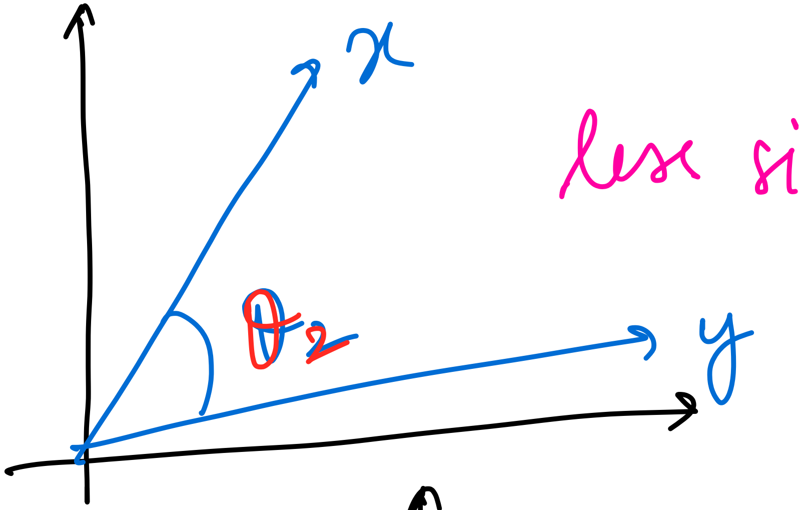
↓

follow rule of Grammer.

