Representing and Accessing [Textual] Digital Information (COMS/INFO 630), Spring 2006
2/7/06: **Lecture guides for 1/26/06 (lecture one)**

Authors: Asif-ul Haque and Benyah Shaparenko (first eight pages); Ari Rabin and Victoria Krafft. I think that the two guides complement each other well.

Note: The problems in the attached lecture guides assume term weights of the form

$$d_j = \frac{(\text{frequency of } v^{(j)} \text{ in } d) \cdot n/(\text{number of documents containing } v^{(j)})}{\sqrt{\sum_j \left[(\text{frequency of } v^{(j)} \text{ in } d) \cdot n/(\text{number of documents containing } v^{(j)})\right]^2}}$$

where n is the number of documents in the corpus. Remember that this is not the only possible or necessarily best-performing "tf-idf" weight. For example, we could impute the idf quantities to the *query's* term weights, as in a later lecture, so that the normalization is based only on the document at hand. Or, we could make use of more complex notions of term frequency and inverse document frequency, for example:

$$d_j = \frac{(1 + \ln(1 + \ln(\text{frequency of } v^{(j)} \text{ in } d)))(\ln((1+n)/(\text{number of documents containing } v^{(j)}))}{(1-s) + s\frac{\text{length of } d \text{ in bytes}}{\text{average byte-length of documents in } C}}$$

used by AT&T at TREC-7 (Singhal, Choi, Hindle, Lewis, Pereira 1999; Singhal 2001). Note the use of pivoted length normalization, which we discuss in later lectures, and the double damping of the raw term frequency.