

March 15, 2023

DSML: DAV Fundamentals

Introduction to NLTK

"ChatGPT can't replace knowledge workers. It doesn't really understand what it's talking about and is not capable of generating new ideas or making hard decisions. It sounds coherent and vaguely insightful, but all it really does is try to sound smart by rephrasing the question its asked."

Days before OpenAI



Days after OpenAI



Knowledge workers:



5 12 86 3 2 2 } Example
031 10110 0xA22 } numbers

O TP: 7177 →

a. $\backslash d\{4\}$
b. $\backslash b\backslash d\{4\}\backslash b$

a. or b.?

Motivation :

Contexts: (a) Politics.
(b) Cricket.

Text →

Very difficult to
sense context

"He won." (a, b)

"He won the match." (b)

"He won a key state" (a)

"The crowd cheered
at his victory" (a, b)

"Sehwag earned the cheers
of the crowd that day." (b)

Images



(a)



(b)



(a)

Motivation : Computer Representation.

"pre-processing"

"He won a key state"

How to convert sentences/
text to a numerical form?

→ Vector?

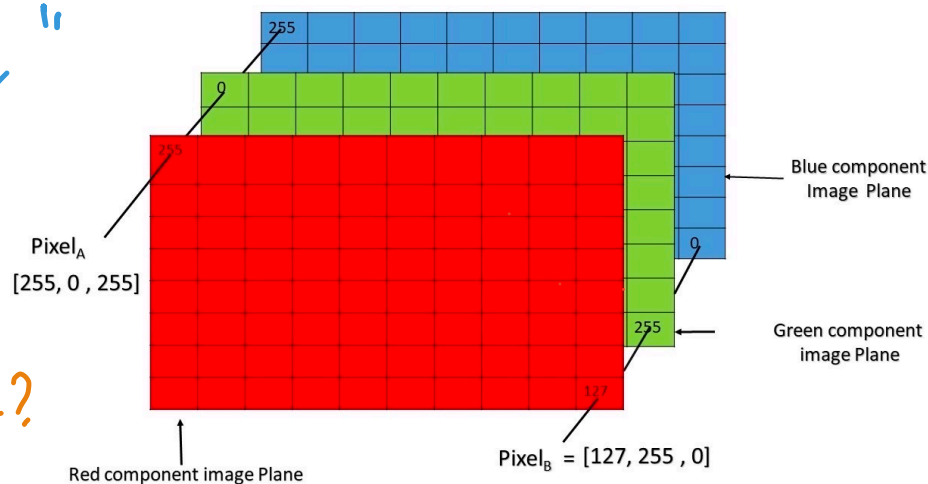
→ Matrix?

→ Graph?

→ Tensor?

Data Structure : Array.

Data types : Numerical.



Pixel of an RGB image are formed from the corresponding pixel of the three component images

"Bag-of-words"

→ This is important when we want to leverage existing optimization algorithms for training.

NLP :

① Google : Search Engine .

↳ Page-rank algorithm

↓

Search for text based on keywords.

granularity : web-page level .

one vector \longrightarrow 1 document / web-pages .

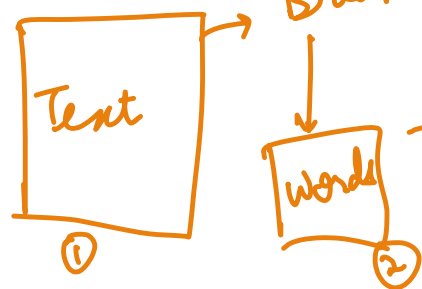
② Amazon : E-commerce website .

↳ Sentiment Analysis .

granularity : sentence level .

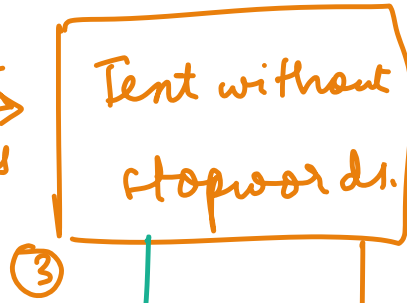
Big Picture :

Break into words - tokenize.



