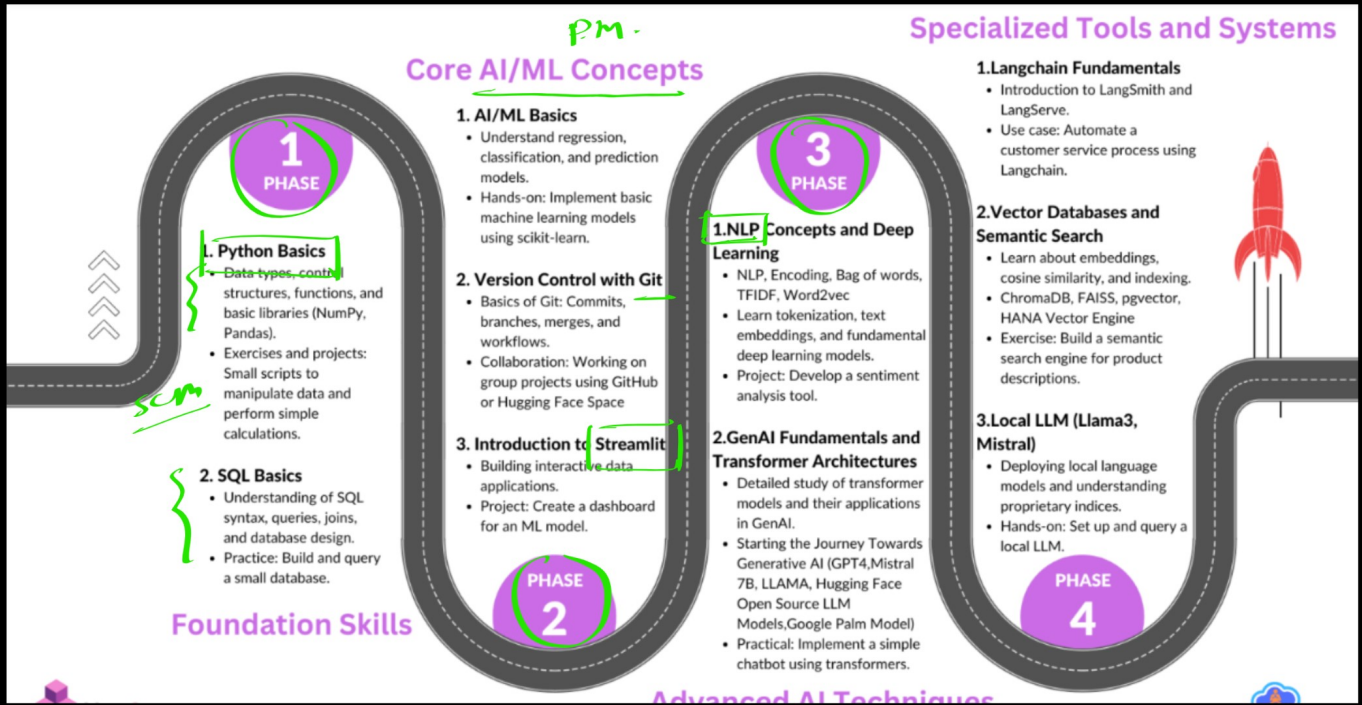


Day 04 - Gen AI Business & SAP.



NLP

Lecture

6 Tokenization ✓

6 Lemmatization ✓

6 Vectorization ✓ → TFIDF

6 Similarity Search → Cosine Similarity

End to End App.

→ FAQ / KBA Bot (chat).

→ SAP BTP

SAP
BTP

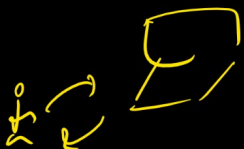
How to create sales order in SAP

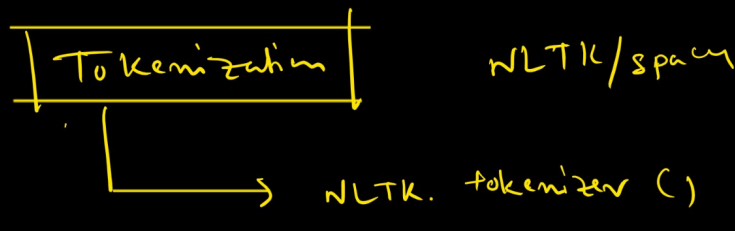
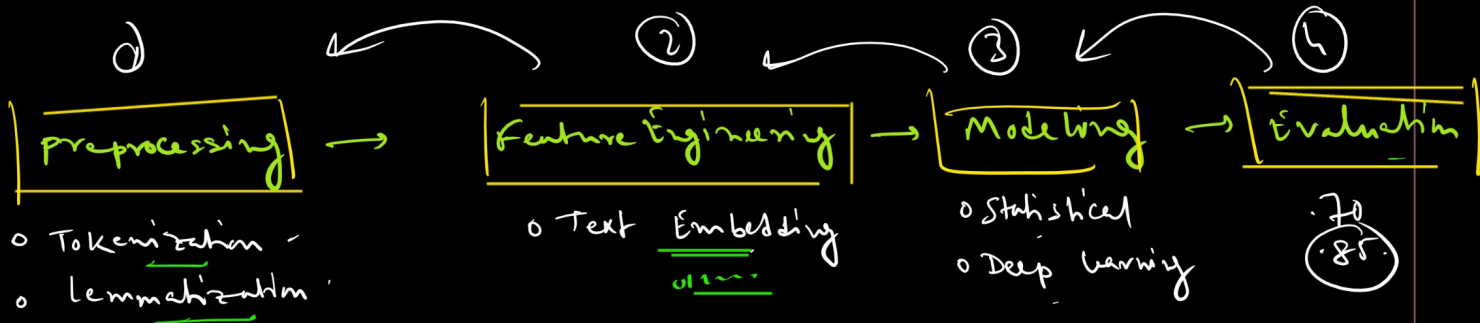
How can I generate a customer Invoice in SAP

