Many studies have used the phrase “term tf weighting” from Robertson’s work, and the second (often called pivoted normalization weighting) from the work done at Cornell University. Most research groups at TREC currently use some variant of these two weightings. Many studies have used the phrase “term tf-Idf weighting” to refer to any term weighting method that uses the probabilistic models developed term weighting schemes which were shown to be effective on the small test collections available then. Inception of TREC provided IR researchers with very large and varied test collections allowing rapid development of effective weighting schemes.

Soon after first TREC, researchers at Cornell University realized that using raw tf of terms is non-optimal, and a dampened frequency (e.g., a logarithmic tf function) is a better weighting metric. [4] In subsequent years, an effective term weighting scheme was developed under a probabilistic model by Steve Robertson and his team at City University, London. [22] Motivated in part by Robertson’s work, researchers at Cornell University developed better models of how document length should be factored into term weights. [29] At the end of this rapid advancement in term weighting, the field had two widely used weighting methods, one (often called Okapi weighting) from Robertson’s work, and the second (often called pivoted normalization weighting) from the work done at Cornell University. Most research groups at TREC currently use some variant of these two weightings. Many studies have used the phrase “term tf-Idf weighting” to refer to any term weighting method that uses tf and idf, and do not differentiate between using a simple document scoring method (like \( \sum_{t \in Q,D} tf \cdot ln\frac{N}{df} \)) and a state-of-the-art scoring method (like the ones shown in Table 1). Many such studies claim that their proposed methods are far superior than “term tf-Idf weighting”, often a wrong conclusion based on the poor weighting formulation used.

### Okapi weighting based document score: [23]

\[
\sum_{t \in Q,D} \frac{N - df}{df + 0.5} \cdot \frac{(k_1 + 1)tf}{(1 - b) + b \cdot \frac{dl}{avdl}} + \frac{(k_3 + 1)qtf}{k_3 + qtf}
\]

This term should be after \( k_1 \)

\( k_1 \) (between 1.0–2.0), \( b \) (usually 0.75), and \( k_3 \) (between 0–1000) are constants.

### Pivoted normalization weighting based document score: [30]

\[
\sum_{t \in Q,D} \frac{1 + ln(1 + ln(tf))}{1 - s + s \cdot \frac{dl}{avdl}} \cdot qtf \cdot ln\frac{N + 1}{df}
\]

\( s \) is a constant (usually 0.20).

Table 1: Modern Document Scoring Schemes