School of Computing and Information Systems
The University of Melbourne
COMP90049
Knowledge Technologies (Semester 1, 2018)
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. **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?