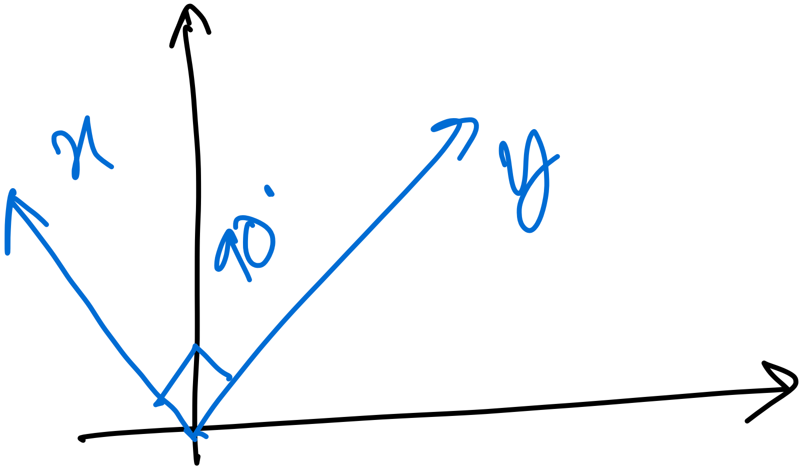


I

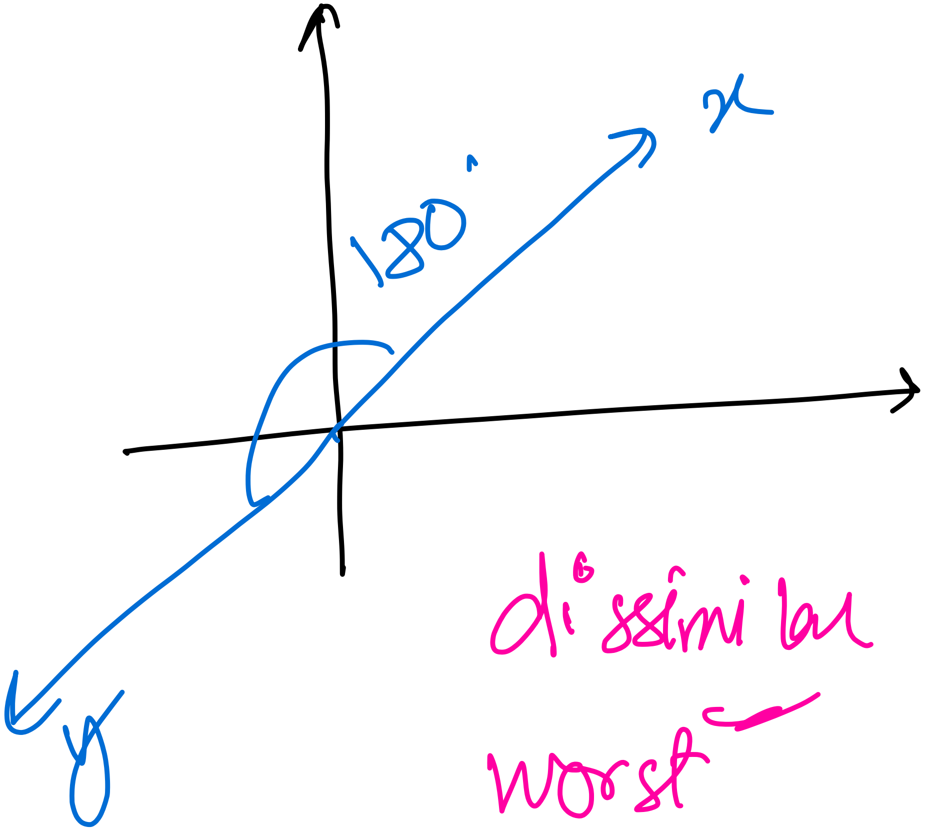
much similar



less similar



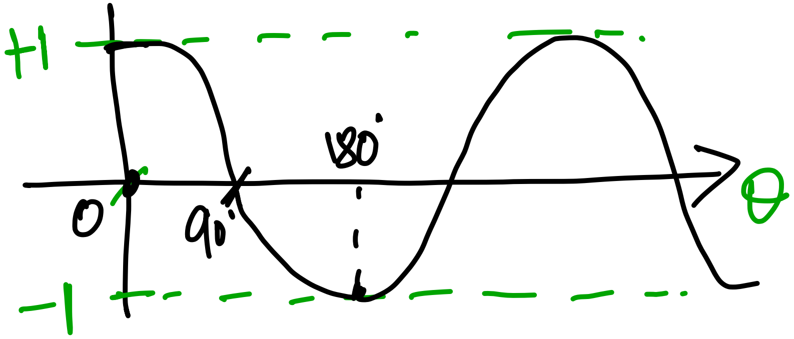
Very very very similar.



dissimilar

worst

(Cosθ)



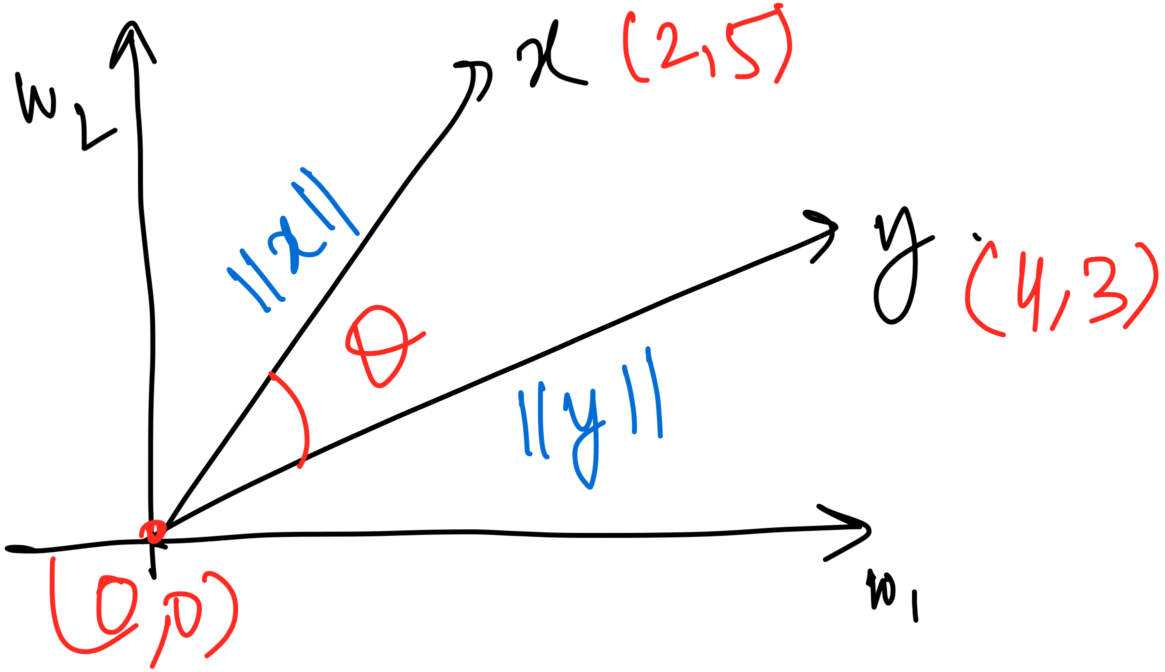
@ $\theta = 0 \Rightarrow +1$

@ $\theta = 90^\circ \Rightarrow 0$

@ $\theta = 180^\circ \Rightarrow -1$

Similarily score $= [-1 \text{ to } +1]$

$\cos\theta$



$x \to (2,5)$

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

norm/magnitude

$x \cdot y = 2 \times 4 + 5 \times 3$

$= 23$

$x \cdot y$ dot product

$\cos\theta \to$ Similarity score $\to$ (-1 to +1)

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