WST540: Quiz 1

Consider that our document collection S has the following documents: $D_1, ..., D_4$:

document	words
D_1	Information retrieval is an important subject.
D_2	The Johnson family has got a golden retriever.
D_3	Information theory uses plenty of theorems from mathematics.
D_4	It provides a golden opportunity for information sharing.

Our dictionary $DICT$ consists of 8 words: $\{w_1 =$ information, $w_2 =$ retrieval, $w_3 =$ subject, $w_4 =$ Johnson, $w_5 =$ golden, $w_6 =$ theory, $w_7 =$ mathematics, $w_8 =$ sharing$\}$. By stemming, "retrieval" and "retriever" are regarded as the same word, and so are "theory" and "theorem".

Problem 1. Let $tf(w, D)$ denote the term frequency of term w in a document D. Give the value of $tf(w_i, D_j)$ for all $1 \le i \le 8$ and $1 \le j \le 4$.

Solution.

	D_1	D_2	D_3	D_4
w_1	1	0	1	1
w_2	1	1	0	0
w_3	1	0	0	0
w_4	0	1	0	0
w_5	0	1	0	1
w_6	0	0	2	0
w_7	0	0	1	0
w_8	0	0	0	1

Problem 2. Let $idf(w)$ denote the inverse document frequency of term w. Give the value of $idf(w_i)$ for all $1 \le i \le 8$.

Solution.

w_1	0.415
w_2	1
w_3	2
w_4	2
w_5	1
w_6	2
w_7	2
w_8	2

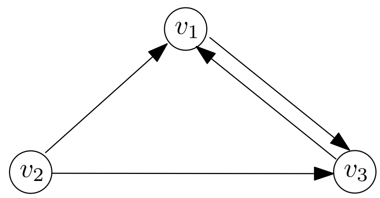
Problem 3. Convert D_1 into an 8-dimensional point according to the tf-idf model.

Solution. $(0.415, 1, 2, 0, 0, 0, 0, 0)$.

Problem 4. Assume that a query (which is a sequence of words) has been converted to a point $(0.415, 1, 2, 0, 0, 0, 0, 0)$. What is the score of D_1 with respect to this query according to the cosine metric?

Solution. 1.

Problem 5. Consider the following graph:



Let v_1 be the first vertex in Google's random surfing model. What is the probability that v_2 is the 10-th vertex visited? Recall that at each step re-seeding happens with probability 15%.

Solution. Observe that v_2 has no incoming edge. Therefore, at any step, the surfer can reach v_2 only through re-seeding. The probability is therefore $15\%/3 = 5\%$.

This problem is canceled because the edge from v_3 to v_1 was missing in the quiz paper.