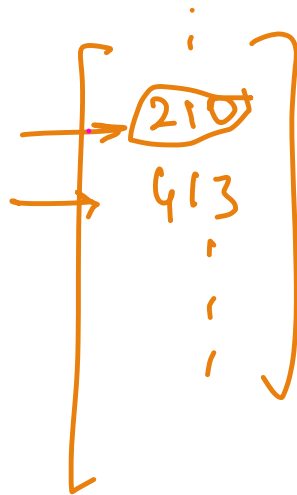
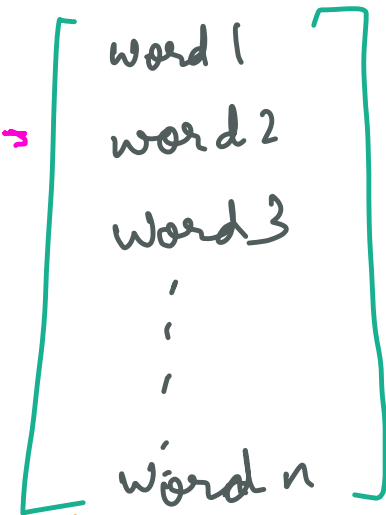
400 x 600

Remove stopwords



has 'n' unique words.

Create a vector as follows.



Get a vector having the frequency of the words at correct position.

→ call
→ called
→ calling → same
→ calls or different?
→ caller



Two ways to do this:

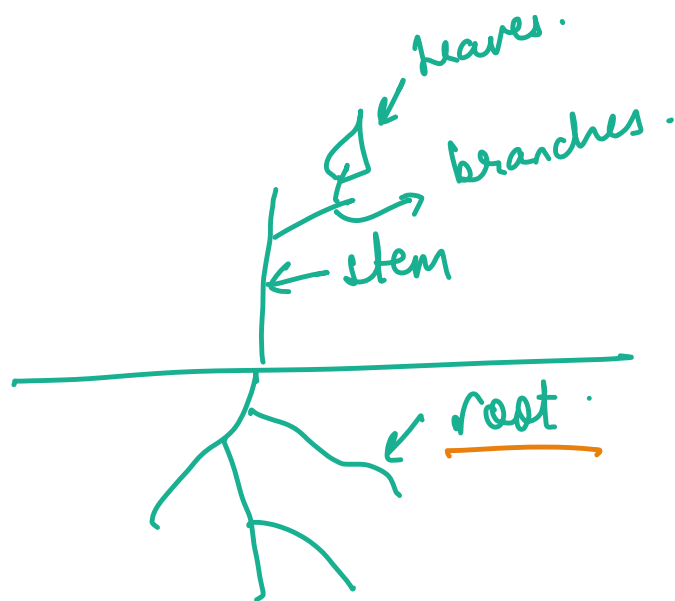
(a) Stemming.

(b) Lemmatization.

I like "animals" because they are good.
↑
more importance.

1] "animals" and animals should be treated differently at the word tokenize stage.

2] Amazon: Instead, we could treat them the same at the word tokenize stage, and later, increase frequency.



"Stemming"

