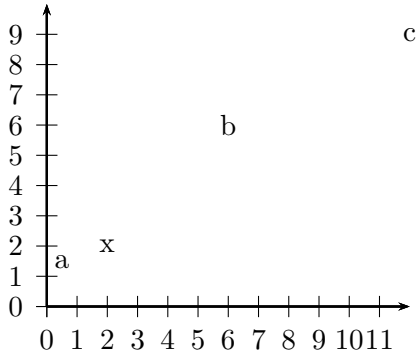
Information Retrieval: Assignment 3

Problem 1. (10 points)

In the figure below, which of the three vectors $\vec{a}$, $\vec{b}$, and $\vec{c}$ is (i) most similar to $\vec{x}$ according to dot product similarity ($\sum_i x_i \cdot y_i$), (ii) most similar to $\vec{x}$ according to cosine similarity ($\sum_i x_i \cdot y_i)/(|x||y|)$), (iii) closest to $\vec{x}$ according to Euclidean distance ($|x - y|$)? The vectors are $\vec{a} = (0.5 \;\; 1.5)^T$, $\vec{x} = (2 \;\; 2)^T$, $\vec{b} = (6 \;\; 6)^T$, and $\vec{c} = (12 \;\; 9)^T$.



Problem 2. (12 points)

Compute the lnc.ltn similarity between the query "John Miller" and the document "John Miller, John Fisher, and one other John" by filling out the empty columns in the table below. Treat "and", "one", and "other" as stop words. N = 100,000,000. df_{John} = 50000, $\text{df}_{\text{Fisher}}$ = 100000, $\text{df}_{\text{Miller}}$ = 10000. What is the final similarity score? What is the Jaccard similarity score between query and document?

	query					document				
word	tf-raw	tf-wght	df	idf	weight	tf-raw	tf-wght	weight	n'lized	product
...	...	...	...	...	...	...	...	...	...	...

Problem 3. (5 points)

If we were to only have one-term queries, explain why the use of weight-ordered postings lists (i.e., postings lists ordered according to weight, instead of docid) truncated at position k in the list suffices for identifying the k highest scoring documents. Assume that the weight w stored for a document d in the postings list of t is the cosine-normalized weight of t for d.

Problem 4. (10 points)

Below is a table showing how two human judges rated the relevance of a set of documents to a particular information need (0 = nonrelevant, 1 = relevant). Let us assume that you've written an IR system that for this query returns the set of documents {2, 5, 6, 7, 8}.

docID	1	2	3	4	5	6	7	8	9	10	11	12
Judge 1	0	0	1	1	1	1	1	1	0	0	0	1
Judge 2	0	0	1	1	0	0	0	0	1	1	1	0

(i) Calculate the kappa measure between the two judges. (ii) Calculate precision, recall, and F_1 of your system if a document is considered relevant only if the two judges agree it is relevant. (iii) Calculate precision, recall, and F_1 of your system if a document is considered relevant if either judge thinks it is relevant.

Problem 5. (5 points)

In Rocchio's algorithm, what weight setting for $\alpha/\beta/\gamma$ does a "find pages like this one" search (the "similar" link that appears for mouse-over on the URL on most Google search results) correspond to?

Problem 6. (10 points)

Suppose that a user's initial query is "cheap CDs cheap DVDs extremely cheap CDs". The user examines two documents, d_1 and d_2. She judges d_1, with the content "CDs cheap software cheap CDs" relevant and d_2 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 9.3 (page 182) what would the revised query vector be after relevance feedback? Assume $\alpha = 1, \beta = 0.75, \gamma = 0.25$.

<u>Due date:</u> Thursday, June 6, 2013, 12:15

Please turn in your assignment in class if possible. Email submissions are only accepted if you have a good reason why you cannot attend the review meeting.