Opera
File
Edit
View
History
Bookmarks
Developer
Window
Help
1 / 5

Convert text to Numeric Vectors.

1. Bag of Words
2. Binary BoW
3. TF-IDF ✓
4. Word2Vec

} n-grams

n=1 unigrams

n=2 bi-grams

n=3 tri-grams.

crickets runs score ✓

→ D1 : I love pasta. Pasta is tasty ✓ → Bigram

D2 : I hate pasta. pasta is cheap ×

→ Trigrams

global warming, glaciers.

Size of the vectors ⇒ No. of features.

← Vocab : list of unique words 8

Vocab :	I	love	pasta	hate	is	tasty	cheap ← , ? 7
D1 :	1	1	2	0	1	1	0
BBoW	1	1	1	0	1	1	1
D2 :	1	0	2	1	1	0	1

Vocab size

	x_1	x_2	x_3	. . .	x_7	. . . .	x_{7800}	y	→
D1	1	1	2	. . .	0	. . . .		+ve	→ Classification / clustering
D2	1	0	2	1 . . 1		. . . .			

TF - IDF

→ Term Frequency - Inverse Document Frequency

$$\text{Term Frequency (TF)} = \frac{\text{No. of times the word occurs in the doc.}}{\text{Total no. of words in the doc.}}$$

$$\text{Document Frequency (DF)} = \frac{\text{No. of documents in which the word occurs.}}{\text{Total no. of documents.}}$$

$$\text{IDF} = \frac{1}{\text{DF}}$$

✓ D1 : The Car is driven on the road. ✓

✓ D2 : The truck is driven on the highway

	Term Freq					TF-IDF	
	D1	D2	DF	IDF	log(IDF)	D1	D2
The	2/7	2/7	2/2	1	0	0 -	0 -
car	1/7	0	1/2	2	0.3	0.043	0 -
truck	0	1/7	1/2	2	0.3	0	0.043
is	1/7	1/7	2/2	1	0	0	0
driven	1/7	1/7	2/2	1	0	0	0
on	1/7	1/7	2/2	1	0	0	0
road	1/7	0	1/2	2	0.3	0.043	0
highway	0	1/7	1/2	2	0.3	0	0.043

	x_1	✓ x_2	✓ x_3	x_4	x_5	x_6	x_7	x_8	y
→ D1	0	0.043	0	0	0	0	0.043	0	+ve
D2	0	0	0.043	0	0	0	0	0.043	-ve

→ (Sentiment Analysis / document Classification) →

→ Word2Vec

→ NLU

→ is there ticket available chennai to Goa

looking for 2 Seats from Chennai to Goa -

→ Deep learning models.

→ Sentence Completion (LSTM)

Many is very clever. She has got good prospects.

→ Machine translation

— grammer.

—

Deep learning

AI.

ANN →

↓

Classification/
Regression

CNN. → Convolutional-

↳ Visual Cortex

↳ RNN.

Reinforn

—
—
—
—

→ Ima-

→ Text processing

512 × 512

1024 × 1024 ⇒ 10,00 000

⇒ 1 Million

1000

Opera
File
Edit
View
History
Bookmarks
Developer
Window
Help
6 / 6

Message	Label.
How r u ?	Ham
Trip for two	spam
.	.
.	.

Vect = Countvectorizer()

⇒ Vect.fit_transform(x_1 ..)

x_1	x_2	x_3	- -	x_{7082}	Label.

clf = linearsvc()

= clf.fit( )

= clf.predict( )

Pipeline ⇒ model.(X_trn)

Mess | Lab → TF-IDF — SVC. →