# A Appendix

Table 1 shows the QQ metrics selected as features from the paper Predicting Query Quality for Applications of Text Retrieval to Software Engineering Tasks text(Mills et al., 2017). The metrics we choose are divided into two categories: specificity and coherency. The specificity of the query is reflected by the query terms' distribution over the collection. The coherency of a query is usually measured as the level of inter-similarity between the documents in the collection containing at least one of the query terms.

Table 1: Selected Query Quality (QQ) metrics

| **Measure** | **Description** | **Formula** |
|---|---|---|
| | **Specificity** | |
| AvgIDF | Average of the Inverse Document Frequency (idf)[1] values over all terms | $\frac{1}{\lvert Q \rvert} \sum_{q \in Q} idf(q)$ |
| MaxIDF | Maximum of the Inverse Document Frequency (idf) values over all query terms | $\max_{q \in Q}(idf(q))$ |
| DevIDF | The standard deviation of the Inverse Document Frequency (idf) values over all query terms | $\sqrt{\frac{1}{\lvert Q \rvert} \sum_{q \in Q} (idf(q) - AvgIDF)}$ |
| AvgICTF | Average Inverse Collection Term Frequency (ictf)[2] values over all query terms | $\frac{1}{\lvert Q \rvert} \sum_{q \in Q} ictf(q)$ |
| MaxICTF | Maximum Inverse Collection Term Frequency (ictf) values over all query terms | $\max_{q \in Q}(ictf(q,d))$ |
| DevTCTF | The standard deviation of the Inverse Collection Term Frequency (ictf) values over all query terms | $\sqrt{\frac{1}{\lvert Q \rvert} \sum_{q \in Q} (ictf(q) - AvgICTF)}$ |
| AvgEntropy | Average entropy[3]values over all query terms | $\frac{1}{\lvert Q \rvert} \sum_{q \in Q} entropy(q)$ |
| MedEntropy | Median entropy values over all query terms | $\underset{q \in Q}{median}(entropy(q))$ |

| | | |
|---|---|---|
| MaxEntropy | Maximum entropy values over all query terms | $\max_{q\in Q} Entropy(q)$ |
| DevEntropy | The standard deviation of the entropy values over all query terms | $\sqrt{\frac{1}{\lvert Q\rvert}\sum_{q\in Q}(entropy(q) - AvgEntropy)}$ |
| Query Scope (QS) | The percentage of documents in the collection containing at least one of the query terms | $\frac{\lvert \cup_{q\in Q} D_q\rvert}{\lvert D\rvert}$ |
| Simplified Clarity Score (SCS) | The Kullback-Leiber divergence of the query language model from the collection language model[4] | $\sum_{q\in Q} P_q(Q)\log\left(\frac{p_q(Q)}{p_q(D)}\right)$ |
| **Coherency** | | |
| AvgVAR | Average of the variances of the query term weights over the documents containing the query term (VAR)[5], over all query terms | $\frac{1}{\lvert Q\rvert}\sum_{q\in Q} VAR(q)$ |
| MaxVAR | Maximum of the variances of the query term weights over the documents containing the query term (VAR), over all query terms | $\max_{q\in Q}(VAR(q))$ |
| SumVAR | Sum of the variances of the query term weights over the documents containing the query term (VAR), over all query terms | $\sum_{q\in Q} VAR(q)$ |
| Coherence Score (CS) | The average of the pairwise similarity between all pairs of documents containing one of the query terms among all documents in the corpus | $\frac{1}{\lvert Q\rvert}\sum_{q\in Q}\left(\frac{\sum_{(d_i,d_j)\in D_q} sim(d_i,d_j)}{\lvert D_q\rvert\cdot(\lvert D_q\rvert - 1)}\right)$ |

| | | |
|---|---|---|
| $q$ - a term in the query; | $D$ - the set of documents in the collection; | $d$ - a document in the document collection $D$; |
| $Q$ - the set of query terms; | $D_t$ - the set of documents containing term $t$; | $tf(t, D)$ - the frequency of term $t$ in all docs; |
| $tf(t, d)$ - the frequency of term $t$ in $d$; | $tf(t, Q)$ - the frequency of term $t$ in the query; | $sim(d_i, d_j)$ - the cosine similarity between the vector-space representations of $d_i$ and $d_j$ |

## References

Mills C, Bavota G, Haiduc S, et al., 2017. Predicting query quality for applications of text retrieval to software engineering tasks. *ACM Trans Softw Eng Methodol*, 26(1):3:1-3:45.
https://doi.org/10.1145/3078841

[1] $idf(t) = \log\left(\frac{|D|}{|D_t|}\right)$

[2] $ictf(t) = \log\left(\frac{|D|}{tf(t,D)}\right)$

[3] $entropy(t) = \sum_{d \in D_t} -\frac{tf(t,d)}{tf(t,D)} \cdot \log_{|D|} \frac{tf(t,d)}{tf(t,D)}$

[4] $p_t(X) = \frac{tf(t,X)}{|X|}$

[5] $VAR(t) = \sqrt{\frac{\sum_{d \in D_t}(w(t,d) - \overline{w}_t)^2}{df(t)}}$,where $w(t, d) = \frac{1}{|d|} \log(1 + tf(t, d)) \cdot idf(t)$, and $\overline{w}_t = \frac{1}{|D_t|} \sum_{d \in D_t} w(t, d)$