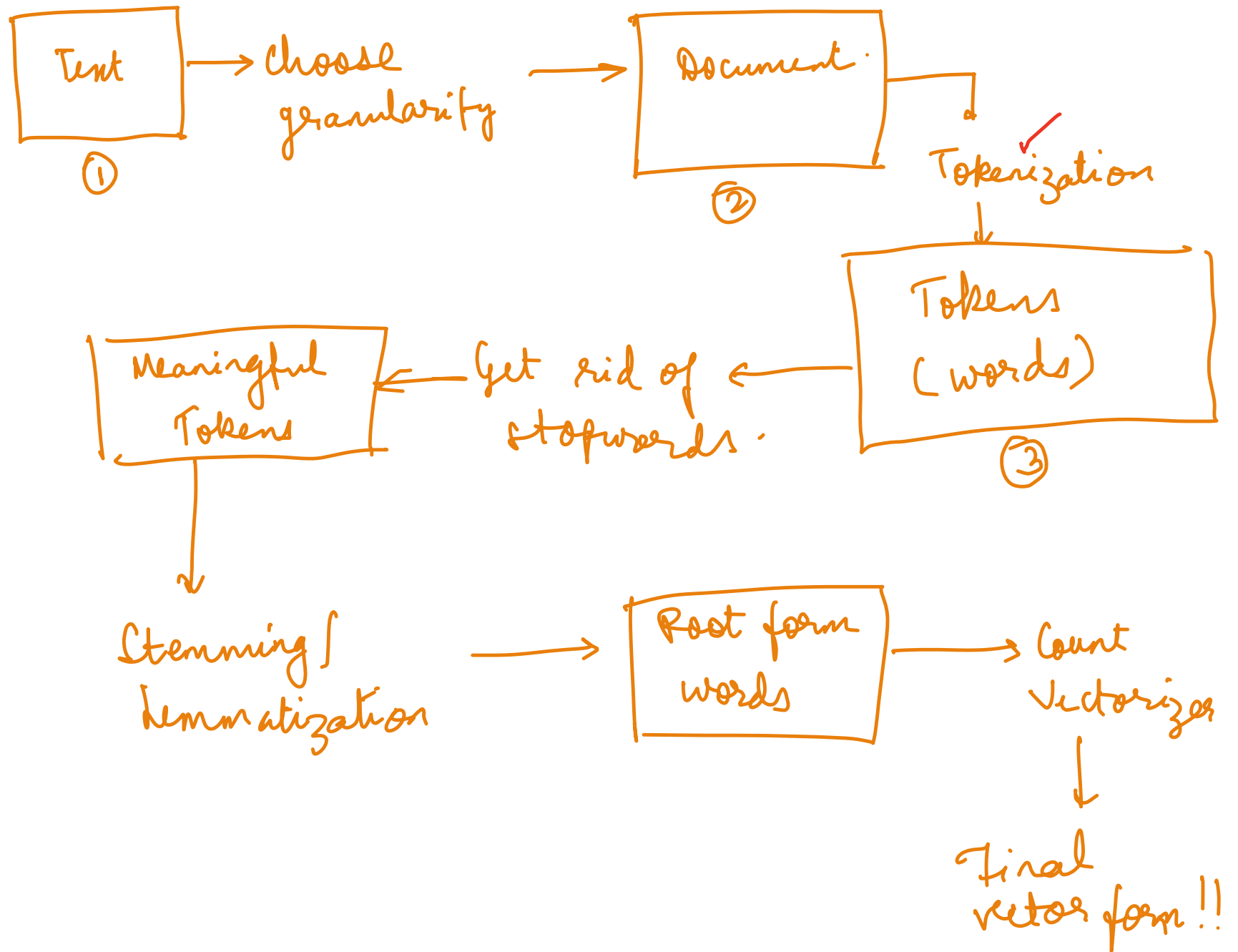
← { eat
eating.

root word - The origin of the word.

stem → Base word.

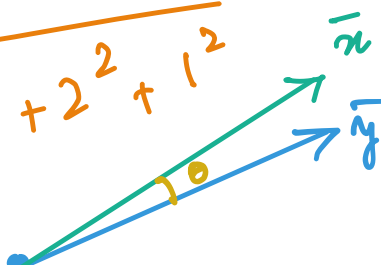
branches → suffixes or prefixes.

lemma — a short rule to be used further.



$$\|\vec{x}\| = \sqrt{\vec{x}^T \vec{x}}$$

①, ②, ③

$$\|\downarrow\| = \sqrt{1^2 + 2^2 + 1^2}$$


Case 1

Both vectors
point in
a "similar
direction"

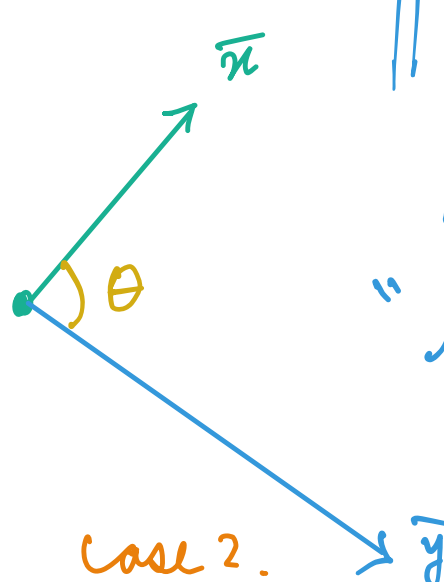
Inner product, dot product

$$\cos(\theta) = \frac{(\vec{x}^T \vec{y})}{\|\vec{x}\| \|\vec{y}\|}$$

$$[1, 2, 1]$$

$$[2, 2, 0]$$

$$2 + 4 + 2 = 8$$



Case 2

they are pointing
in different
direction.

"norm"
length of
the vector.

Cosine similarity : Checks how similar two vectors are based on the angle between them.

word 2 vec \rightarrow NN algo for obtaining
meaningful vector forms.

\rightarrow uses word associations
(Eg: how many times
has "tin" appeared with "man")

Count Vectorizes \rightarrow Very basic way of
doing the same.