Computer Science 572 Exam
Prof. Horowitz
Monday, April 25, 2016, 10:00pm – 11:45am

Name: **Student Id Number:**

1. This is a closed book exam.
2. Please answer all questions.
3. There are a total of 50 questions. Each question is worth 2 points.
4. **Place your answer immediately below or to the right of the question. Limit answers to ONE SENTENCE.**

1. what is the title of the textbook for this class?

2. Name one of the authors of the textbook

3. AltaVista, Lycos and InfoSeek are (circle the best answer):
- apps for an iPhone
- early web search engines
- information retrieval systems

4. David Filo and Jerry Yang are (circle the best answer):
- founders of Google
- creators of spreadsheets
- founders of Yahoo

5. Google first appeared on the web in (circle the best answer):
- 1998
- 2004
- 2010

6. Google retains a user's entire query history? (circle the best answer):
- True
- False

7. What is the second largest web search engine in terms of revenue?

8. Approximately how many websites are there? (circle the best answer):
- 1 million
- 10 million
- 100 million
- 1 billion
- 10 billion
- 100 billion

9. Which content type is NOT indexed by Google? (circle the best answer):
- swf
- xlsx
- rtf
- svg
- None of the above as all of them are indexed

10. The definitions of Precision and Recall both have "the number of relevant items retrieved" as their numerator. What is the denominator for Precision?

11. The definitions of Precision and Recall both have "the number of relevant items retrieved" as their numerator. What is the denominator for Recall?

12. What information is contained in the robots.txt file?

13. In what directory does one find the robots.txt file on a website, if there is one?

14. What is the better strategy for crawling a website? (circle the best answer):
- breadth first search
- depth first search

15. Name a cryptographic hashing method:

16. A cryptographic hash function of file X has three main properties:
it is easy to compute
it is difficult to find a file that has the same hash value,
and what is the third property?

17. Distance measures are defined by 4 properties:
1. no distances are negative
2. $d(x,y) = 0$ *iff* $x = y$
3. $d(x,y) = d(y,x)$
what is the fourth property?

18. Given two sets A and B, define the Jaccard similarity.

19. In one sentence what is the English description of tf-idf?

20. crawler4j was downloaded from what online source repository?

21. what formula or law is expressed this way?
$\log(y) = \log(k*x^c)$

22. Give three examples of stopwords:

23. In one sentence define Heaps Law.

24. Define token

25. Define tokenization

26. Stemming is a method for normalizing tokens. What is the name of a famous stemming algorithm?

27. A phonetic algorithm identifies words that sound the same but are spelled differently. Name one.

28. In class you saw the Block Sort-Based Indexing algorithm. What was the algorithm attempting to minimize?

29. True or False, Google does NOT permit Boolean operators in queries?

30. Given the query
a OR "b c" d
where a, b, c, and d represent keywords being searched for, fully parenthesize the query as Google would do it. Insert any implied Boolean operators.

31. Describe in one sentence the search results that are returned by Google if you enter the query
`filetype:  pdf`

32. In one sentence define cloaking.

33. in the line
<meta name=robots content="noindex,follow,noarchive">
explain the meaning of noindex, follow and noarchive

34. In the Google AdWords system what does CTR stand for?

35. In one sentence what is Google AdSense?

36. Lucene/Solr uses two methods for ranking results. What are they?

37. True or false, Solr includes explicit operators for AND, OR, NOT?

38. True or False, Solr will return results in XML format?

39. Several heuristic techniques were presented for speeding up the computation of search results. Mention two of them:

40. Name the three types of spelling errors:

41. What is a popular data structure for organizing a lexicon that is especially useful to implement autocomplete or prefix matching?

42. An equivalent term for the word shingles is?

43. What is the difference between hard clustering and soft clustering?

44. In one sentence define: Dendrogram

45. In the class we have often mentioned Google, Yahoo, and Bing as the major web search engines. However, others have also been mentioned. Name three:

46. Let True Positive (TP) be defined as an invalid click that is correctly identified as invalid; let True Negative (TN) be a valid click that is correctly identified as valid; let False Positive (FP) be a valid click that is incorrectly identified as invalid; and let False Negative (FN) be an invalid click that is incorrectly identified as valid. Using the above terms define: **accuracy rate** and **error rate**:

47. Given a set of N queries and AVGPrec(N) the average precision of each query, what is the formula for the Mean Average Precision?

48. In the formula for Discounted Cumulative Gain, how are documents appearing lower in a search result list penalized?

49. True or False, Page rank is calculated for each search query and highly relevant pages will always have a higher page rank?

50. We looked at two algorithms for classifying documents into groups. What are they called?