Problem 1

A. Explain why it is important for a crawler to detect whether two pages that it has downloaded are “near duplicates”.

B. Give two reasons that a crawler would want to record the URLs of all the near duplicate pages it has downloaded, rather than discard them.

C. How does the existence of near duplicate pages affect the computation of PageRank?

Problem 2

The use of inverse document frequency (IDF) in the vector model of document retrieval can lead to the following anomaly: There can exist documents D and E and collections A and B, where both A and B contain both D and E, and a query Q such that:

- If Q is posed in the context of collection A, D is judged to be more relevant to Q than E.
- If Q is posed in the context of collection B, E is judged to be more relevant to Q than D.

A. Explain how this can happen.

B. Can this happen if only term frequency is used? Explain your answer.

C. Argue using an example that in some cases this may actually be reasonable.

Problem 3

Suppose that P, Q, and R are different web pages. Explain how it can happen that adding a link from P to Q can raise the PageRank of R. Explain how it can happen that adding a link from P to Q can lower the PageRank of R. In both cases, you should show a specific graph where this happens, though you need not work out the actual numerical values.

Problem 4

Describe two measures for evaluating the quality of a ranked list of results. Explain how these relate to user models.

Problem 5

Clustering of web pages can be done either based on the entire text of the web pages or based on snippets; and it can be done online or offline.

A. It sounds like there should be four combinations, but in fact there are only three. Which combination is impossible? Why?

B. Describe the advantages and disadvantages of the remaining three possibilities.
Problem 6

A. Suppose for query Q, sponsor A has bid $1.00 per click, sponsor B had bid $0.50 per click, sponsor C has bid $0.20 per click, and that the search engine has a minimal charge of $0.01 per click. They all specify a total budget of $50 per day. Assume that the search engine company uses second-price auction. On one day, there are 200 clicks on the top sponsored link on the results page, and no clicks on any of the lower links. What does each of the companies owe the search engine company?

B. Google does not in fact rank bids purely on their value; it also incorporates a quality score. Using the above example, explain how this can cost Google money in the short term. What is the long term advantage to Google?