National University of Singapore
School of Computing
CS3245: Information Retrieval
Tutorial 6

Probabilistic IR and Web Search

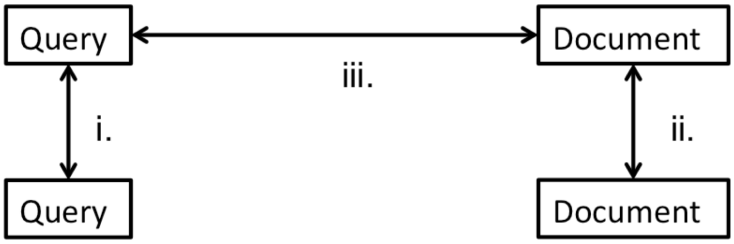
Readings: IIR Chapters 11, 12, 20 & 21

1. **Probabilistic and Language Model IR.** The language model approach to IR states that we estimate the relevance of a document to a query, $P(d|q)$, as the probability that a random sample of the document would generate the query $P(q|d)$.

   (a) We computed in lecture how likely a document generates a query using a language model. However, the language model for documents is in fact not a very good model for queries. Do you know why?

   (b) Would it be better to use a language model for a query to generate a document? Justify your answer.

   (c) The below diagram illustrates three different sources of information: i. query-to-query similarity, ii. document-to-document similarity, and iii. query-to-document similarity. In class, we have been limiting our discussion to the final, third relationship. Describe briefly how the other two relationships (i and ii) could be used to improve IR performance, when you have access to the appropriate sources (e.g., a large collection of queries for i. and a large collection of documents for ii.).



(The following questions are for self-study after the final lecture.)

2. **Spamdexing.**

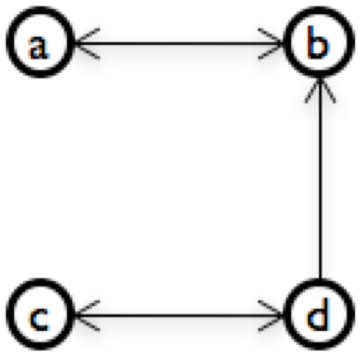
   One form of spamdexing is to manipulate documents in order to artificially raise or lower the relevance of a document to query. For example a malicious user may use comment spamdexing to add comments to a news article to promote a commercial website or lower the article contents' relevance for important keywords.

   Defend or refute the following statement and justify your argument. *Standard IR models that treat documents as simple bags of words without ordering or positional information are more prone to spamdexing than IR models that can differentiate words at different positions.*

3. **PageRank.**

   (This was a previous final exam question in AY 2015/2016.)

(a) Calculate PageRank on the following graph, where the damping factor is set to 0.85 (15% teleportation rate). Assume a uniform initial vector. Show all work for three iterations of PageRank, where results are calculated to two decimal places.



(b) Which edge can be added to the graph in a) to make the rankings of the graph most uniform? Justify either numerically or in prose reasoning.

(c) Which edge can be added to the graph in a) to make the rankings of the graph favor the dominant node even more? Justify either numerically or in prose reasoning.

(d) Say we have prior evidence that a certain sets of nodes in the graph are more important. Thus far, PageRank considers all nodes with equal (uniform) weight. Consider a modification to add node weights, where we can assign higher initial weight to important nodes at the expense of the lower weights of unimportant nodes. Describe the influence this scheme would have on the final node ranking.