

## Count Vectorizer vs TFIDF

- 1) Count Vectorizer:

Color	
	Red
	Red
	Yellow
	Green
	Yellow

	Red	Yellow	Green
	1	0	0
	1	0	0
	0	1	0
	0	0	1

So as you can see the count vectorizer is merely count the frequency of given word in particular line. So how many unique words is in our corpus that will be our column and in given line how man times a given one is repeated.

- 2) Tfifd Vectorizer:

**Text vectorization - TF-IDF**

		<i>tokens</i>									
		face	person	guide	lock	cat	dog	sleep	micro	pool	gym
		0	1	2	3	4	5	6	7	8	9
<i>documents</i>	D1		0.05					0.25			
	D2	0.02			0.32					0.45	
...											

TF-IDF sparse matrix example

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(max_df=0.5, max_features=1000,
                             min_df=2, stop_words='english')
tfidf_corpus = vectorizer.fit_transform(text_corpus)
```

TF-IDF with scikit-learn

So, In **TfidfVectorizer** we consider **overall document weightage** of a word. It helps us in dealing with most frequent words. Using it we can penalize them. TfidfVectorizer weights the word counts by a measure of how often they appear in the documents.

- **Term Frequency:** This summarizes how often a given word appears within a document.
- **Inverse Document Frequency:** This downscals words that appear a lot across documents.

TF-IDF are word frequency scores that try to highlight words that are more interesting, e.g. frequent in a document but not across documents.

The [TfidfVectorizer](#) will tokenize documents, learn the vocabulary and inverse document frequency weightings, and allow you to encode new documents. Alternately, if you already have a learned CountVectorizer, you can use it with a [TfidfTransformer](#) to just calculate the inverse document frequencies and start encoding documents.

The same create, fit, and transform process is used as with the CountVectorizer.

### Example

```
sample = ['problem of evil',
          'evil queen',
          'horizon problem']
```

### CountVectorizer

	<b>evil</b>	<b>horizon</b>	<b>of</b>	<b>problem</b>	<b>queen</b>
<b>0</b>	1	0	1	1	0
<b>1</b>	1	0	0	0	1
<b>2</b>	0	1	0	1	0

### TfidfVectorizer

	<b>evil</b>	<b>horizon</b>	<b>of</b>	<b>problem</b>	<b>queen</b>
<b>0</b>	0.517856	0.000000	0.680919	0.517856	0.000000
<b>1</b>	0.605349	0.000000	0.000000	0.000000	0.795961
<b>2</b>	0.000000	0.795961	0.000000	0.605349	0.000000