

Summary

- Smoothing of $p(w|d)$ is necessary for query likelihood
- General idea: smoothing with $p(w|C)$
 - The probability of an unseen word in d is assumed to be proportional to $p(w|C)$
 - Leads to a general ranking formula for query likelihood with TF-IDF weighting and document length normalization
 - Scoring is primarily based on sum of weights on matched query terms
- However, how exactly should we smooth?