R1 :- This pasta is tasty → positive

R2 :- This pasta is not tasty → Negative

Stop words →

	pasta	tasty
R1	1	1
R2	1	1

$V_1 = V_2$

R1 :- Pasta tasty —
R2 :- Pasta tasty — } both are same
R1 = R2

→ Sparsity
→ OOV

* Don't apply stopwords

Vocab :-

	This	pasta	is	tasty	not
R1	1	1	1	1	0
R2	1	1	1	1	1

[11110]
[11111]

Which word is more significant ?

n-grams - uni-gram (BOW)

n_gram_range = (1,1) — Bag of words

n = 1 = BOW

n = 2 = bi-grams — pair of words

n = 3 = tri-grams → 3 consecutive words

R1 : This Pasta is tasty

R2 : This Pasta is not tasty

vocab :- Bi-grams = pair of words

	This Pasta	Pasta is	is tasty	is not	not tasty
R1	1	1	1	0	0
R2	1	1	0	1	1

[1 1 1 0] V1

[1 1 0 1] V2

R1 ≠ R2

V1 ≠ V2

Data Augmentation

10 words → 100 words

n-grams

R1 This is pasta is tasty.

R2 This is pasta is not tasty.

Vocab

This is pasta is tasty not This is pasta pasta is is tasty
1 2 3 4 5 6 7 8

This is pasta is pasta is tasty pasta is not is
10 11 12

This is pasta is tasty This is pasta is not pasta is
14 15 16

TF-IDF - Term Frequency - Invers... Fr...

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

TF = probability

$$0 \leq TF \leq 1$$

Sent 1 :- nice cat

Sent 2 :- nice dog

Sent 3 :- cat dog nice

TF

Vocab	S1	S2	S3
cat	1/2	0	1/3
dog	0	1/2	1/3
nice	1/2	1/2	1/3

$$TF = \frac{\#\text{ of rep of words in a sent}}{\#\text{ of words in sent.}}$$

IDF → Inverse Documents frequency

$$IDF = \log\left[\frac{\#\ \text{Documents}}{\#\ \text{Documents containing word}}\right] + 1$$

↓

for case zero

Vocab	IDF Comes
cat	$\log(3/2)+1$
dog	$\log(3/2)+1$
nice	$\log(3/3)=\log(1)=0+1$

TF * IDF

	F1 cat	F2 dog	F3 nice
Sen1	$\frac{1}{2}\log(\frac{3}{2})+1$	$0+1$	$\frac{1}{2}*0=0+1$
Sen2	$0+1$	$\frac{1}{2}\log(\frac{3}{2})+1$	$0+1$
Sen3	$\frac{1}{3}*\log(\frac{3}{2})+1$	$\frac{1}{3}\log(\frac{3}{2})+1$	$0+1$

Freq

cat = 2

Dog = 2

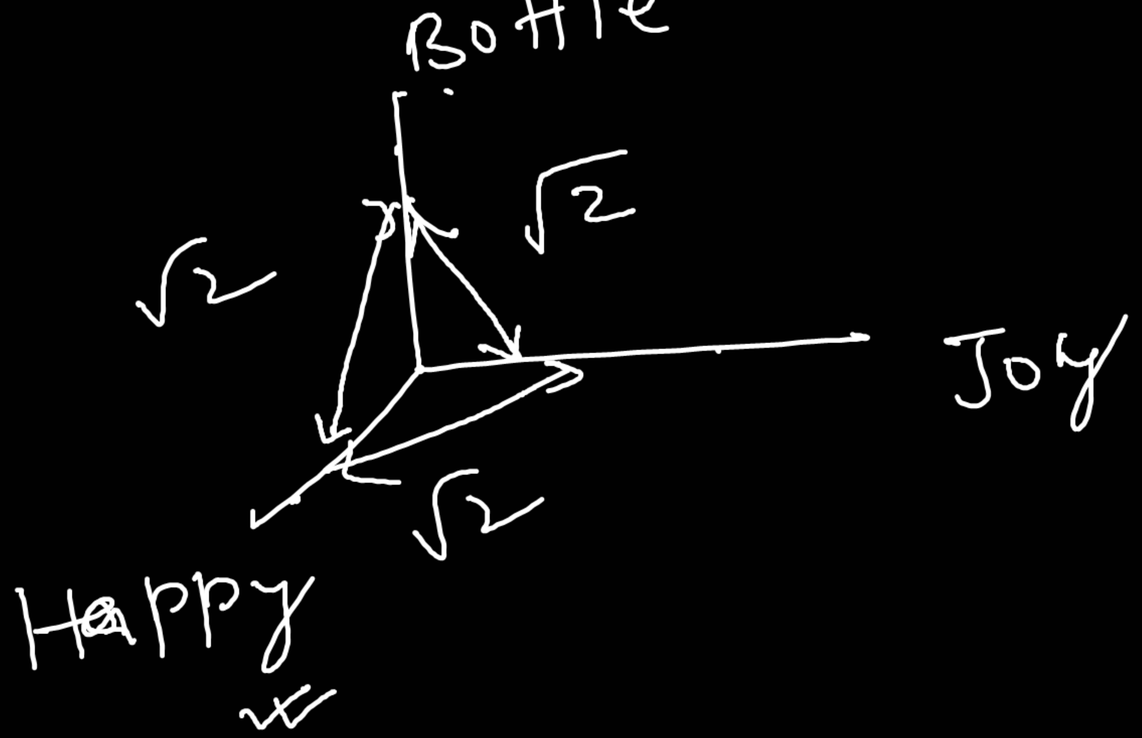
nice = 3

Adv

-> Better interpret
-> n-gram

Dis-adv

-> Sparsity
-> OOV
-> Semantic meaning



Dense
OOV
gone
Deep learning
word2vec - 2014
Google
2024
Glove — 2015