School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 1, 2017)
Workshop exercises: Week 6

1. Given a document set made up of five documents, with the indicated term frequencies $f_{d,t}$:

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

   calculate the document ranking for the (conjunctive) query `apple lemon`, based on the **language model** approach to IR given in the lectures, using $\mu = 1$, and then $\mu = 3$:

$$S(q,d) \quad = \quad \prod_{t \in q} \left(\frac{|d|}{|d|+\mu} \cdot \frac{f_{d,t}}{|d|} + \frac{\mu}{|d|+\mu} \cdot \frac{F_t}{F} \right)$$

2. We ran four different systems, which each returned documents for a single query. We then judged whether each result returned was relevant (1) or not relevant (0):

Rank	1	2	3	4	5	6	7	8	9	10	11	12
System A	0	1	1	1	0	0	0	1	0	1	1	1
System B	1	0	1	0	1	0	1	0	1	0	0	0
System C	0	1	0	1	0	1	0	1	0	1	0	1
System D	1	1	1	1	1	0	0	0	0	0	0	0

   (a) Find the precision of each of the four systems. What about recall?

   (b) Rank the systems according to P@1, P@3, P@6, P@12.

   (c) Make a reasonable assumption, and then find the AP score for each system.

3. What is **pooling**, and why is it important in IR evaluation?

4. What are the four primary components of a **Web–scale Information Retrieval engine**? Briefly describe our goal in each of them.

5. Recall the (hypothetical) method of **crawling** given in the lectures:

   (a) Would this method be *effective* at solving the problem of crawling? Why or why not?

   (b) Would this method be *efficient* at solving the problem of crawling? Why or why not?

6. "Canonicalisation" (of text) typically comprises "tokenisation" and "normalisation." What are these generally accepted as referring to?
   (Note the terminology is not used consistently in the literature; for example "tokenisation" occasionally refers to all three ideas.)

   (a) What are some issues that arise when canonicalising text written in English?

   (b) (EXTENSION) What are some issues that might arise when canonicaling text written in other languages?