School of Computing and Information Systems
The University of Melbourne
COMP90049
Knowledge Technologies (Semester 1, 2018)
Workshop exercises: Week 6

1. Identify the two (sometimes three) components of a **TF-IDF model**. Indicate the rationale behind them as in, why would they contribute to a "better" result set?

2. Many TF-IDF models are possible; consider the following one:

$$w_{d,t} = \begin{cases} 1 + \log_2 f_{d,t} & \text{if } f_{d,t} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$w_{q,t} = \begin{cases} \log(1 + \frac{N}{f_t}) & \text{if } f_{q,t} > 0 \\ 0 & \text{otherwise} \end{cases}$$

   (a) Construct suitable vectors for the five documents in the collection below, and then use the **cosine measure** to determine the document ranking for the (conjunctive) query `apple lemon`:

DocID	`apple`	`ibm`	`lemon`	`sun`
Doc_1	4	0	0	1
Doc_2	5	0	5	0
Doc_3	2	5	0	0
Doc_4	1	2	1	7
Doc_5	1	1	3	0

   (b) If Documents 4 and 5 were the only **truly** relevant documents in the collection, calculate $P@1$, $P@3$, and $P@5$ for the above system.

   (c) (Extension) Do you expect the document ranking to be different if we had instead used the TF-IDF model below? Why or why not?

$$w_{d,t} = \frac{f_{d,t}}{f_t}$$

$$w_{q,t} = \begin{cases} 1 & \text{if } f_{q,t} > 0 \\ 0 & \text{otherwise} \end{cases}$$

3. What does **Recall** correspond to, in an Information Retrieval context? Why is Recall not usually considered to be a useful evaluation metric for an IR Engine? How does this relate to $P@k$?