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):

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

Shortcomings of lexical-based similarity measures

- Invariant to sequence order: randomly shuffling the order of the words in each document yields the same result.

- Doesn’t capture word synonymy, polysemy, or context

- Highly sensitive to text processing and tokenization scheme that is used

- Given the sparsity of the term document matrix, should we expect any distance-based similarity metric to be meaningful?