Due: 2018/05/11 (before class)

Homework 9

Problem 1. (20 points) In Rocchio's algorithm, what weight setting for $\alpha/\beta/\gamma$ does a "Find pages like this one" search correspond to?

Solution:
Rocchio's formula:

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

Reasonable values might be $\alpha = 1$ $\beta = 0.75$ $\gamma = 0.15$
So in our case
$\alpha = 0$ $\beta = 1$ $\gamma = 0$

Problem 2. (20 points) Why is positive feedback likely to be more useful than negative feedback to an IR system? Why might only using one nonrelevant document be more effective than using several?

Solution:
Relevance feedback has been shown to be very effective at improving relevance of results. Its successful use requires queries for which the set of relevant documents is medium to large. Full relevance feedback is often onerous for the user, and its implementation is not very efficient in most IR systems. In many cases, other types of interactive retrieval may improve relevance by about as much with less work.
Positive feedback means data has been found by the information retrieval system. Hence the extracted data can be used in may ways, A negative feedback implies that no data found so the IR fails.
The idea of feedback system to tell the IR system which documents are relevant to the user and to maximize the return of such documents. Even if the IR system is not explicitly told that documents d1, d2, etc. are nonrelevant, the IR system tries to maximize the precision based upon the relevant documents feedback which is more important and sufficient in most cases. If several nonrelevant documents are used for feedback calculation, some of them might bear some similarity with a few relevant documents (assuming that the user marks them as nonrelevant based on a few words or sentences depicted and doesnt process sufficient knowledge about them). In such a case, some of the properties of relevant documents might not be conveyed properly to the IR system which results in low precision output. There are low chances of such a problem if only 1 non- relevant is used.

Problem 3. (30 points) Suppose that a user's initial query is "cheap CDs cheap DVDs extremely cheap CDs". The user examines two documents, d1 and d2 . She judges d1 , with the content "CDs cheap software cheap CDs" relevant and d2 with content "cheap thrills DVDs" nonrelevant. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length-normalize vectors. Using Rocchio relevance feedback as in Equation below what would the revised query vector be after relevance feedback? Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$.

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

Solution:
Term Frequencies:

word	query q	d_1	d_2
CD's	2	2	0
cheap	3	2	1
DVDs	1	0	1
extremely	1	0	0
software	0	1	0
thrills	0	0	1

For $1.0 * \underset{q}{\rightarrow} + 0.75 * \underset{d}{\rightarrow} + 1 - 0.25 * \underset{d_2}{\longrightarrow}$
we have,
$(\frac{7}{2}\ \frac{17}{4}\ \frac{3}{4}\ 1\ \frac{3}{4} - \frac{1}{4})^T \implies (3.5\ 4.25\ 0.75\ 1\ 0.75\ \ 0.25)^T$
Negative weights are set to 0. The Rocchio vector thus is:
$(3.5\ 4.25\ 0.75\ 1\ 0.75\ 0)^T$

Problem 4. (30 points) Omar has implemented a relevance feedback web search system, where he is going to do relevance feedback based only on words in the title text returned for a page (for efficiency). The user is going to rank 3 results. The first user, Jinxing, queries for: "banana slug". And the top three titles returned are:
1. banana slug Ariolimax columbianus
2. Santa Cruz mountains banana slug
3. Santa Cruz Campus Mascot

Jinxing judges the first two documents Relevant, and the third Not Relevant. Assume that Omar's search engine uses term frequency but no length normalization nor IDF. Assume that he is using the Rocchio relevance feedback mechanism, with $\alpha = \beta = \gamma = 1$. Show the final revised query that would be run. (Please list the vector elements in alphabetical order.)

Solution:

	query q	d_1	d_2	d_3
Ariolimax	0	1	0	0
banana	1	1	1	0
Campus	0	0	0	1
columbianus	0	1	0	0
Cruz	0	0	1	1
Mascot	0	0	0	1
mountains	0	0	1	0
Santa	0	0	1	1
slug	1	1	1	0

Using Rocchio relevance feedback mechanism, with $\alpha = \beta = \gamma = 1$
After changing negative components back to 0, we get:
$\underset{qr}{\longrightarrow} = [\frac{1}{2}, 2, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, 2]$