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

   - The main difference here is the existence of people — users. Because people are wildly divergent, the notion of a relevant result in information retrieval depends on contextualising the data to the particular user (which may be very difficult, because we have an imperfect model of the user and indeed the user has an imperfect model of their needs!). Whereas with data retrieval, there is a particular unit of data (bitstream) that we need to access in memory or on a hard drive, and there is generally no ambiguity.

2. (Extension) How many books are there in an average library? How many words are there in an average library? How many documents are there on the World Wide Web? How many words?

   - These aren't straightforward questions to answer. But, to an order of magnitude, a small city library might have about 10K books; a larger one, maybe 30K. I might estimate the word count of a typical book to be about 50K (many are longer; many are shorter; there are varying definitions of "word"), which would situate a library as carrying roughly 1G words. The US Library of Congress catalogues about 2.3M books, so perhaps 100G words.
   - As of 2008, Google claimed to index 1T unique urls (`http://googleblog.blogspot.com.au/2008/07/we-knew-web-was-big.html`); by 2012, this had supposedly risen to 30T (`http://www.google.com/insidesearch/howsearchworks/thestory/`); now, it would hypothetically be much larger. But maybe take all of this with a grain of salt! `:-)` Estimating the number of words on the Web is even harder — Google tells me that the mean document size is about 400KB, but much of that isn't going to be text. I might ballpark about 1000 words (roughly 6KB of the 400KB) per document, which might be upwards of 10000T words (or more!, but probably less)!

3. Identify some different types of "informational needs."

   - Rehashing the lecture slides:
     - Requests for informations, e.g. "global warming"
     - Factoid questions, e.g. "melting point of lead"
     - Topic tracking, e.g. "Dutch elections" [as in, the most recent ones]
     - Navigational, e.g. "University of Melbourne home page"
     - Service or transactional, e.g. "Mac powerbook"
     - Geospatial, e.g. "Carlton restaurant"
   - This isn't an exhaustive list. Nor is it non-overlapping: for example, most queries can be construed as being navigational in nature (as the user is likely to click through to a relevant document), and many are informational as well.

4. 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?

   - An information retrieval engine returns documents with respect to a query, usually based on the presence of the keywords (from the query) in the document.
   - A document is relevant for a query if it meets the user's information need (approximated by the query) — note that there is no requirement for the query keywords to be present!

5. Identify some differences between **Boolean** querying and **ranked** querying, in an Information Retrieval context.

   - Boolean: documents match if they contain the terms (and don't contain the `NOT` terms; i.e. the Boolean formula evaluates to `TRUE`); matching is Yes/No; repeatable, auditable, controllable; queries allow expression of complex concepts
   - Ranked: based on evidence that the document is on the same topic as the query; matching is gradiated (to come up with a ranking!); different models give different results; queries are easy to write and results are easy to read for non-specialists

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

   - More weight is given to documents where the query terms appear many times (TF)
   - Less weight is given to terms that appear in many documents (IDF)
   - Less weight is given to documents that have many terms (not present in all models)