Representing and Accessing [Textual] Digital Information (COMS/INFO 630), Spring 2006
2/21/06: **Lecture guides for 2/2/06 (lecture three)**

Authors: Siavash Dejgosha and Ricardo Hu (first 11 pages); Nick Hamatake and Min Park (subsequent pages).

Some additional questions on the topic of the lecture are below. (Not to say that there's anything wrong with the questions posed in what's attached; these are just some other thoughts that people raised in class.)

1. Recall the proposal in class of an alternate method to simplify the pivoted length normalization equation. The proposal was motivated by observing that if b' is a constant with respect to d, then

$$\begin{aligned} \mathrm{norm}'(d) &= m'\mathrm{norm}(d) + b' \\ &\overset{\mathrm{rank}}{=} \frac{m'}{b'}\mathrm{norm}(d) + 1. \end{aligned}$$

   If we adopt this "non-pivoted" approach — note that the pivot never made an appearance — is there a "pedagogically lucid" way to "insert" the corpus-average normalization factor $\overline{\mathrm{norm}}$ so as to get a intuitive normalization for documents of "appropriate" (that is, average) length (with respect to normalization)?[1]

2. As was asked in class, is it possible to adopt something similar to Zobel's "arrival rate" technique — or to do something else entirely — to estimate whether the length distribution of relevant documents in the TREC sampled pool is significantly biased with respect to the length distribution of relevant documents among the entire corpus?

[1] Of course, it's not completely clear that the derivation in class could be classified as "pedagogically lucid", either.