# CSCI 6370 IR and Web Search

## Homework Assignment One

Total points: 100. Due: Monday, June 8.

## Problem

This assignment is designed for you to get familiar with the basic vector space model in IR. You are to calculate measures such as tf-idf and similarity between a query and a set of documents using different measures, given some basic statistic about the documents and the index terms (key words).

**Note:** It might be easier to either use a spread-sheet program or write a program to do the computation and the sorting.

1. Table 1 lists the index terms and their appearances in a set of documents.

Table 1: Term Frequencies

| **Doc/Term** | **retrieval** | **database** | **computer** | **text** | **information** |
| --- | --- | --- | --- | --- | --- |
| D1 | 4 | 10 | 2 | 0 | 1 |
| D2 | 3 | 0 | 7 | 4 | 5 |
| D3 | 7 | 2 | 4 | 6 | 8 |

We also know that the total number of documents in the set is 1000. Table 2 shows the document frequencies of these terms.

Table 2: Document Frequencies

| **Term** | **retrieval** | **database** | **computer** | **text** | **information** |
| --- | --- | --- | --- | --- | --- |
| **Frequency** | 100 | 70 | 220 | 80 | 110 |

Compute tf-idf for each of the (doc, term) pairs listed in Table 1. List your results in sorted order from the largest value of tf-idf to the smallest value.

Table 3: Term Frequency-Inverted Document Frequencies

| **Term** |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **tf / idf** |  |  |  |  |  |

2. Assume we use the tf-idf as the weight in the vector space model, write down the document-term matrix using the results generated from the above problem. Remember a document-term matrix has terms as its columns and documents as its rows.

3. Now assume we have a query Q = “computer information”, compute the similarity based on the inner product similarity and the cosine similarity for each of the documents listed in Table 1. Which document is the most relevant in each of the similarity measures? Which one is the least relevant?