TF-IDF

Inverse document frequency (IDF)

- $N = |D|$ – total number of documents in corpus
- $|\{d \in D: t \in d\}|$ – number of documents where the term t appears
- $\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D: t \in d\}|}$

TF-IDF

- $\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$
- A high weight in TF-IDF is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents