
Information Retrieval and Text Mining: Assignment 3

Problem 1. (10 points)

One measure of the similarity of two vectors is the Euclidean distance between them: $|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^M (x_i - y_i)^2}$. Given a query q and documents d_1, d_2, \dots , we may rank the documents d_i in order of increasing Euclidean distance from q . Show that if q and the d_i are all normalized to unit vectors, then the rank ordering produced by Euclidean distance is identical to that produced by cosine similarities.

Problem 2. (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 3. (10 points)

Find an example of a Google search where (i) on the measure of regular relevance, each of the first 10 hits is relevant to the query and (ii) on the measure of marginal relevance, at least one hit has zero relevance. Submit a printout of the results of the Google search and mark the marginally nonrelevant result.

Problem 4. (10 points)

Run a query on Google with at least 10 hits and analyze the dynamic summaries of the first 10 hits. Make at least three suggestions as to how one could improve the dynamic summaries returned by Google.

Problem 5. (10 points)

In Rocchio's algorithm, what weight setting for $\alpha/\beta/\gamma$ does a "Find pages like this one" search 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 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 what would the revised query vector be after relevance feedback? Assume $\alpha = 1, \beta = 0.8, \gamma = 0.2$.

Due date: Tuesday, June 9, 2009, 15:45 (turn assignments in before class in V38.03 or by email)