School of Computing and Information Systems
The University of Melbourne
COMP90049 Knowledge Technologies (Semester 2, 2018)
Workshop exercises: Week 6

1. What are the four tasks into which a Web-scale information retrieval engine is usually divided? Briefly summarise each one.

2. When parsing Web pages:

   (a) What is "tokenisation"? What are some common problems that arise in tokenisation of English text? What about other languages?

   (b) What is "stemming"? How might it be different in languages other than English?

3. 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 function $w_{d,t} = f_{d,t} \times \frac{N}{f_t}$

     (b) Using the weighting functions $w_{d,t} = 1 + \log_2 f_{d,t}$ and $w_{q,t} = \log_2(1 + \frac{N}{f_t})$
   - Suppose there is an accumulator limit of 2 and that query terms are processed in order of decreasing rarity — are the same top two documents found? What are the similarities of the top two documents?

4. What is "phrase querying"? What extra information would you need to store in your inverted index to accomplish phrase querying? How much extra space would that require? Indicate how you would use that information to execute a phrasal query like `"to the commission"` on the data set above.

5. When evaluating a query, why should we begin with the rarest terms (those with greatest $w_{q,t}$)? Is this more true for ranked or Boolean querying?

6. What is "link analysis"? Briefly describe the PageRank algorithm and the rationale behind it.