School of Computing and Information Systems
The University of Melbourne
COMP90042
WEB SEARCH AND TEXT ANALYSIS (Semester 1, 2019)

Workshop exercises: Week 4

Discussion

1. Discuss the process of static **inverted index construction** and how it can be used to perform in incremental index construction.
2. Why is a **logarithmic** index layout useful? What are the disadvantages of such an index structure?
3. Based on the following top-6 retrieval results from a collection of 100 documents, and the accompanying binary relevance judgements

doc	score	relevance
a	0.4	0
b	1.2	0
c	2.2	1
d	0.5	1
e	0.1	1
f	0.8	0

   compute the following evaluation metrics:

   (a) precision@3
   (b) average precision (do you need to make any assumptions about the document collection?); and
   (c) rank-biased precision (RBP), with $p = 0.5$
   (d) plot the precision-recall graph, where you plot (precision, recall) point for the top k documents, $k = 1, 2, \ldots 6$.
   (e) what are the strengths and weaknesses of the methods above for evaluating IR systems?

4. How can a retrieval method be learned using supervised machine learning methods? Consider how to frame the learning problem, what data will be required for supervision, and what features are likely to be useful.

Catch-up

- What is top-K retrieval and why is it useful?
- What are some of the reasons a search engine would want to offer a query completion service?

Get ahead

- The lecture links to a blog post discussing a more advanced index merging scheme used by Lucene. Read up on why Lucene uses this method compared to the logarithmic method discussed in the lecture.
- In general, similarity measures that we have discussed up till now are quite simplistic as they only take into consideration term frequency statistics. Modern search engines use many more advanced features to determine the relevance of a document. Can you think of some features that might be useful?