what is the process to place a purchase order.

chatbot.

01 11 2023.





He in a good boy

x x

stopwords

NLTK.corpus.stopwords()

✓ Lemmatization

Identity base works.

Is too much eating is harmful?

ate → eat

linguistic knowledge.

nltk . stem . wordnet lemmatization()

Stemming ✗

talking — talk
walking — walk
adjustable — adjust. } Rule based.

Vectorization

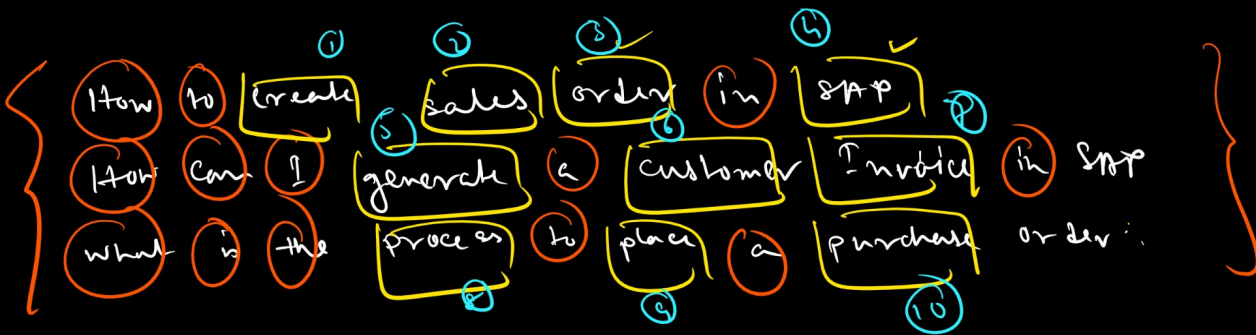
TF-IDF

o Bag of words

o word 2 Vec

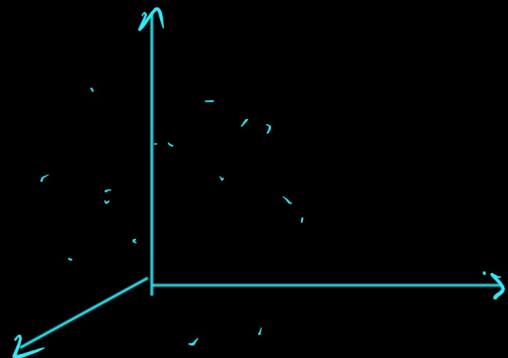
o Embedding models.

Techniques to do Vectorization.



10 unique words

Dataframe



S1 Boy is good. \longrightarrow Boy good
 S2 girl is good. \longrightarrow girl good
 S3 Boy and girl are good. \longrightarrow boy girl good.

boy girl good

$$TF = \frac{\text{No. of Representation of words in sentence}}{\text{No. of words in sentence.}}$$

$$IDF = \log_e \left(\frac{\text{No. of sentence}}{\text{No. of sentence containing the word}} \right)$$

	TF		IDF
	S1	S2	S3
good	1/2	1/2	1/3
boy	1/2	0	1/3
girl	0	1/2	1/3

$\log_e \left(\frac{3}{3} \right) = 0$
 $\log_e \left(\frac{3}{2} \right)$
 $\log_e \left(\frac{3}{2} \right)$

Final TF-IDF

	good	boy	girl
S1	0	$1/2 \log (3/2)$	0
S2	0	0	$1/2 \log (3/2)$
S3	0	$1/3 \log (3/2)$	$1/3 \log (3/2)$

Q \rightarrow How is the boy \rightarrow 0111011

Co-Sign Similarity

(s1) Hello world.

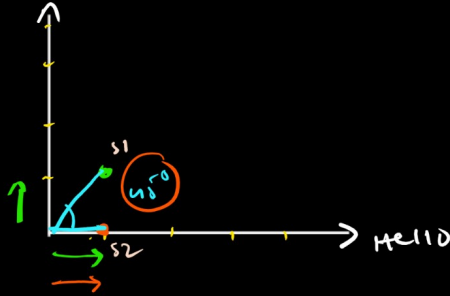
(s2) Hello.

