Representing and Accessing [Textual] Digital Information (COMS/INFO 630), Spring 2006
3/9/06: **Lecture guide for 2/16/06 (lecture seven)**

Authors: Chris Danis and Brian Rogan. (There is only one guide for this lecture.)

The first of the additional questions below addresses a common confusion.

1. Recall the specific example of language-model estimation we considered in lecture, in which for a particular document d, the language model induced from d has parameters

$$\theta_j(d) = \frac{tf_j(d) + \mu \frac{tf_j(C)}{|C|}}{|d| + \mu}.$$

Suppose that we have the following two-document corpus:

d':	there isn't any there there
d'':	there you go again

where $v^{(1)}$ is "there" and $v^{(2)}$ is "any", and consider the following argument:

Set $\mu = 1$. *We see that*

$$\begin{aligned} \theta_1(d') &= \frac{3 + 1 \cdot \frac{4}{2}}{5+1} = \frac{5}{6} \\ \theta_2(d') &= \frac{1 + 1 \cdot \frac{1}{2}}{5+1} = \frac{1}{4} \end{aligned}$$

But this means that $\theta_1(d') + \theta_2(d') > 1$. *So the estimation method must be incorrect.*

What is wrong with this argument?

2. (*adapted from a question by Siavash Dejgosha and Ricardo Hu*) We've discussed scoring documents by the probability assigned to the query by the language model induced from the documents (this is commonly known as scoring by *query likelihood*), where the induction method is that described above and in lecture.

What ranking of the documents results if the query contains a term that does not appear in the corpus (but happens to be in the term index, let us say)? Can we fix the problem solely by altering our method of language-model estimation in some reasonable fashion?