# Representation Revisited

- Document-term frequency matrices work!

- However, there are some issues:
  - Longer documents will tend to have higher term counts.
  - Terms that appear frequently across the corpus aren't as important.

- We can improve upon this representation if we can achieve the following:
  - Normalize documents based on their length.
  - Penalize terms that occur frequency across the corpus.

# Term Frequency

- Let *freq(t, d)* be the count of the instances of the term $t$ in document $d$.
- Let *TF(t, d)* be proportion of the count of term t in document $d$.
- Mathematically, where $n$ is the number of distinct terms in document $d$:

$$TF(t,d) = \frac{freq(t,d)}{\sum_i^n freq(t_i,d)}$$

# Inverse Document Frequency

- Let $N$ be the count distinct documents in the corpus.

- Let $count(t)$ be the count of documents in the corpus in which the term $t$ is present.

$$IDF(t) = \log\left(\frac{N}{count(t)}\right)$$

# The Mighty TF-IDF

- Combine $TF$ and $IDF$ to enhance document-term frequency matrices:

$$TF\text{-}IDF(t,d) = TF(t,d) * IDF(t)$$