WST540: Exercise List 1

Consider that our document collection S has the following documents: $D_1, ..., D_5$:

document	words
D_1	Data Base System Concepts
D_2	Introduction to Algorithms
D_3	Computational Geometry: Algorithms and Applications
D_4	Data Structures and Algorithm Analysis on Massive Data Sets
D_5	Computer Organization

Our dictionary $DICT$ consists of 8 words: $\{w_1 = \text{data}, w_2 = \text{system}, w_3 = \text{algorithm}, w_4 = \text{computer}, w_5 = \text{geometry}, w_6 = \text{structure}, w_7 = \text{analysis}, w_8 = \text{organization}\}$. We consider that, by stemming, "computer" and "computational" are regarded as the same word, and so are "algorithms" and "algorithm".

Problem 1. Let $tf(w, D)$ denote the term frequency of term w in a document D as defined in our lecture notes. Give the value of $tf(w_i, D_j)$ for all $1 \leq i \leq 8$ and $1 \leq j \leq 5$.

Problem 2. Let $idf(w)$ denote the inverse document frequency of term w as defined in our lecture notes. Give the value of $idf(w_i)$ for all $1 \leq i \leq 8$.

Problem 3. Convert each document in S into an 8-dimensional point according to the tf-idf model as defined in our lecture notes.

Problem 4. Assume that we have received a query with terms "Geometry Algorithm Concepts". Convert the query to an 8-dimensional point.

Problem 5. Rank the documents in descending order of their relevance to the query in Problem 4 according to the cosine metric.