SIMS 202 Assignment 6
Due Wednesday November 26

Please place a hardcopy of your assignment in Prof. Hearst's mailbox in the office on the first floor, or else bring a hardcopy to class on Tuesday November 25th. There are four questions in total.

Practice with Sigma Notation

(1) Recall the semantics 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 another example

$$n = 3; \qquad s = \sum_{i=1}^{n} a_i * a_{i+1}$$

means s is assigned the sum of $a_1 * a_2 + a_2 * a_3 + a_3 * a_4$. And

$$n = 3; \qquad s = \sum_{i,j=1}^{n} a_i * b_j$$

means s is assigned the sum of $a_1 * b_1 + a_2 * b_2 + a_3 * b_3$.

For the problems below you may use a calculator or computer if you like. You may want to show the main intermediate stages of the computation if you're unsure about how to do the work.

Compute s for the following three formulas (be sure to check the boundaries for the indices).

(a) $n = 10; \qquad s = \sum_{i=0}^{n-1} i^2$

(b) $m = 10; \qquad s = \sum_{j=1}^{m} -j$

(c) $n = 4; \qquad a_i = i; \qquad b_j = j + 2; \qquad s = \sum_{i,j=0}^{n-1} a_i * b_j$

Computing Term Weights

(2) For a collection C consisting of N documents, use 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

(a) It is always the case that $f_k >= n_k$. Why is this?

(b) Assume 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?

(c) Assume the term "user" occurs in the document D_2 one time 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?

Computing Document Similarity

(3) To compare the similarity of two documents, or a document and a query (where the query is viewed as a document) use the following similarity comparison formula:

$$sim(D_i, D_j) = \frac{\sum_{k=0}^{M-1} w_{ik} * w_{jk}}{\sqrt{\sum_{k=0}^{M-1} (w_{ik})^2 * \sum_{k=0}^{M-1} (w_{jk})^2}}$$

Use the weighting formulae from (2).

Say we have a query consisting of the term "information" and the term "retrieval", and that

n_k for "information" is 120
n_k for "retrieval" is 100
N is 1000.

(a) Compute the similarity value between the query and each of the documents D_1, D_2, and D_3, which have the following characterstics:

Document D_1 contains "information" 3 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" 11 times.

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

(b) Discuss the results briefly.

(c) Draw a graph showing the vectors for the queries and the documents. Place "information" on the x-axis and "retrieval" on the y axis (you can show either the tf or the w. Also draw the vector for the query. Does the graph correspond with your results for part (a)?

(d) What would the results above look like if we just used tf for the term weights, without multiplying by idf?

Computing Precision and Recall

(4) You are an information technologist asked to decide which of three systems to choose for your client. Along with cost factors and assessment of the user interface, you want to assess the relative strengths of the systems' ranking algorithms (called Bear, Cardinal, and Wolf).

Assume you know, for a document collection and a set of queries, what all the relevant documents in the collection are. You only have binary (yes/no) relevance judgements for each (query, document) pair. You also know the order in which the systems rank the documents. This information appears on the next two pages (and can be found online in an excel spreadsheet on the NT server Newt in the Groups folder in is202 $->$ Excel $->$ prec-recall.xls).

(a) For each system, compute the average precision (over all queries) at recall intervals of 20%. (That is, precision at 20% recall, 40% recall,..., 100% recall.) Show the results in a table, and also graph the results for all three systems on one plot (connect the points for each system with a smooth line).[1] You can either draw the graph by hand or use a program to help you.

(b) Now compute the average precision (over all queries) at four different document cutoff levels, showing the results in a table. Choose cutoff levels that help facilitate comparison of the three systems.

(c) Based on these results, which ranking algorithm do you recommend?

[1] Don't worry about how to interpolate between the points.

Relevance judgements for 25 documents for three queries. 1 indicates relevant, 0 indicates not relevant.

DocId	Query0	Query1	Query2
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0
5	1	0	0
6	0	0	0
7	0	0	0
8	0	1	0
9	0	1	0
10	0	1	0
11	0	1	0
12	0	1	0
13	0	0	0
14	0	0	1
15	0	0	1
16	0	0	1
17	0	0	1
18	0	0	1
19	0	0	0
20	0	0	0
21	0	0	0
22	0	0	0
23	0	0	0
24	0	0	0
25	0	0	0

Ordering of 25 documents for each of three queries by each of three ranking algorithms. Documents are identified by document ID. The document shown on row 1 is the highest rank, that on row 2 is the second-highest ranked, etc. Assume there are no ties.

	Query 0			Query 1			Query 2		
	Bear	Cardinal	Wolf	Bear	Cardinal	Wolf	Bear	Cardinal	Wolf
1	5	7	11	12	16	7	1	5	23
2	2	16	2	14	23	23	18	24	9
3	1	9	10	8	3	3	11	13	25
4	15	18	21	23	13	13	16	19	18
5	9	4	1	9	4	4	19	17	1
6	19	17	5	5	14	8	10	12	2
7	3	8	22	20	17	17	23	22	11
8	6	5	9	21	11	25	21	4	10
9	18	11	20	4	7	16	3	1	19
10	4	15	3	19	21	21	6	8	16
11	12	12	23	3	18	18	22	11	12
12	8	10	4	10	10	22	12	3	22
13	11	6	19	7	19	19	17	16	15
14	17	2	6	18	22	10	2	23	8
15	7	19	15	11	20	20	13	25	24
16	20	20	25	17	1	5	24	18	14
17	10	21	18	24	2	2	25	2	7
18	21	25	16	13	12	12	14	20	13
19	16	22	7	2	15	15	5	14	17
20	23	1	8	16	5	1	4	10	20
21	13	23	17	6	8	14	7	6	21
22	22	13	12	22	6	24	9	21	4
23	24	24	13	15	9	9	15	15	5
24	25	3	14	25	25	11	20	7	6
25	14	14	24	1	24	6	8	9	3