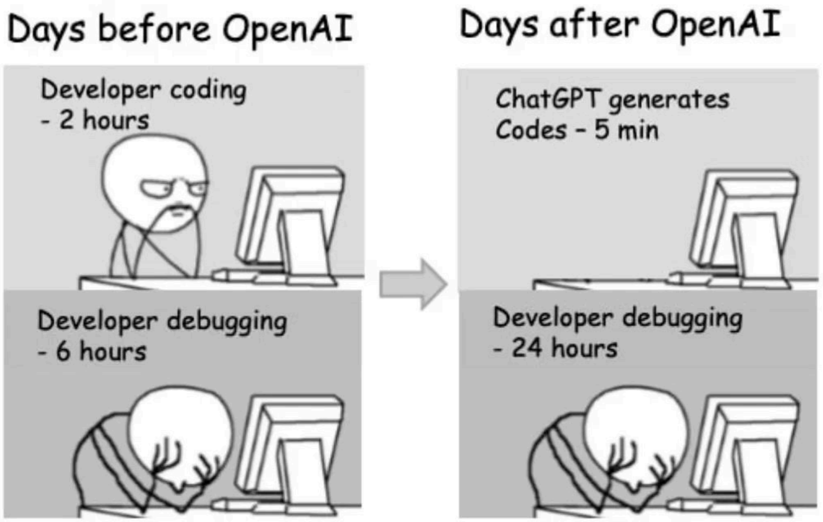
February 4, 2023

DSML: CC Fundamentals

# Introduction to NLTK


Days before OpenAI
Developer coding - 2 hours
Developer debugging - 6 hours
Days after OpenAI
ChatGPT generates Codes - 5 min
Developer debugging - 24 hours

"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."

Knowledge workers:

51286     322

031     275

10110     0xA22     OTP: 7177 →

a. \d{4}

b. \b\d{4}\b

a. or b.?

# Motivation :

## Text

## Images

" He won by scoring a century "

" He won the election by securing a 51% vote share "

" Sehwag managed to secure India's victory by scoring the most runs"

# Motivation : Computer Representation.

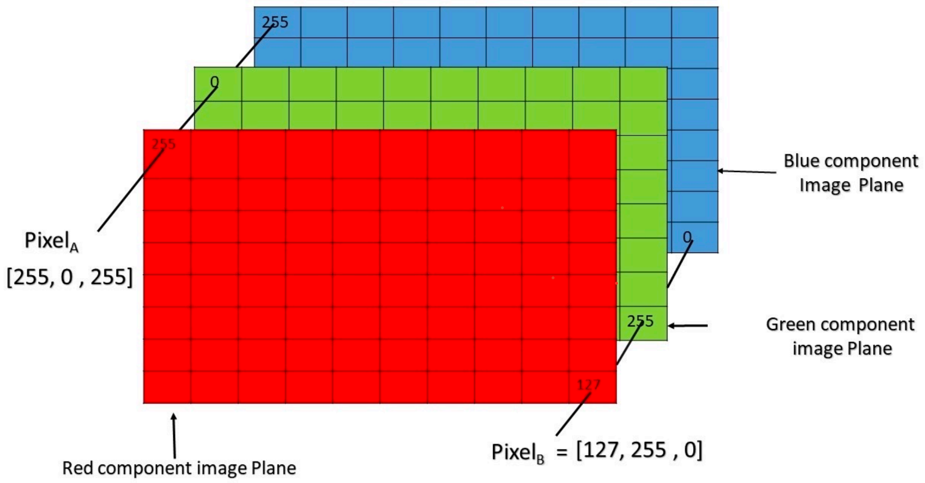
str ?

int ?

float ?

bool ?

numpy array ?


255
0
255
0
255
127
Pixel$_A$
[255, 0 , 255]
Blue component Image Plane
Green component image Plane
Red component image Plane
Pixel$_B$ = [127, 255 , 0]

