TF-IDF weighting

- TF stands for *term frequency*
- IDF stands for *inverse document frequency*
- Given a collection of documents, D, a word, w, and a document from the collection, $d \in D$, the relevance score, $W_{d,w}$, assigned to that word-document pair is shown below:

$$W_{d,w} = (1 + \log f_{D,w}) \cdot \log \frac{|D|}{f_{d,w}}$$

where $f_{d,w}$ = number of documents containing term w

$f_{D,w}$ = frequency of term w in corpus D

- There are several variants of this weighting scheme, the above is most popular.

- TF-IDF score for a query (q) - document (d) pair:

$$score_{q,d} = \sum_{w \in q \cap d} W_{d,w}$$

Pointwise Mutual Information (PMI)

- Measures the probability of two words, w_1, w_2, being found in same document, $d \in D$, normalized by the product of each term's probability of being found in a document.

$$PMI(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \qquad where \qquad p(w) = \frac{\sum_{d \in D} \mathbf{1}\{w \in d\}}{|D|}$$

$$p(w_1, w_2) = \frac{\sum_{d \in D} \mathbf{1}\{w_1, w_2 \in d\}}{|D|}$$

- Because the word-document matrix is sparse, many of the PMI matrix entries end up being large negative numbers; these values aren't meaningful. For this reason it is common to only consider the positive entries in the matrix using Positive Pointwise Mutual Information (PPMI):

$$PPMI(w_1, w_2) = \max\left(PMI(w_1, w_2), 0\right)$$