Computer Science 572 Midterm
Prof. Horowitz
Thursday, March 8, 2012, 2:00pm – 3:00pm

Name: **Student Id Number:**

1. This is a closed book exam.
2. Please answer all questions.
3. There are a total of 40 questions. Each question is worth 2 1/2 points.
4. **Place your answer immediately below the question.**

1. [2 1/2 pts] Name the 3 major phases that a search engine goes through:

2. [2 1/2 pts] When a search engine crawls the web and visits a website for the first time, what is the first file the crawler should look for?

3. [2 1/2 pts] Where is the file in your answer above located?

4. [2 1/2 pts] The terms "TF" and "IDF" are used in information retrieval. What do the terms stand for?

5. [2 1/2 pts] Given sets S and T, define the Jaccqard Similarity of S and T.

6. [2 1/2 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.

7. [2 1/2 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?

8. [2 1/2 pts] Google offers a variety of special operators that can be used to narrow a search. Define the following ones: filetype, site, allinanchor.

9. [2 1/2 pts] True or false: Google does auto completion and spelling correction at the same time.

10. [2 1/2 pts] What is the purpose of the Levenshtein Algorithm?

11. [2 1/2 pts] Define "case folding".

12. [2 1/2 pts] Define "stemming".

13. [2 1/2 pts] Define "stop words" and provide three examples

14. [2 1/2 pts] Define the term-document incidence matrix

15. [2 1/2 pts] Inverted indices typically include for each term a list of documents where the term occurs. Some inverted indices also include the position of a word in a document which adds a lot of additional space. What advantage is gained by including the position of a word?

16. [2 1/2 pts] Write out the 3-grams for the phrase:
"Fourscore and seven years ago our fathers brought forth a nation"

17. [2 1/2 pts] State Zipf's Law

18. [2 1/2 pts] What is the Soundex Algorithm?

19. [2 1/2 pts] What is Porter's Algorithm?

20. [2 1/2 pts] What is de-duplication and give two examples of why it needs to be done.

21. [2 1/2 pts] What might a search engine do to quickly check that a newly discovered web page is identical to one that was already seen and indexed?

22. [2 1/2 pts] Cho and Garcia-Molina describe strategies for distributed crawlers and they define three different ways the crawlers can interact: independent, dynamic assignment, or static assignment. In a few words define each.

23. [2 1/2 pts] What are the names of the Google and Yahoo! web crawlers?

24. [2 1/2 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?

25. [2 1/2 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?

26. [2 1/2 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?

27. [2 1/2 pts] How is the failure of a Map worker handled in the MapReduce framework?

28. [2 1/2 pts] Google has two programs for advertisers, one that places ads next to search engine results and one that places ads on a website. Name both of the programs.

29. [2 1/2 pts] When Google must decide how to order the ads for a given query phrase, what formula does it use?

30. [2 1/2 pts] Two improper techniques used to enhance a web pages ranking in search results are cloaking and page jacking. Define them both.

31. [2 1/2 pts] When an advertiser bids on a set of keywords he can ask for 3 different types of matches. Mention two of them and define them.

32. [2 1/2 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?

33. [2 1/2 pts] What is a “tracking pixel”?

34. [2 1/2 pts] Suppose the Pepsi Cola company wants to bid on the words Coca Cola whenever they are entered as a query, so a Pepsi Cola ad will appear. Is this legal?

35. [2 1/2 pts] When a search engine gets a query such as “what are the movie times for The Artist, how are they able to identify the local movie theaters?

36. [2 1/2 pts] What is meant by Google’s Universal Search?

37. [2 1/2 pts] Do Google, Yahoo and Bing track ALL clicks by users, or just clicks on ads?

38. [2 1/2 pts] What is Google’s reason for not telling an advertiser why each and every click was marked as valid?

39. [2 1/2 pts] List the four main features/functions that Apache Tika provides.

40. [2 1/2 pts] What format does Apache Tika use to represent and return the parsed content of a document stream?