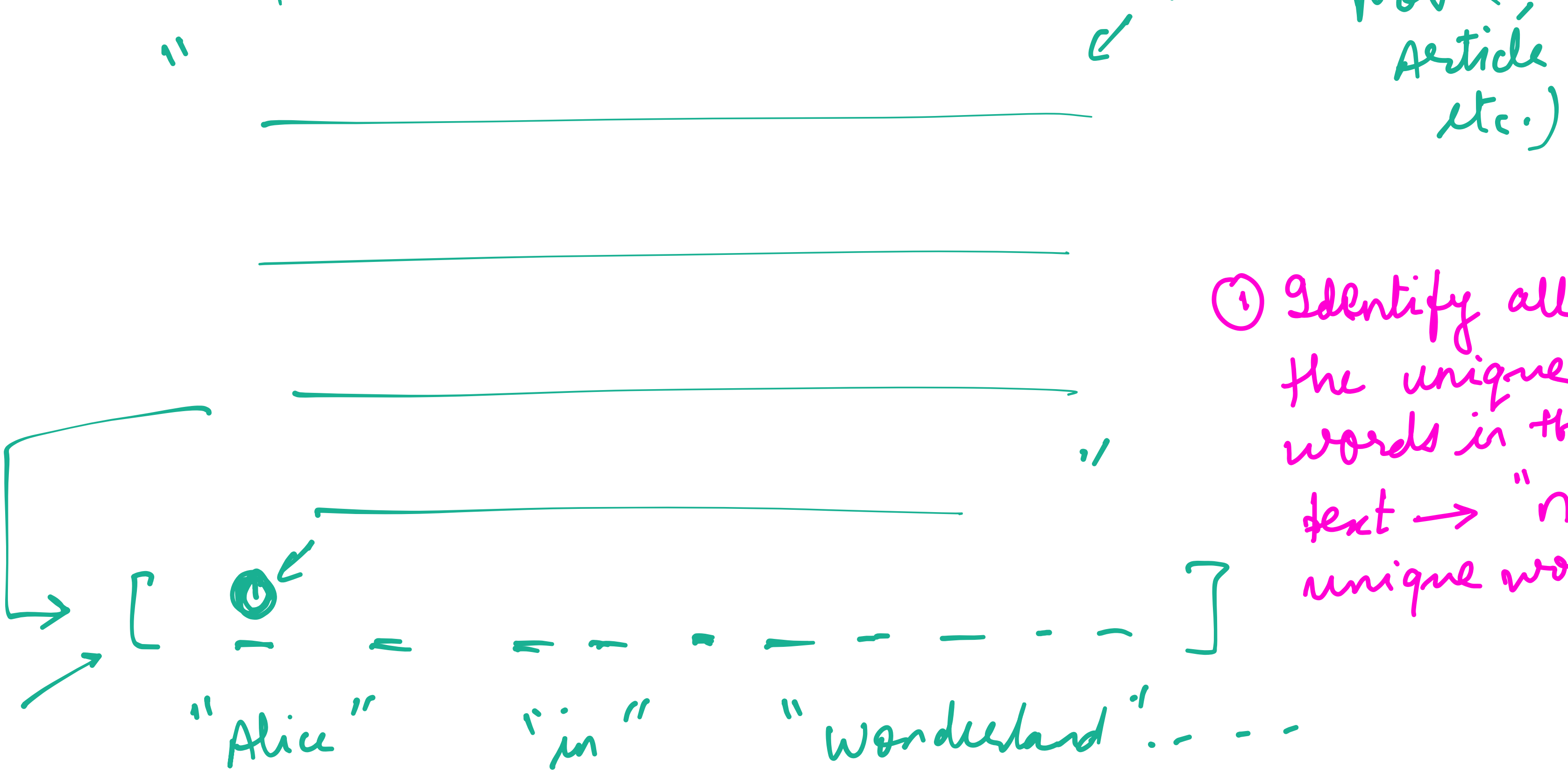
Pixel of an RGB image are formed from the corresponding pixel of the three component images

Goal: Machine Learning, Optimization.

Fundamental challenge : Make matrices or vectors out of text!!

**Tokenization:** Convert a sentence into component **words**.

"Bag of words"



① Identify all the unique words in the text → "n" unique words.

[[3, 3, 1],<br>
[1, 2, 1],<br>
[3, 3, 9]]

```
['gangs wasseypur great movie .', 'success movie depends performance acto
rs .', 'new movies releasing week .']
```

