Computer Science 572 Exam
Prof. Horowitz
Tuesday, November 29, 2016, 9:00pm – 10:30am

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 the question. Limit answers to ONE SENTENCE.**

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

2. This semester we have seen at least a dozen short videos by two professors from Stanford. Name one of them

3. Google recently announced that it is indexing how many pages?

4. True or False, Google retains a user's entire query history?

5. 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

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

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

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

9. In one sentence define “cloaking”

10. This question has two parts. 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? Secondly, what algorithm did they determine would produce the highest quality pages in the shortest time?

11. True or False: Cho and Garcia-Molina showed that in order to optimize the freshness of a web-crawl, we should crawl pages as fast as possible.

12. A cryptographic hash function of a file 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?

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

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

15. In one sentence define Heaps Law.

16. State Zipf's Law

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

18. An inverted index is often split into two parts. Name them.

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. 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?

22. The HITS algorithm forms what type of graph when ranking pages?

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

24. In one sentence what is Google AdSense?

25. What is a “tracking pixel”?

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

27. 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?

28. Briefly describe the difference between a broad match and an exact match in the context of AdWords.

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

30. Lucene builds an inverted index from documents it parses. Is the inverted index positional?

31. What is the Soundex Algorithm?

32. Google places an implicit Boolean operation between the terms of a query. What is it?

33. Define “edit distance”.

34. The LevenShtein algorithm assigns what weights to the operations of insertion, deletion, and substitution?

35. What data structure is very helpful when used to catch spelling errors as the user types?

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

37. Given two vectors A = (A_1, . . . , A_N) and B = (B_1, . . . ,B_N) representing two documents, define their cosine similarity

38. The k-means++ algorithm uses a different method than the k-means algorithm for choosing the initial clusters. What is that method?

39. Mention one possible criteria for determining when the k-means algorithm can terminate.

40. Is the Aglomerative clustering algorithm top-down or bottom up?

41. What set of points does K-means clustering use to identify a cluster?

42. WordNet uses the following terms: synset, hypernym, hyponym and meronym. In one sentence define one of them.

43. In relation to capturing clicks on the search result pages, Google and Bing differ in what way?

44. Name the four types of protection for intellectual property

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

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

47. In one sentence define the contiguity hypothesis

48. True or False, in vector space classification, it doesn’t matter if one uses Euclidean distance or cosine similarity?

49. Which vector space classification method uses centroids to define the boundaries of regions?

50. Of the two vector classification algorithms discussed in class, which one decomposes the set of documents into Voroni cells?