School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 1, 2019)
Workshop exercises: Week 7

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

2. 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?

3. When **tokenising** text, we often **canonicalise** it. What are these generally accepted as referring to?
   (Note the terminology is not used consistently in the literature.)

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

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

4. Assume that we have crawled the following "documents":

```
(1) The South Australian Tourism Commission has defended a marketing strategy
which pays celebrities to promote Kangaroo Island tourism to their followers
on Twitter.
(2) Mr O'Loughlin welcomed the attention the use of Twitter had now attracted.
(3) Some of the tweeting refers to a current television advertisement promoting
Kangaroo Island.
(4) Those used by the Commission have included chef Matt Moran, TV performer
Sophie Falkiner and singer Shannon Noll.
(5) He said there was nothing secretive about the payments to celebrities to
tweet the virtues of a tourism destination.
(6) Marketing director of SA Tourism, David O'Loughlin, said there was no ethical
problem with using such marketing and it might continue to be used.
(7) Depending on their following, celebrities can be paid up to $750 for one
tweet about the island.
```

- Parse each document into terms.
- Construct an inverted index over the documents, for (at least) the terms `and`, `australia`, `celebrity`, `commission`, `island`, `on`, `the`, `to`, `tweet`, `twitter`
- Using the vector space model and the cosine measure, rank the documents for the query `commission to island on twitter`

  (a) Using the weighting functions $w_{d,t} = f_{d,t}$ and $w_{q,t} = \frac{N}{f_t}$

  (b) Using the weighting functions $w_{d,t} = 1 + \log_2 f_{d,t}$ and $w_{q,t} = \log_2(1 + \frac{N}{f_t})$

5. When querying, what is an **accumulator**? What is the main problem, if we wish to use them? What heuristics can we use to solve this problem?