Computer Science 572 Exam
Prof. Horowitz
Wednesday, October 3, 2018, 8:00am – 8:50am

Name: **Student Id Number:**

1. This is a closed book exam.
2. Please answer all questions.
3. There are a total of 32 questions. Each question is worth 3 points.
4. **Place your answer immediately below the question. Limit answers to ONE SENTENCE.**

1. Name three document types indexed by Google whose format is binary. Name the program suffix that represents the document type.

2. The purpose of the Soundex Algorithm is to take multiple spellings of the same name and map them to a unique id. What is the format of the Soundex Algorithm id?

3. According to the class notes and the textbook, when parsing documents, is a token different than a word? Yes or No?

4. Name two different text encoding standards.

5. What is the relationship between frequency and rank of text in a document?

6. A positional index includes the actual starting position of the words in a document; however, this greatly increases the size of the postings lists. What is a positional index especially good for?

7. When entering a three word query in the Google search box, represented by the three letters: a b c, does Google retrieve documents that contain a OR b OR c, or does it retrieve only documents that contain a AND b AND c?

8. Given two vectors $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$ define the cosine of the angle between the two vectors.

9. When computing the ranked retrieval of a set of documents it is not necessary to return them all, only a subset of the ones with the highest rank. What data structure can be used to speed up the determination of the documents with highest rank?

10. Define the F-measure, *F*, as a function of *R* (recall) and *P* (precision)

11. Given a query that returns four relevant documents X1,X2,X3,X4 and six irrelevant documents Y1,Y2,Y3,Y4,Y5,Y6, in the following order:

    X1,Y1,X2,X3,Y2,Y3,Y4,X4,Y5,Y6

    compute the recall and precision at each of the ten positions of the query results (two decimal places are sufficient)

Result	X1	Y1	X2	X3	Y2	Y3	Y4	X4	Y5	Y6
Recall										
Precision										

12. Porter's stemming algorithm suffers from over-stemming and under-stemming. Define each of them using a single sentence.

13. Crawler4j allows for multiple threads to be used as crawlers. The constant `numberOfCrawlers` specifies the maximum number of crawler threads the program will allow. In what class is `numberOfCrawlers` assigned a value?

14. As part of the exercise you had to make sure that crawler4j would actually visit binary encoded files. To accomplish that you had to assign `setIncludebinaryContentInCrawling();` what is the entire line of code that successfully causes binary files to be crawled?

15. Crawler4j uses the Pattern.compile( ); routine to accept a regular expression that defines what files should be avoided by the crawler. Write a regular expression that can be used by Pattern.compile() and that specifies that the crawler should skip css and js files. Make sure your pattern matches the end of a URL's input string.

16. Crawler4j recognizes many HTTP status codes. Define the following status codes: 201, 301, 401 and 501.

17. The class notes define the distance of two sets, *d(A, B)* and the similarity of two sets *s(A, B)*. How are *d(A, B)* and *s(A, B)* related? A formula relating the two is sufficient.

18. What can Google do to force a webpage to appear lower down in its search results?

19. Given a set of web pages and a very important page, say X in that set, how should someone structure (link together) the set of web pages so page X gets the highest PageRank?

20. Name three types of information YouTube asks the uploader to provide when a video is uploaded?

21. YouTube makes use of Google's worldwide data centers to deliver its content. Websites that lack a global infrastructure can purchase access to a commercial Content Distribution Network. Name one of these companies.

22. In analyzing the effectiveness of autocompletion algorithms the authors used a general statistical technique that is used to evaluate any process that produces a set of responses to a sample of queries. What is the name of the statistical measure?

23. Does Google support stemming?

24. Does Google support the use of wildcards, e.g. * in a query?

25. Name three Google operators that can be used in a query

26. Provide one sentence that describes how the following query behaves:
    allintext:orbi eero google wifi

27. Google follows a 4-step process in changing their search algorithm. What are those 4 steps?

28. Google has several "special" websites that offer searches to specific types of content. Name two of them:

29. The video by Prof. Ullman is entitled *locality sensitive hashing*? In one sentence describe what the algorithm is designed to do.

30. In the class notes at least two techniques were mention for avoiding spider traps. Mention any one of them?

31. Below is the formula for PageRank

    $PR(A) = (1-d) + d\ (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$

    where A, T1, . . . , Tn are web pages. Define the following three entities: *PR(A)*, *C(T1)*, and *d*

32. All web pages are given an initial PageRank value. One approach assigns zero initially to all pages. Another approach assigns a small non-zero value like 0.15 to all pages. Does the eventual PageRank depend upon the initial value?