CS630 Lecture Practice Problems

Lecturer: Lillian Lee
Scribes: Chris Danis (cgd3) & Brian Rogan (bcr6)

Lecture 7: 16 February 2006

1. Provide an intuition on how the language modeling approach to IR differs from the earlier (RSJ) model that we discussed in class.

   ANSWER:
   The language modeling approach considers the probability of a user providing a specific query given that the user's information need is satisfied and we're considering a certain document. In contrast, the RSJ model models the probability of generating a specific document given that a certain query was specified and the user's information need is satisfied.

2. Recall that when we were deriving the language modeling scoring function in class, an intermediate step of the equation was:

   $$P(Q = q|R = y, D = d)P(R = y|D = d)$$

   During the derivation we assumed we could either ignore the $P(R = y|D = d)$ term, or assume that R and D were independent, which allowed us to ignore this term under rank. We did admit, however, that this term could be important under certain circumstances. Provide an example where the inclusion of the $P(R = y|D = d)$ term could be useful.

   ANSWER:
   As we discussed in lecture, $P(R = y|D = d)$ provides us a measure of a document's intrinsic probability of satisfying the user's information need independent of a specific query (the a priori probability of relevance). There are a variety of circumstances under which this could be useful. The Internet is one environment where such a term could be very useful, and in fact Google's "PageRank" technology serves a function very much like this term: certain documents are considered to be better answers to the set of all possible queries because they are considered to be "more important" or "reputable" documents in terms of the entire corpus. In a sense, you can think of this term as a "credibility" term, and it can be used to normalize a collection where some documents are considered to be vastly more or less important than others. Of course, other applications are possible.

3. Why do we say that the term $\frac{tf_j(C)}{|C|}$ feels "anti-idf" in nature?

   ANSWER:
   The idea of using a *idf* term is that documents which contain terms that occur in few other documents (and therefore are distinctive) should receive more weight in ranking than documents which contain terms that are common to many documents. The term above provides exactly the opposite intuition: it rewards terms which are present in many documents over those that occur in very few documents.

4. In lecture we also said that the above term $\left(\frac{tf_j(C)}{|C|}\right)$ makes sense in a generative model, but as a ranking model it seems illogical. We discussed why it was illogical as a ranking model in the last question. Why does it make sense as a generative model?

   ANSWER:
   Recall that that in our multinomial topic model, the term $\frac{tf_j(C)}{|C|}$ is part of the term θ_j. To put it in context it is part of the equation:

$$P_{\vec{\theta}(d)}(\vec{q}) = f(\vec{q}, d) \cdot k(\vec{q}) = k(\vec{q}) \prod_j \left(\frac{tf_j(d) + \mu \times tf_j(C)/|C|}{|d| + \mu} \right)$$

   We gave the intuition that $f(\vec{q}, d)$ was the probability of seeing the "sorted version" of q in document d. In this case, the probability a term of the query is contained in the document should increase proportionally to the likeliness of seeing that that term in the corpus as a whole.