1. [4 pts] The terms “TF” and “IDF” are used in information retrieval. What do the terms stand for?

2. [4 pts] Recall and Precision are two measures of the effectiveness of an Information Retrieval system. If A is the number of relevant records retrieved, B is the number of relevant records NOT retrieved, and C is the number of irrelevant records retrieved, define Recall and Precision in terms of A, B, and C.

3. [4 pts] Google offers a variety of special operators that can be used to narrow a search. Define the following ones: filetype, site, allinanchor.

4. [4 pts] When an advertiser decides to bid on a set of keywords, e.g. “European cars”, Google, Bing and Yahoo allow the advertisers to match keywords in several ways. Mention and describe two, each in a single sentence.

5. [4 pts] What is a “parked domain”?

6. [4 pts] Write out the 3-grams for the phrase: “Fourscore and seven years ago our fathers brought forth a nation”

7. [4 pts] State Zipf’s Law

8. [4 pts] Suppose there are only two web pages, each with only one link that points to the other web page. What will be the PageRank of each page?
9. [4 pts] As a website grows and adds more pages with more links to web pages outside of the website, how is the total PageRank of the website affected?

10. [4 pts] The HITS Algorithm developed by Jon Kleinberg identifies two types of web pages that have special significance. What are these two types of web pages?

11. [4 pts] True or False, Google, Yahoo, and Bing record all user clicks, both on ads and on organic search results.

12. [4 pts] When investigating click fraud, there are both online tests and offline tests. Give an example of an online test and an offline test.

13. [4 pts] Suppose one advertiser bids $1.00 for his ad to be displayed and a second advertiser bids $0.50 for his ad to be displayed and all other factors affecting ads are identical. If the first advertiser’s ad is clicked on how much does he pay Google?

14. [4 pts] What is a “tracking pixel”?

15. [4 pts] Define “cloaking”

16. [4 pts] What effect does the following line in a web page have?

    `<meta name=robots content="noindex,nofollow">`

17. [4 pts] When creating an index of documents search engines make use of case folding, stemming and stop words. In one sentence define each of these three terms.

18. [4 pts] Kleinberg’s “Authoritative Sources in a Hyperlinked Environment” was covered in the Essential Papers section of the course and in lecture material. Consider Twitter as a set of hyperlinked resources – would you consider Twitter an authority or a hub in today’s world? Explain your answer.
19. [4 pts] In an inverted index why is it important to use numeric identifiers as opposed to URLs? Explain your answer (5 pts).

20. [4 pts] Name three reasons that it is important to detect mirrors during deduplication (10 pts).

21. [4 pts] You are building a new inverted index system for your web search engine company Acme La Vista that is a focused vertical search engine for Italian Food restaurants. Your inverted index includes 3 documents ($D1$…$D3$) fetched by your crawler, which are decomposed into the following words:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$D1$</td>
<td>Italian luigis Fetuccini pasta Heineken</td>
</tr>
<tr>
<td>$D2$</td>
<td>Dona Marie Italian sausage tortellini</td>
</tr>
<tr>
<td>$D3$</td>
<td>Fetuccini butter salt oregano garlic</td>
</tr>
</tbody>
</table>

22. [4 pts] Derive the term dictionary (“index file”) and the postings file for the above 3 documents.

23. [4 pts] How many iterations does Page Rank require to compute the ranking of all documents in an index? Explain your answer.

24. [4 pts] How many characters does BING wait before auto suggesting a keyword to search on?

25. [4 pts] A study of how to design a web page crawler to locate the best quality pages was done by Cho and Garcia-Molina. What measure of quality did they use? What algorithm did they determine would produce the highest quality pages in the shortest time?