Representing and Accessing [Textual] Digital Information (COMS/INFO 630), Spring 2006
3/9/06: **One lecture guide for 2/21/06 (lecture eight)**

Note: this cover sheet is a version that was updated on 3/10/06 for posting on the course homepage.

Authors: Peter Babinski and David Lin. (The other guide for this lecture is forthcoming.)

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

1. Recall that with the specific language-model induction method we considered, ranking by query likelihood means that the score of a given document d with respect to a query q is based on the following:

$$\prod_j \left[\frac{tf_j(d) + \mu \frac{tf_j(C)}{|C|}}{|d| + \mu} \right]^{tf_j(q)} . \tag{1}$$

   We were at first alarmed to note an apparently "anti-IDF" quantity appearing, namely, $tf_j(C)/|C|$. Then, we noted that through some ranking-preserving algebraic manipulations (following Zhai and Lafferty (2001), who made the observation that "the use of $[tf_j(C)/|C|]$ as a reference smoothing distribution has turned out to play a role very similar to the well-known IDF"), we were able to cause an "IDF-like" term to appear (if $\mu \neq 0$). However, some of us found it alarming that we should be able to switch from an "anti-IDF"- to an "IDF"-based scoring function solely through mathematically justified rewritings.

   Resolve the mystery by explaining why $tf_j(C)/|C|$ in Equation 1 never was "anti-IDF" in the first place, in the sense that the more frequent $v^{(j)}$ is in the corpus, the less the frequency of $v^{(j)}$ in d affects the ranking that d receives according to Equation 1 relative to the other documents in the corpus.

2. Show that the "two-state Hidden Markov Model" language model induces a proper probability distribution over all possible term sequences (the zero-length sequence should also be included).

   Hint: first show that the probability mass assigned to the entire set of length-k sequences, for fixed k, is $s(1-s)^k$, where s is the "stopping" probability.

References

Zhai, Chengxiang and John D. Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR*, pages 334–342.