Bag of Words
and
TFIDF

Bag of Words

- Corpus:
  - ‘I really really like dogs’
  - ‘I really really hate dogs’
  - ‘I like apples’

Bag of Words

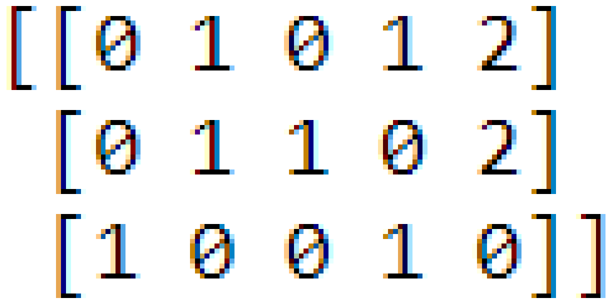
- Corpus:
  - ‘I really really like dogs’
  - ‘I really really hate dogs’
  - ‘I like apples’

```
['apples', 'dogs', 'hate', 'like’, 'really']
```

Bag of Words

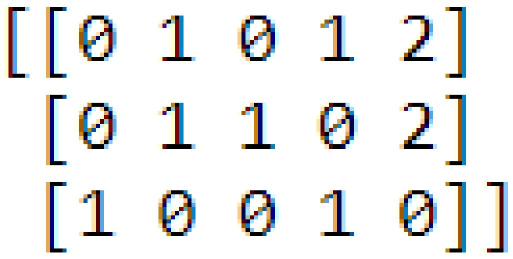
- Corpus:
  - ‘I really really like dogs’
  - ‘I really really hate dogs’
  - ‘I like apples’

```
['apples', 'dogs', 'hate', 'like’, 'really']
```



```
[[0 1 0 1 2]
 [0 1 1 0 2]
 [1 0 0 1 0]]
```

TFIDF (Term Frequency Inverse Document Frequency)



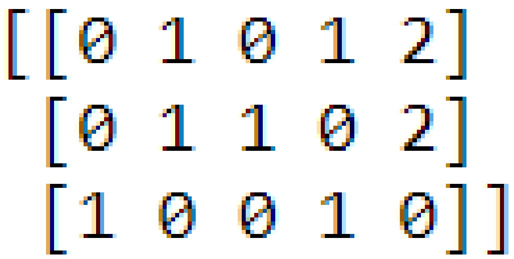
```
[[0 1 0 1 2]
 [0 1 1 0 2]
 [1 0 0 1 0]]
```

```
['apples', 'dogs', 'hate', 'like’, 'really']
```

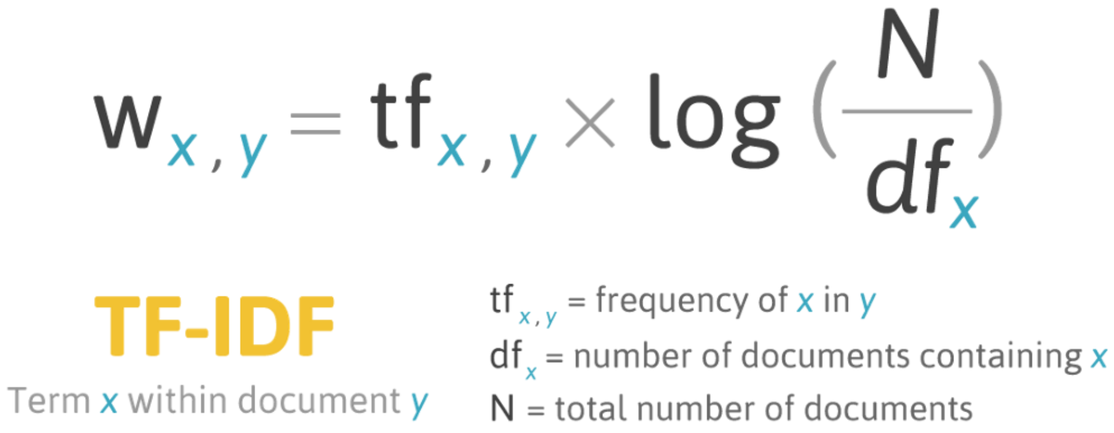
Corpus:

‘I really really like dogs’
‘I really really hate dogs’
‘I like apples’

TFIDF (Term Frequency Inverse Document Frequency)



```
[[0 1 0 1 2]
 [0 1 1 0 2]
 [1 0 0 1 0]]
```



$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

```
['apples', 'dogs', 'hate', 'like’, 'really']
```

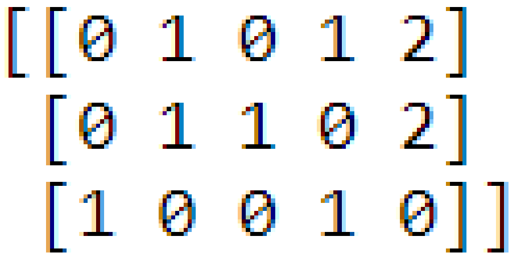
Corpus:

‘I really really like dogs’

‘I really really hate dogs’

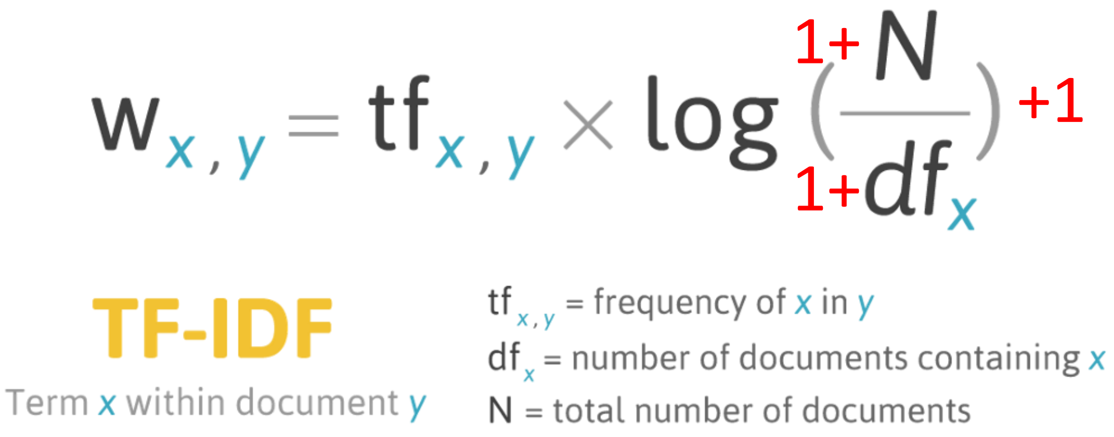
‘I like apples’

TFIDF (Term Frequency Inverse Document Frequency)



```
[[0 1 0 1 2]
 [0 1 1 0 2]
 [1 0 0 1 0]]
```

```
['apples', 'dogs', 'hate', 'like', 'really']
```



$$w_{x,y} = tf_{x,y} \times \log\left(\frac{1+N}{1+df_x}\right) +1$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Corpus:

‘I really really like dogs’

‘I really really hate dogs’

‘I like apples’

Example 1:

y = I like apples

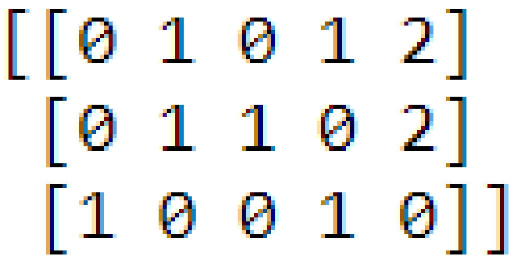
x = apples

$W_{x,\,y} = 1 * (\log_e(1+3/1+1) + 1) = 1.69314$

x = like

$W_{x,\,y} = 1 * (\log_e(1+3/1+2) + 1) = 1.28768$

TFIDF (Term Frequency Inverse Document Frequency)



```
[[0 1 0 1 2]
 [0 1 1 0 2]
 [1 0 0 1 0]]
```

```
['apples', 'dogs', 'hate', 'like’, 'really']
```

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{1+N}{1+df_x}\right)+1$$

TF-IDF

Term *x* within document *y*

$tf_{x,y}$ = frequency of *x* in *y*

df_x = number of documents containing *x*

N = total number of documents

Corpus:

‘I really really like dogs’

‘I really really hate dogs’

‘I like apples’

L2 normalization:

$$\hat{\mathbf{u}} = \frac{\mathbf{u}}{|\mathbf{u}|}$$

Example 1:

y = I like apples

x = apples

$W_{x,\,y}$ = 1 * (log$_e$(1+3/1+1) + 1) = 1.69314

x = like

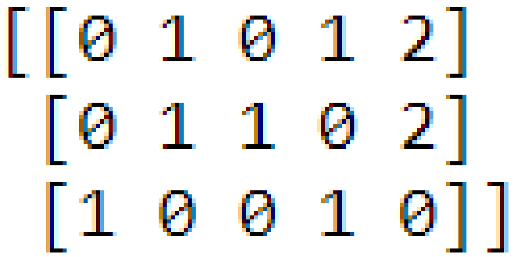
$W_{x,\,y}$ = 1 * (log$_e$(1+3/1+2) + 1) = 1.28768

u = [1.693 0 0 1.288 0]

|u| = root(1.693^2 + 0^2 + 0^2 + 1.288^2 + 0^2) = 2.127

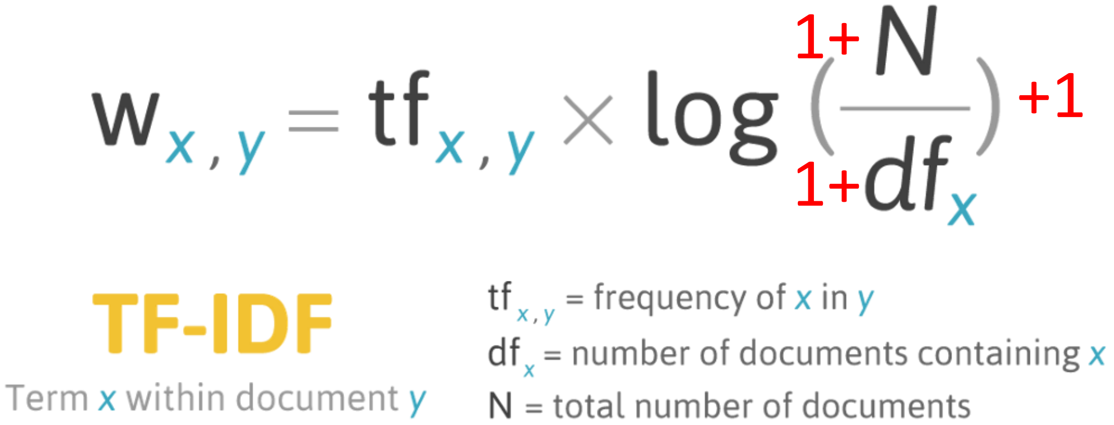
û = [0.80, 0, 0, 0.61, 0]

TFIDF (Term Frequency Inverse Document Frequency)



```
[[0 1 0 1 2]
 [0 1 1 0 2]
 [1 0 0 1 0]]
```

```
['apples', 'dogs', 'hate', 'like’, 'really']
```



$$w_{x,y} = tf_{x,y} \times \log\left(\frac{1+N}{1+df_x}\right) + 1$$

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Corpus:

‘I really really like dogs’

‘I really really hate dogs’

‘I like apples’

L2 normalization:

$$\hat{\mathbf{u}} = \frac{\mathbf{u}}{|\mathbf{u}|}$$

Example 2:

y = I really really like dogs

x = dogs

$W_{x,\,y} = 1 * (\log_e(1+3/1+2) + 1) = 1.288$

x = like

$W_{x,\,y} = 1 * (\log_e(1+3/1+2) + 1) = 1.288$

x = really

$W_{x,\,y} = 2 * (\log_e(1+3/1+2) + 1) = 2.575$

u = [0, 1.288, 0, 1.288, 2.575]

$|u| = \text{root}(0^2 + 1.288^2 + 0^2 + 1.288^2 + 2.575^2) = 3.154$

û = [0, 0.41, 0, 0.41, 0.82]