↓

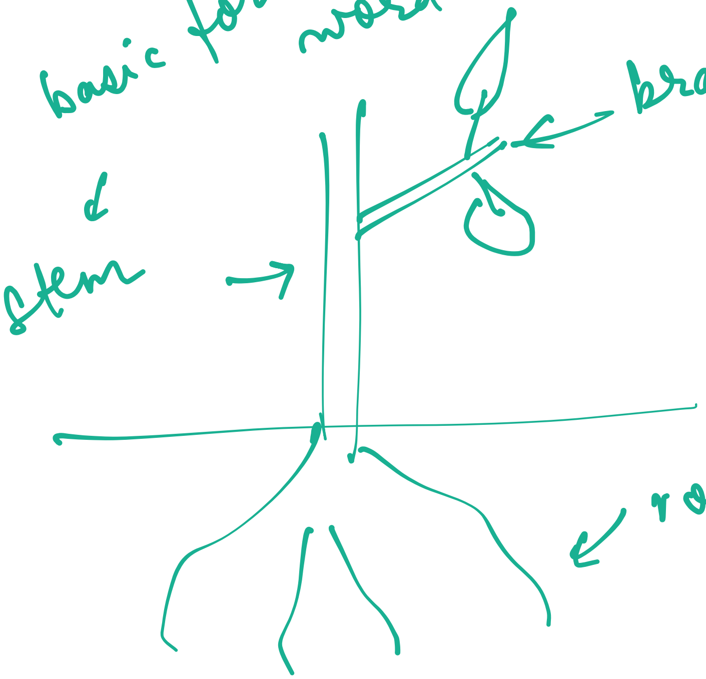
" " , " " → 12 unique word.

[

basic form of the word.

↓

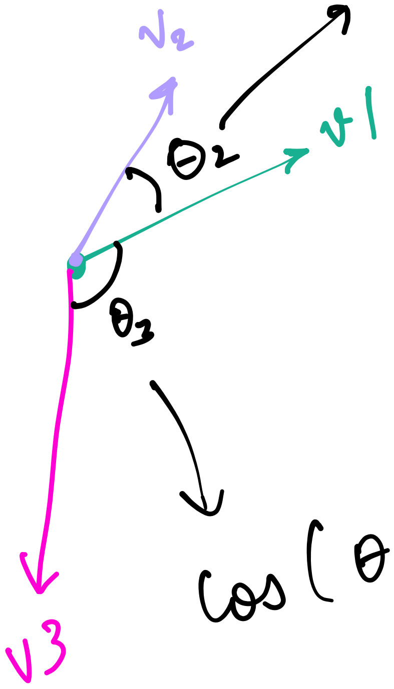
stem →



← branches — tense, plural form etc.

