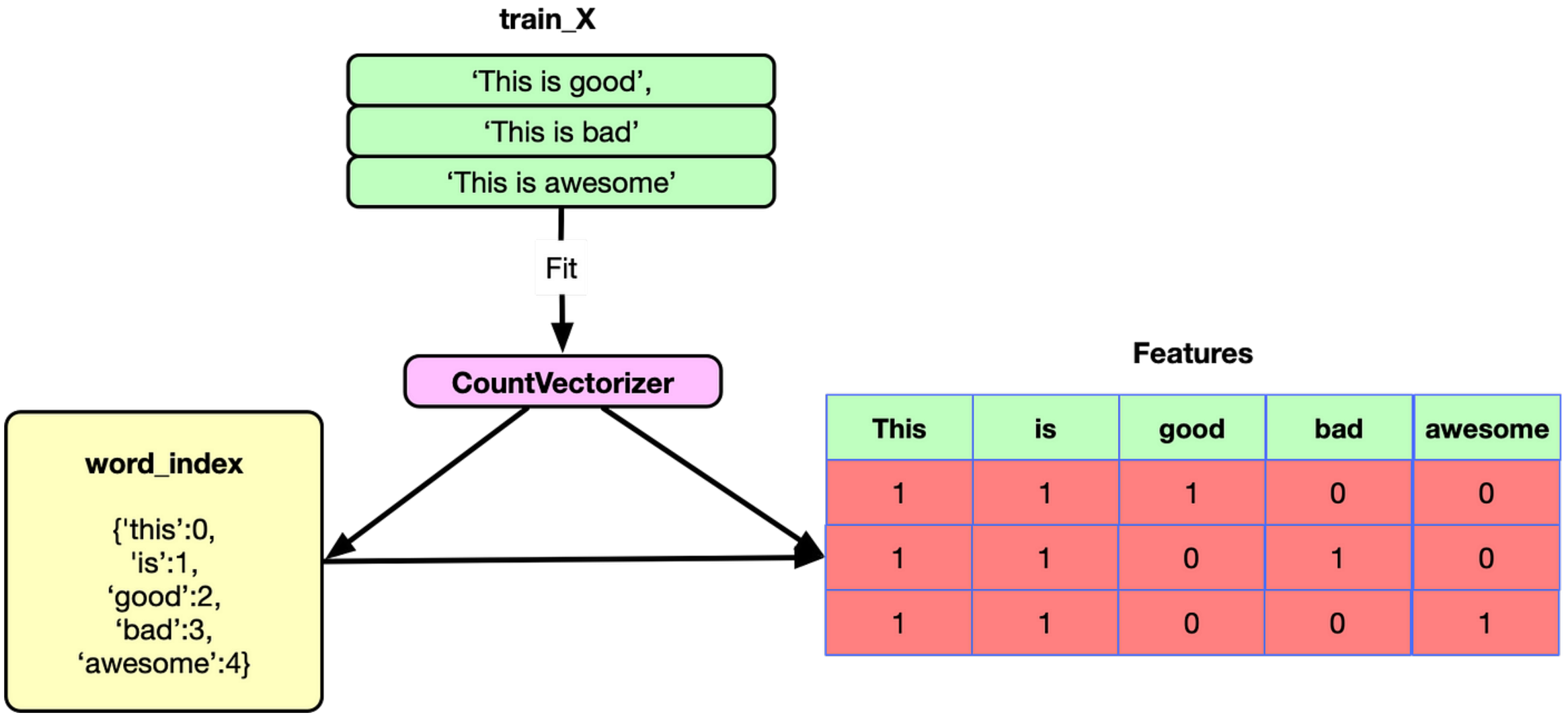
train_X
'This is good',
'This is bad'
'This is awesome'
Fit
CountVectorizer
word_index
{'this':0,
'is':1,
'good':2,
'bad':3,
'awesome':4}
Features
This
is
good
bad
awesome
1
1
1
0
0
1
1
0
1
0
1
1
0
0
1

$$TFIDF\ score\ for\ term\ i\ in\ document\ j = TF(i,j) * IDF(i)$$

where

$IDF = Inverse\ Document\ Frequency$

$TF = Term\ Frequency$

$$TF(i,j) = \frac{\text{Term i frequency in document j}}{\text{Total words in document j}}$$

$$IDF(i) = \log_2 \left(\frac{\text{Total documents}}{\text{documents with term i}}\right)$$

and

$t = Term$

$j = Document$

<TF-IDF 계산 방법>

$$TF(t, d) = \frac{\text{문서 d 에서 단어 t 가 등장한 횟수}}{\text{문서 d 에 등장한 모든 단어의 수}}$$

$$IDF(t, D) = \log \left(\frac{\text{총 문서의 개수}}{\text{단어 t를 포함하는 문서의 수}}\right)$$

TF-IDF(t,d,D) = TF(t, d) * IDF(t, D)

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0
On	1/7	1/7	log(2/2) = 0	0	0
The	1/7	1/7	log(2/2) = 0	0	0
Road	1/7	0	log(2/1) = 0.3	0.043	0
Highway	0	1/7	log(2/1) = 0.3	0	0.043