Problem 1

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

Answer: Near duplicate pages should not be returned as different results on a results page.

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.

Answer:

- The URLs need to be stored in the “seen” cache so that they are not downloaded again in this crawl.
- The URLs are used as part of the seed collection for the next crawl. (By the time of the next crawl, the pages may no longer be near duplicates.)
- The multiple URLs need to be taken into account in terms of PageRank as in part (C).

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

Answer: For the purposes of PageRank, all duplicates and near duplicates should be considered as a single node. E.g. if there are two copies A.1 and A.2, then these should be viewed as a single conceptual node A, where the links into A are the union of the links into A.1 and A.2.

Problem 2

The use of inverse document frequency (IDF) in the vector model of document retrieval can lead to the following anomaly: There can exists 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.

Answer: Suppose Q consists of two query terms, X and Y; that X is frequent in D but rare in E and Y is frequent in E but rare in D; and that X is rare in documents in A but common in B, and Y is common in A but rare in B. Then the IDF for X in A is large and the IDF for Y in A is small, so X will be weighted more than Y for queries over A, so D will be ranked higher than E for queries over A. The reverse is true for queries over B.

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

Answer: No. If only term frequency is used, then the collection doesn’t enter into the calculation at all.

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

Answer: This question is actually hopelessly difficult. An answer would have to satisfy three constraints:
• The informational needs of a typical searcher through A are different from the informational needs of a typical searcher through B.

• A and B must overlap in two documents D and E.

• The two users doing the searches on the query "X Y" in the two different collections are less interested in the word that is common in the collection and more interested in the word that is uncommon.

With enough work one could probably contrive a reasonably plausible scenario, but it is certainly not a reasonable exam question.

What is true is that this consideration can work well in clustering. If you are clustering the documents in a topic specific collection A, then there is no much point in giving high weight to documents characteristic of the topic of A, because they don’t give any useful discrimination. E.g. if you are clustering documents in the Journal of Infectious Diseases, you do not want to give high weight to "virus" because the relative frequency of the word in given document probably does not indicate much about the subtopic of the document.

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.

Answer:
First case: Initial graph:

```
P
Q --> R
```

Adding a link from P to Q raises the PageRank of Q and thus indirectly the PageRank of R.

First case: Initial graph:

```
P --> R
Q
```

If you add a link from P to Q, then P’s “contribution of importance” is divided between Q and R rather than going exclusively to R, so the PageRank of R decreases.

Problem 4

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

- Precision at 10: The fraction of results on the first results page that are relevant to the query. This corresponds to a model in which the user looks at all the relevant results on the first page; it is (very roughly) the total amount of information obtained from that.
- Rank of first relevant item (small is good). This corresponds to a model in which the user goes down the list of results until reaching one that is relevant. It is (roughly) the time spent skipping irrelevant results.

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?

Answer: Offline based on snippets is impossible, since the selection of snippets is query-specific.

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

Answer:

- Entire text offline has the maximal amount of information and can afford comparatively computationally expensive techniques, but creates a fixed clustering system with no query specificity.
- Entire text online has maximal amount of information and is query-specific but is likely to be slow, because all pages must be downloaded.
- Snippet online is query-specific and fast, but may miss important clues omitted from the snippet.

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 clocks on any of the lower links. What does each of the companies owe the search engine company?

Answer: First A gets 98 clicks, paying $0.51 per click (second price). At that point A’s budget is exhausted, so A is no longer bidding. A owes the search engine $49.98. The next 102 clicks go to B, which pays $0.21 per click, so B owes the search engine $21.42. (Thanks to Michal Novemsky for correcting my initial wrong answer.)

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?

Answer: Suppose that A and B’s web pages are so lousy that C’s quality score outweighs its low bid and C is ranked first. Then C pays only $0.01 per click (search engine minimum) so it owns Google only $2.00 for the 200 clicks, so Google has lost $68. In the long run, Google figures that keeping end users happy by maintaining high quality on the highly-ranked sponsored links is worth the short-term loss.