CAB431 Tutorial (Week 5): IR models
Solution

****

Task 1. Solution (discussion points):

There are lots of variant formulations and combinations! Whatever formulation is used, the unit-length-normalized TF*IDF scores are the precomputed and stored, so that similarity comparison is just a dot product.

Term Frequency (*tf*):

The term frequency *tf* in *tf*idf* can be the raw term frequency $f_{d,t}$ (the number of term *t*'s appearance in document *d*). However, a term that occurs 10 times is not generally 10 times as important as a term that occurs once. Therefore, an alternative formulation of the *tf* in a document *d* can be:

$$1 + \log(f_{t,d})$$

Inverse Document Frequency (*idf*):

If *N* is the number of documents in a given document collection *C* (or a dataset), and df_t is the number of documents that contain term *t*. Then the *idf* of term *t* in a collection *C* is defined as:

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t}$$

For example, suppose *C* includes 10 documents, and a word "tutorial" appears in three documents. Then, mathematically, its Inverse-Document Frequency, idf_t= log(10/3).

Smoothing and Document-Length-Normalized version:

$$tfidf(t, d) = \frac{(1+\log(f_{t,d}))\cdot\log\frac{N}{df_t}}{\sqrt{\sum_{i=1}^{T}\left[(1+\log(f_{i,d}))\cdot\log\frac{N}{df_i}\right]^2}}$$

where *N* = |*C*|, and *T* is the total number of terms in collection *C*.

Task 2. Solution

	Term1	Term2	Term3	Term4	Term5
D1	3	0	0	5	7
D2	5	3	4	6	0
D3	0	0	5	4	6
D4	9	0	0	1	2
D5	0	1	0	3	2
D6	3	0	2	4	4
df	**4**	**2**	**3**	**6**	**5**

Task 4. Solution

	Relevant	**Non-relevant**	**Total**
$d_1 = 1$	$r_i = 3$	$n_i - r_i = 1$	$n_i = 4$
$d_1 = 0$	$R - r_i = 0$	$(N-R)-(n_i - r_i) = N - n_i - R + r_i = 3$	$N - n_i = 3$
Total	$R = 3$	$N-R = 4$	$N = 7$

	Relevant	**Non-relevant**	**Total**
$d_2 = 1$	$r_i = 0$	$n_i - r_i = 4$	$n_i = 4$
$d_2 = 0$	$R - r_i = 3$	$N - n_i - R + r_i = 0$	$N - n_i = 3$
Total	$R = 3$	$N-R = 4$	$N = 7$

	Relevant	**Non-relevant**	**Total**
$d_3 = 1$	$r_i = 2$	$n_i - r_i = 1$	$n_i = 3$
$d_3 = 0$	$R - r_i = 1$	$N - n_i - R + r_i = 3$	$N - n_i = 4$
Total	$R = 3$	$N-R = 4$	$N = 7$