↙ root → where the word comes from

# "cosine similarity"

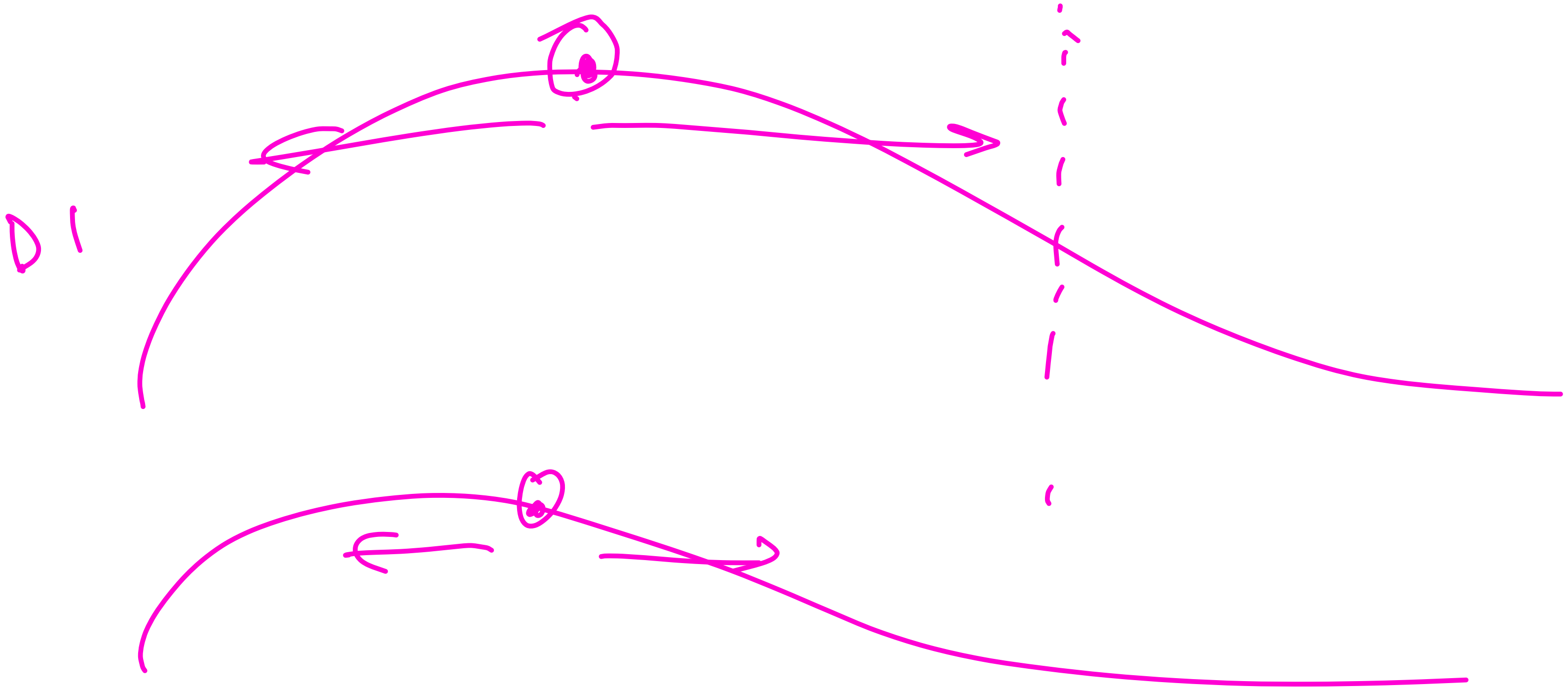


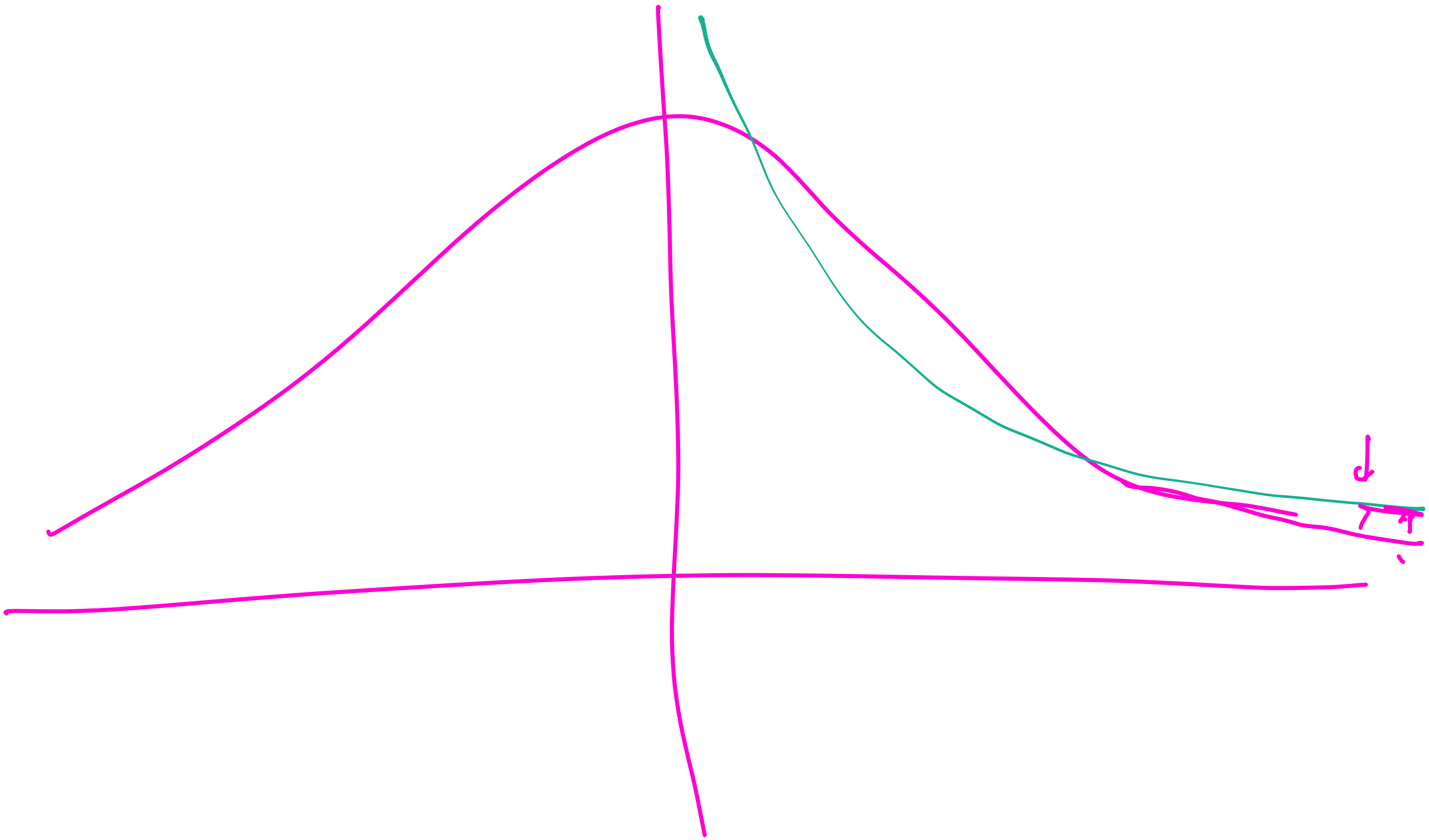
$\cos(\theta_2)$ ↓ cosine similarity between v1 & v2.

