

## Review Questions

1. Illustrate the operation of a search engine like Google (as described in Brin and Page 1998) by providing a high-level trace of the processing it would do for the query “mobile telephone ergonomics”. Note places where parallelism is exploited.
2. Give pseudocode for finding the 10 most similar documents to a query using the vector space model. Briefly describe the data structure(s) your code assumes.
3. There are many 2-word pairs in English, including some, like “ice cream” and “platform independent” which should be treated as single terms. If there is a 2-word sequence  $x y$ ,  $P(x) = 10^{-4}$ ,  $P(y) = 10^{-3}$  and  $P(xy) = 10^{-5}$ , should  $x y$  be treated as a single term? Explain.
4. Explain how good personalization can compensate for weaknesses in the computation of query-document similarity. Also explain how it can compensate for weaknesses in the page quality estimate (PageRank etc.)
5. Query expansion is usually done without notifying the user what’s going on. Give reasons both for and against making this more visible to the user.
6. Explain how a language model contributes to speech recognition.
7. Name three important ways in which commercial search engines differ from freeware IR systems.
8. Discuss the advantages and disadvantages of structured queries as a way to access a database of scientific articles.
9. The Library of Congress devotes a lot of effort to cataloging books. Can they stop now that we have search engines? Why or why not?

