

# TF-IDF and Hard Information Metrics

## TF-IDF Equation

### Unigrams

For libraries ending with “unigram,” the term frequency-inverse document frequency (TF-IDF) is calculated using Loughran and McDonald’s (2011) formula for unigrams:

$$\omega_{ij} = \begin{cases} \frac{(1+\log(\text{tf}_{i,j}))}{(1+\log(\alpha_i))} \log\left(\frac{N}{\text{df}_i}\right) & \text{if } \text{tf}_{i,j} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Where: -  $N$  represents the total number of documents in the corpus, -  $\text{df}_i$  is the number of documents containing at least one occurrence of the  $i$ -th unigram, -  $\text{tf}_{i,j}$  is the raw count of the  $i$ -th unigram in the  $j$ -th document, -  $\alpha_i$  is the average unigram count in the document.

The log transformation helps to attenuate the impact of high-frequency unigrams, and the term  $\log\left(\frac{N}{\text{df}_i}\right)$  adjusts the impact based on the unigram’s commonality (Loughran & McDonald, 2011).

### Bigrams

For libraries ending with “bigram,” the TF-IDF calculation is adjusted for bigrams using Loughran and McDonald’s (2011) formula:

$$\omega_{ij} = \begin{cases} \frac{(1+\log(\text{tf}_{i,j}))}{(1+\log(\alpha_i))} \log\left(\frac{N}{\text{df}_i}\right) & \text{if } \text{tf}_{i,j} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Where: -  $N$  represents the total number of documents in the corpus, -  $\text{df}_i$  is the number of documents containing at least one occurrence of the  $i$ -th bigram, -  $\text{tf}_{i,j}$  is the raw count of the  $i$ -th bigram in the  $j$ -th document, -  $\alpha_i$  is the average bigram count in the document.

The log transformation helps to attenuate the impact of high-frequency bigrams, and the term  $\log\left(\frac{N}{\text{df}_i}\right)$  adjusts the impact based on the bigram’s commonality (Loughran & McDonald, 2011).

### Mixed N-gram Library

For libraries that do not end with “unigram” or “bigram” and are identified solely by the library name, the TF-IDF formula is adjusted to include all keywords, whether they are unigrams, bigrams, or trigrams:

$$\omega_{ij} = \begin{cases} \frac{(1+\log(\text{tf}_{i,j}))}{(1+\log(\alpha_i))} \log\left(\frac{N}{\text{df}_i}\right) & \text{if } \text{tf}_{i,j} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Where: -  $N$  represents the total number of documents in the corpus, -  $\text{df}_i$  is the number of documents containing at least one occurrence of the  $i$ -th keyword (either a unigram or a bigram), -  $\text{tf}_{i,j}$  is the raw count

of the  $i$ -th keyword (either a unigram or a bigram) in the  $j$ -th document, -  $\alpha_i$  is the average unigram count in the document.

The log transformation helps to attenuate the impact of high-frequency words/bigrams, and the term  $\log\left(\frac{N}{df_i}\right)$  adjusts the impact based on the keyword’s commonality (Loughran & McDonald, 2011).

## Hard Information

We adopt the methodology of Campbell et al. (2024) to quantify hard information in 10-K reports, as detailed in their study, “Number of Numbers: Does Quantitative Textual Disclosure Reduce Information Risk” (Campbell et al., 2024). Hard information is defined as the proportion of quantitative data within the annual report (10-K report) or within a specific section of the report, such as the business section or management discussion and analysis section.

The formula for measuring hard information is:

$$\text{Hard Information} = \frac{N(\text{Numbers})}{N(\text{Numbers}) + N(\text{Words})}$$

Where: -  $N(\text{Numbers})$  represents the total count of numerical data (numbers) in the 10-K report, -  $N(\text{Words})$  represents the total word count of the report.