Representing and Accessing [Textual] Digital Information (COMS/INFO 630), Spring 2006
3/16/06: **Midterm**

Do not begin (i.e., don't look beyond this cover page) until told you may do so.

This exam consists of five main questions. Therefore, if you find yourself taking more than 15 minutes on a question, you may want to go on and come back to that question later.

Write your answers in the exam booklet(s) provided. **Show and explain all work**: solutions without adequate explanation will be considered incomplete.

Please turn in this set of questions with your exam booklet. (You will get it back; it's just that sometimes people accidentally write on the question sheets instead of in the "blue"book.)

Your name (please print): _____

(1) This question concerns the RSJ document scoring function. Assume binary attribute variables. Suppose we were to estimate $P(A_j = 1|R = y)$ for terms $v^{(j)}$ **that are in the query** by

$$\frac{N^{(j)} + \alpha}{N + \alpha}, \tag{EQ-1}$$

where $N^{(j)}$ is the number of documents containing $v^{(j)}$, N is the total number of documents, and α is a non-negative constant. (Note that the denominator is just a normalization factor.)

a) Give a short (roughly three-sentence) justification for estimate EQ-1. *Hints*: you may use the fact that EQ-1 is increasing in α; also, try first considering the case where $\alpha = 0$.

b) Either briefly (one to three sentences) explain how EQ-1 differs in a substantive way from the Croft/Harper estimate, or explain why the two estimates are essentially equivalent.

(A sample answer that does not satisfy the "substantive" requirement: "The Croft/Harper estimate depends on different quantities". A sample, probably incorrect answer that *does* satisfy the "substantive" requirement: "EQ-1 assumes that term frequencies follow a Poisson distribution, whereas Croft/Harper assumes that term counts follow a Gaussian distribution.")

c) Derive the scoring function that results if we plug estimate EQ-1 for $P(A_j = 1|R = y)$ into the RSJ model (and use the estimates from class for the quantities not directly related to $P(A_j = 1|R = y)$). Remember to provide adequate explanations.

d) Is the scoring function you derived substantively different from the one that follows from using the Croft/Harper assumption? Explain. (Your answer to the above subproblem must be reasonable to receive credit for this subproblem.)

---

(2) Here, we modify the setting that resulted in our second derivation of the LM-based approach. Assume a finite set of document-topic language models $t_1, t_2, \ldots, t_n$, where the parameters for each t_i are known. Suppose that the system is issued a query whose semantics is, "A document is relevant if it was generated by t_1 **or** by t_2". You should consider the query to be fixed and to be not a term sequence and hence not "generatable" by an LM (for instance, perhaps the system gets information requests through the user clicking on some checkboxes).

a) Derive a scoring function that results from expanding $P(R = y|D = d)$ based on the information just given, where it is required that most, if not all, of the quantities in your function can be directly estimated in a reasonable way. For each quantity, be sure to explain either how you would estimate it, justifying your choice, or why a problem arises (despite good-faith effort on your part) in estimating it.

Hint: note that topic LMs may "generate" documents that are not in the corpus.

b) Is there any advantage in this case to deriving a scoring function from

$$\frac{P(R = y|D = d)}{P(R = n|D = d)} \tag{EQ-2}$$

instead of $P(R = y|D = d)$? Briefly explain your answer. Your derivation above must be correct to receive credit for this subproblem.

(3) Here we consider a way to incorporate Poisson distributions, discussed in class in the context of classic probabilistic retrieval, with the LM-based approach.

Let the query q be fixed, and for simplicity assume that q and all the documents in the corpus are of the same length. Suppose we have a method that produces, for each document d, a value $\mu_j(d)$ for the parameter of a Poisson distribution induced from d for the term frequency of $v^{(j)}$; hence, we have

$$P_d(A_j = a) = \frac{1}{e^{\mu_j(d)}} \frac{[\mu_j(d)]^a}{a!}. \tag{EQ-3}$$

And, our **scoring policy** is to score a document d by taking the product over all $v^{(j)}$ of the probability according to P_d of seeing $v^{(j)}$ occur $\mathrm{tf}_j(q)$ times.

a) Briefly explain why the following choice is justifiable:

$$\mu_j(d) = \mathrm{tf}_j(d) + \lambda \frac{\mathrm{tf}_j(C)}{|C|} \tag{EQ-4}$$

where λ is a non-negative constant. Your answer should indicate that you understand the intuitive "meaning" of the parameter for a Poisson.

b) Show that the scoring policy described above gives rise to a scoring function in which an "idf-like" quantity is readily apparent (please indicate it in your answer). Remember to show all work and provide adequate explanations.

