

SIMS 202 Assignment 7

Due Tuesday November 10

Please bring a hardcopy of your assignment to class on Tuesday November 10th. There are three questions in total.

Practice with Sigma Notation

(1) Recall the semantics of sigma notation. For example,

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

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

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

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

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

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).

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

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

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

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 $f_k \geq n_k$. 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:

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

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 70

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 characteristics:

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.

(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 ?