Variant tf-idf functions
**Sublinear tf scaling**

“It seems unlikely that twenty occurrences of a term in a document truly carry twenty times the significance of a single occurrence.”

\[
w_{t,d} = \begin{cases} 
  1 + \log tf_{t,d} & \text{if } tf_{t,d} > 0 \\
  0 & \text{otherwise}
\end{cases}.
\]

\[
w_{f-idf_{t,d}} = w_{f_{t,d}} \times idf_t.
\]
Maximum tf normalization

- tf weights of all terms in a document are normalized by maximum term frequency
- $a$: a smoothing term to weaken the contribution of the second term, usually set to 0.4
- $ntf_{t,d} = a + (1 - a) \frac{tf_{t,d}}{tf_{\text{max}}(d)}$
- Motivation: to mitigate that longer documents with higher term frequency get higher relevance than shorter documents with same content.
Document and query weighting schemes

\[
\text{score}(q, d) = \frac{\vec{V}(q) \cdot \vec{V}(d)}{|\vec{V}(q)||\vec{V}(d)|}.
\]

<table>
<thead>
<tr>
<th>Term frequency</th>
<th>Document frequency</th>
<th>Normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>n (natural)</td>
<td>n (no)</td>
<td>n (none)</td>
</tr>
<tr>
<td>l (logarithm)</td>
<td>t (idf)</td>
<td>c (cosine)</td>
</tr>
<tr>
<td>a (augmented)</td>
<td>p (prob idf)</td>
<td>u (pivoted unique)</td>
</tr>
<tr>
<td>b (boolean)</td>
<td></td>
<td>b (byte size)</td>
</tr>
<tr>
<td>L (log ave)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{tf}_{t,d} & = 1 + \log(\text{tf}_{t,d}) \\
\text{tf}_{t,d} & = 0.5 + \frac{0.5 \times \text{tf}_{t,d}}{\max_{t}(\text{tf}_{t,d})} \\
\begin{cases} 
1 & \text{if } \text{tf}_{t,d} > 0 \\
0 & \text{otherwise}
\end{cases} \\
1 + \log(\text{avg}_{t}(\text{tf}_{t,d}))/\log(q) \\
\end{align*}
\]

\[
\begin{align*}
\text{idf} & = \log \frac{N}{\text{df}_t} \\
\text{prob idf} & = \max\{0, \log \frac{N-\text{df}_t}{\text{df}_t}\}
\end{align*}
\]

Mnemonic for weight combination: ddd.qqq

Example: Inc.ltc

- Document vector
- Query vector
- Log weighted tf
- No idf
- Cosine normalization
Pivoted normalized document length

Long docs:
- verbose relative weights of terms not affected
- multi-topic relative weights of terms affected

Compensation: make doc. length independent of terms and doc. freqs
Pivoted normalized document length

Probability of relevance = doc. length averaged over all queries
Pivoted normalized document length

For each bucket in documents by length: compute the fraction of relevant documents. Plot this fraction against the median document length of each bucket.

Cosine normalization distorts relevance vis-à-vis the true relevance, at the expense of longer documents.
Pivoted normalized document length

Rotate the cosine normalisation curve counter clockwise to better match the relevance per document length curve.

\[ a|\vec{V}(d)| + (1 - a)piv \\
au_d + (1 - a)piv \]
Pivoted document length normalization problems

When relevance does not depend on doc. length, or when this dependency is too complex,

use doc. length as an ML feature

\[
\text{score}(q, d) = \frac{\langle \vec{V}(q) \cdot \vec{V}(d) \rangle}{|\vec{V}(q)||\vec{V}(d)|}
\]