Hint: for the purposes of this problem, treat $\prod_j e^{\mu_j(d)}$ as a normalization factor; call it $\mathrm{norm}(d)$ and separate it out as much as possible, as early as possible.

---

(4) Here, we re-examine one of the considerations behind the derivation of the Okapi/BM25 weighting function.

For simplicity, assume that all terms have the same IDF — hence, ignore IDF in your work — and that no term occurs in the query twice.

a) What were the criteria behind the (initial) choice of the following quantity

$$\frac{\mathrm{tf}_j(d)}{k + \mathrm{tf}_j(d)} \tag{EQ-5}$$

as the term-frequency portion of the Okapi/BM25 scoring function?

Your answer should consist of properties of functions (e.g., "non-negativity") and the reason these properties are required (e.g., "because the quantity is to be plugged into a log function").

b) Consider the usual VSM scoring function with cosine (L_2) normalization, ignoring the IDF quantities. Explain whether or not the VSM term weights (which you should give explicitly) satisfy the same criteria as you specified in the previous subproblem, and thus can be justified on the same grounds as EQ-5. (Your previous response must be correct to receive credit for this subproblem.)

(5) This question concerns length normalization in the VSM.

Let us use "hat" and "double-hat" notation as follows: for a vector $\vec{x}$, $\widehat{\vec{x}}$ denotes the L_1-normalized version and $\widehat{\widehat{\vec{x}}}$ denotes the L_2-normalized version. Furthermore, for simplicity, assume that for a document d we set $\vec{d} = (\mathrm{tf}_1(d), \mathrm{tf}_2(d), \ldots, \mathrm{tf}_m(d))^T$. Explain how scoring a document d with the function

$$\vec{q} \cdot \widehat{\vec{d}} \tag{EQ-6}$$

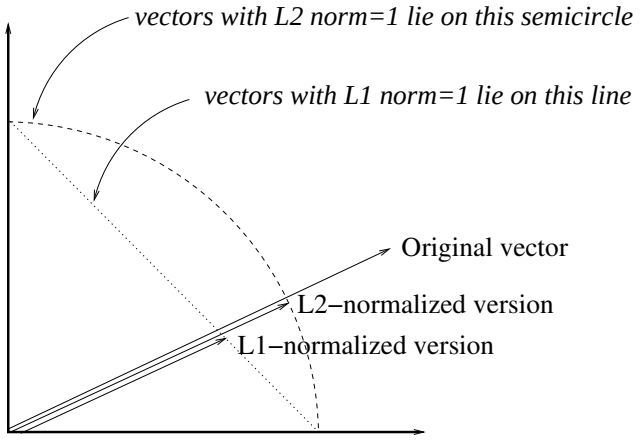
penalizes certain types of long documents more than scoring with the function

$$\vec{q} \cdot \widehat{\widehat{\vec{d}}} \tag{EQ-7}$$

does. Explicitly specify what types of documents incur the extra penalty, and discuss whether the extra penalization has a beneficial or a harmful effect.

Hints: choose (or draw) two specific document vectors d_1 and d_2 such that their relative ranking with respect to some query vector changes depending on which scoring function is used.

Also, note that for any vector $\vec{x}$, $||\vec{x}||_2 \leq ||\vec{x}||_1$, with the maximum difference between the two norms occurring when $\vec{x} = (\beta, \beta, \ldots, \beta)$ for some non-zero constant β, and the minimum difference (namely, 0) happening when exactly one of the entries of $\vec{x}$ is non-negative. In two dimensions, we can see this in the following figure:



– End of exam. Have a good break! –