1. Can a web page author claim his page is copyrighted if he forgets to insert a “Copyright ©” notice statement on the page?

2. Name the four types of protection for intellectual property.

3. Google was successfully sued by the United States Federal government for offering ads by Canadian pharmacies. What did Google do that was wrong?

4. Define and describe what is a “term-document incidence matrix”.

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. Write your answers in the space provided immediately below the question only.
5. What do the terms AdWords and AdCenter refer to?

6. 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. Name two of these methods, and describe each one in a single short sentence.

7. How is the failure of a Map worker handled in the Map/Reduce framework?

8. What is a “tracking pixel”?
9. State Zipf’s Law

10. When investigating click fraud, there are both online tests and offline tests. Give an example of:
   i) an online test.
   ii) an offline test.

11. What effect does the following line in a web page have?

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

12. 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.
13. The terms “TF” and “IDF” are used in information retrieval. What do the terms stand for?

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

15. Google and Bing both allow advertisers to restrict where their ads will be seen; the restriction can be by country, by state, by city. Name one way to accomplish this.

16. What is Hadoop?

17. With respect to search engines, what does the term “relevance feedback” refer to?

18. What is the Soundex Algorithm?

19. 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?

20. 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?
21. True or False? Google, Yahoo, and Bing record all user clicks, both on ads and on organic search results.

22. When Google must decide how to order the ads for a given query phrase, what formula does it use?

23. 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?

24. 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?

25. What does DMCA stand for?

26. 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?

27. What is a way to guarantee that an advertiser’s ad will appear at the top of a Google or Bing results page?

28. Define “cloaking”.
29. What is de-duplication? Give two examples of why it needs to be done.

30. What is Google’s reason for not telling an advertiser why each and every click was marked as valid?

31. What is a “parked domain”?

32. Write out all of the 3-grams for the following phrase:
   “Fourscore and seven years ago our fathers brought forth a nation”
33. Google offers a variety of special operators that can be used to narrow a search. Define the following:
   i) filetype:
   ii) site:
   iii) allinanchor:

34. Some browsers now include a feature that prevents “third-party cookies” from being placed on a browser. Name the three parties involved.

35. The HITS Algorithm developed by Jon Kleinberg identifies two types of web pages that have special significance. Name and describe these two types of web pages.

36. When creating an index of documents search engines make use of case folding, stemming and stop word removal. Briefly define these three terms in one sentence each.
37. List the four main features/functions that Apache Tika provides.

38. Define Kendall’s Tau distance in words, i.e. without using mathematical symbols.

39. Given two sequences of length $n$, what is their maximum Kendall Tau distance?

40. Define Spearman’s footrule distance for two lists of $n$ items without using mathematical symbols.