Hello world.

(s1)	1	1	(1,1)
(s2)	1	0	(1,0)

Matrix

world.



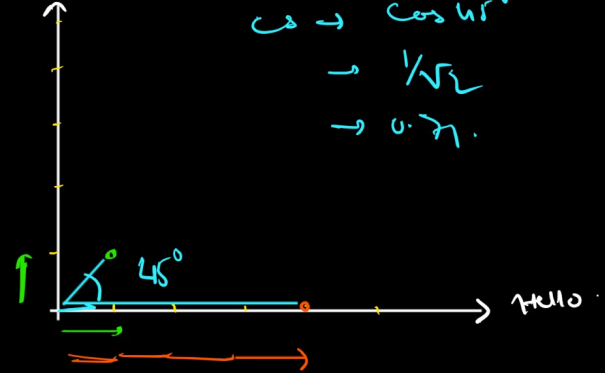
Cosign Similarity

$$\cos \theta \Rightarrow \cos 45^\circ \Rightarrow \frac{1}{\sqrt{2}} = 0.71.$$

(s1) Hello world.

(s2) Hello Hello Hello Hello.

World.



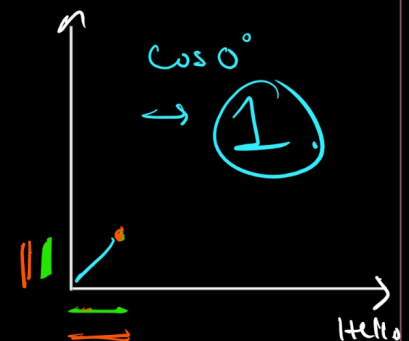
$$\begin{aligned} \cos \theta &\rightarrow \cos 45^\circ \\ &\rightarrow \frac{1}{\sqrt{2}} \\ &\rightarrow 0.71. \end{aligned}$$

(s1) Hello world

(s2) Hello world.

	H	W
(s1)	1	1
(s2)	1	1

World.



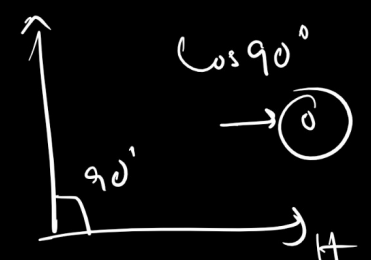
$$\cos 0^\circ \rightarrow 1.$$

(s1) Hello

(s2) world.

	H	W
(s1)	1	0
(s2)	0	1

W.



$$\cos 90^\circ \rightarrow 0.$$

0 \longrightarrow 1.

cosign Similarity Math

$$CS = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}}$$

n = no. of words.

(S1) Hello world.

(S2) Hello.

	H	W	
A	- (1)	- (1)	(S1)
B	- (1)	- (0)	(S2)

$$\frac{(1 \times 1) + (1 \times 0)}{\sqrt{1^2 + 1^2} \times \sqrt{1^2 + 0^2}}$$

$$= \frac{1}{\sqrt{2}} = (0.71)$$

(S1) How to create sales order
 (S2) How to create purchase order.

4 unique words.

create Sales order

purchase

create

sales

order

purchase.

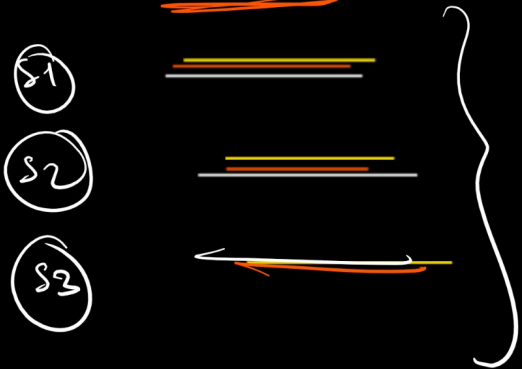
	create	sales	order	purchase.	
(S1)	1	1	1	0	A
(S2)	1	0	1	1	B

$$(1 \times 1) + (1 \times 0) + (0 \times 1) + (1 \times 1)$$

$$\sqrt{1^2 + 1^2 + 0^2 + 1^2} \times \sqrt{1^2 + 0^2 + 1^2 + 1^2}$$

$$= \frac{2}{2.99} \approx 0.66.$$

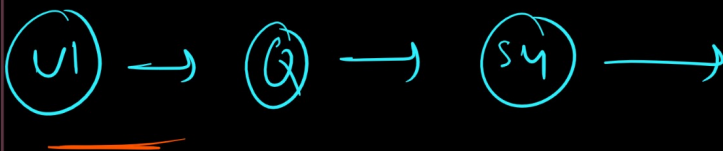
Dataset



Vectorization

$$[01, 011, 111, 011]$$

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$



$$[0111011]$$

Vectorization

Similarity Search

→ Ans.