$\cos(\theta_3)$ → cosine similarity between v1 & v3

$$\cos(\theta) = \frac{\vec{v}_1 \cdot \vec{v}_2}{\|\vec{v}_1\| \cdot \|\vec{v}_2\|}$$ ✓

[ 2 , 3 , 1 , 5 , 4 ] →

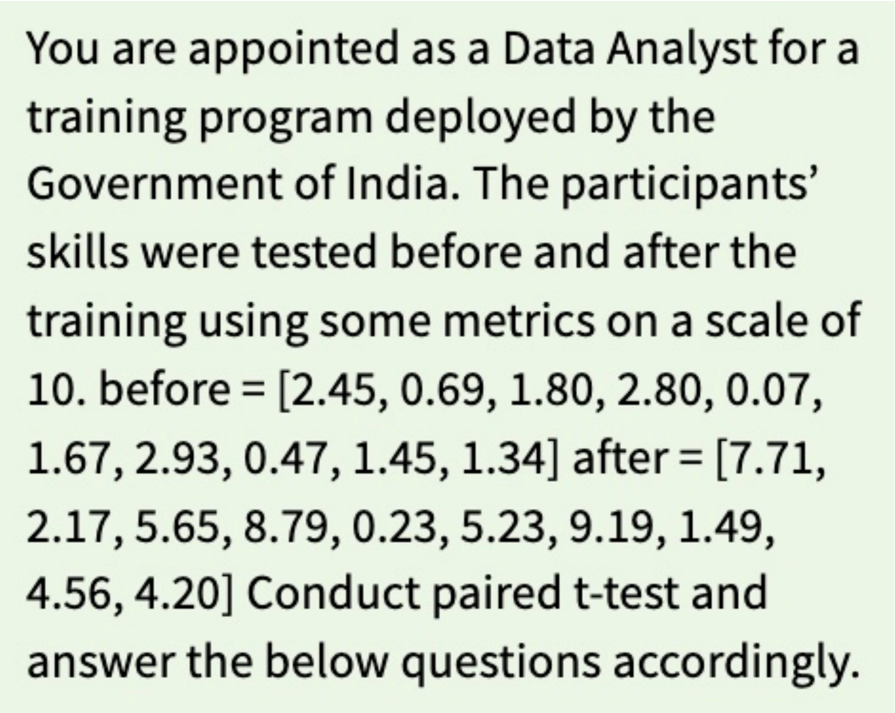




↗ Parametric test → Assumes we know the
parameters of the "population"
data.
distribution .

↗ Non - parametric test → opposite.

$p =$



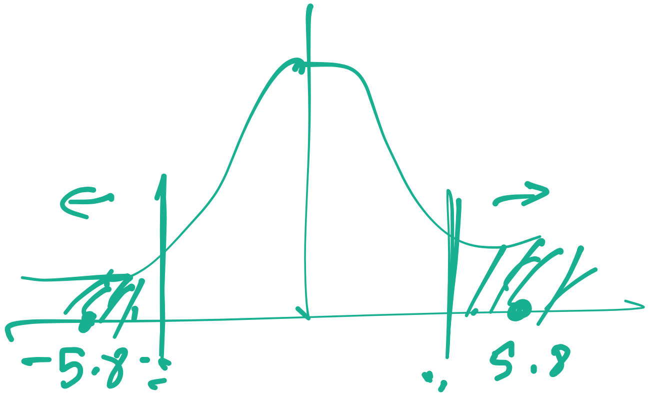
You are appointed as a Data Analyst for a training program deployed by the Government of India. The participants' skills were tested before and after the training using some metrics on a scale of 10. before = [2.45, 0.69, 1.80, 2.80, 0.07, 1.67, 2.93, 0.47, 1.45, 1.34] after = [7.71, 2.17, 5.65, 8.79, 0.23, 5.23, 9.19, 1.49, 4.56, 4.20] Conduct paired t-test and answer the below questions accordingly.

-ve ← before, after.

+ve ← after, before

paired → 2tailed ✓

paired → left/right.



* Use same dataset to explain different H.T. types. (maybe loan.csv)