School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 2, 2019)
Workshop exercises: Week 5

1. What is the difference between "data retrieval" and "information retrieval"? Why is the latter a **knowledge task**, but the former is not?

2. In the context of an **information retrieval** engine, what does it mean for a document to be **returned** for a query? What does it mean for a document to be **relevant** for a query?

3. Identify some differences between **Boolean** querying and **Ranked** querying.

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

5. 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
Doc₁	4	0	0	1
Doc₂	5	0	5	0
Doc₃	2	5	0	0
Doc₄	1	2	1	7
Doc₅	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.

6. 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?