SIMS 202 Assignment 4
Due Tuesday October 9th

Based on an assignment by Prof. Marti Hearst

Please bring a hardcopy of your assignment to class on Tuesday October 9th. There are two questions in total (both with a number of sub-questions).

Introduction

This assignment assumes that you are familiar with the meaning of sigma notation. For example,

$$n = 10; \qquad s = \sum_{i=0}^{n-1} i$$

means s gets assigned the sum of all the integers from 0 to 9, inclusive, or $0 + 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 = 45$. The index is i and its boundaries are from 0 to $n - 1$.

As in recent assignments in Prof. Hearst's course on the foundations of Software Design.

1 Computing Term Weights

For a collection C consisting of N documents, consider the following term weight formulae:

$w_{ik} = tf_{ik} * idf_k$
$idf_k = log(N/n_k)$

where

T_k = term k in collection C
tf_{ik} = frequency of term T_k in document D_i
n_k = the number of documents in C that contain term T_k
f_k = the total frequency of term T_k in all documents of C
M = the number of unique terms in C
idf_k = inverse document frequency of term T_k in collection C
w_{ik} = the weight of term T_k in document D_i

Be sure to show your work.

(a) It is always the case that $f_k \geq n_k$. Briefly explain why.

(b) Say the term "user" occurs in the document D_1 twelve times and in the collection C in five documents, and that the collection consists of 100 documents. What is the weight of this term in this document?

Now say for the same collection the term "user" occurs in the document D_2 one time. What is the weight of this term in this document?

2 Computing Document Similarity

(a) Assume the documents D_1, D_2, and D_3, which have the following characterstics:

Document D_1 contains "information" 20 times and "retrieval" 3 times.

Document D_2 contains "information" 1 time and "retrieval" 15 times.

Document D_3 contains "information" 12 times and "retrieval" 10 times.

Also assume that

"information" occurs in 120 documents in the collection

"retrieval" occurs in 70 documents in the collection

N is 1000

Draw a graph showing the vectors for the raw frequency counts. Place "information" on the x-axis and "retrieval" on the y axis.

(b) Assume the query consists of the two words "information" and "retrieval". Compute the similarity value between the query and each of the documents D_1, D_2, and D_3.

To compare the similarity of two documents, or a document and a query (where the query is viewed as a document) use the weighting formula below to compute each w_{ik} and the following similarity comparison formula.

(This weighting formula normalizes the term weights.)

$$w_{ik} = \frac{tf_{ik} * log(N/n_k)}{\sqrt{\sum_{k=0}^{M-1} (tf_{ik})^2 * [log(N/n_k)]^2}}$$

$$sim(D_i, D_j) = \sum_{k=0}^{M-1} w_{ik} * w_{jk}$$

(Hint: if a term does not occur in a document or query then its weight is zero.)

Be sure to show your work. Discuss the results briefly.

(c) Draw a graph showing the normalized vectors for the documents (represent the documents in terms of their normalized weights). Place "information" on the x-axis and "retrieval" on the y axis. Also draw the vector for the query. Does the graph correspond with your results for part (b)? How is this related to part (a)?

(d) What would the results above look like if we just used tf for the term weights, without multiplying by idf? What would they look like if "information" had occurred in 400 documents in the collection instead of 120?