Representing and Accessing [Textual] Digital Information (COMS/INFO 630), Spring 2006
2/14/06: **One lecture guide for 1/31/06 (lecture two)**

Note: this cover sheet is a version that was updated on 2/21/06 for posting on the course homepage.

Authors: Chris Danis and Brian Rogan. (The other group took on some interesting empirical experiments that we are still working out the details of.)

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

1. Recall our discussion of the normalization function $norm^B(d)$ (using the notation from the attached lecture-guide's problems), defined as $\max_j tf_j(d)$. In class, it was suggested that one could use a normalization function similar in spirit but potentially less sensitive to outlier counts: use the kth largest raw term frequency as normalization function. However, this would introduce a free parameter.

   What would be the effect of using the *average* raw term frequency in the document as normalization term, noting that doing so avoids introducing an extra parameter?

2. What we termed "cosine normalization" might also be termed "L_2" normalization. Recall that under the L_2 norm, the length of a vector $\vec{x} \in \Re^m$ is given by $\sqrt{\sum_{j=1}^{m} x_j^2}$.

   You might ask what the effect of considering other norms would be. For example, the L_1 norm of $\vec{x}$ is given by $\sum_{j=1}^{m} |x_j|$. One way to think of L_1 normalization is as converting the vector entries into parameters for a multinomial distribution; at any rate, it certainly means that the vector's entries sum to one (when the entries are non-negative and at least one is positive).

   Show that the rankings produced by L_1 normalization are equivalent to those produced by L_2 normalization, assuming the commonly-used VSM scoring function $\cos(\vec{q}, \vec{d})$.