Computer Science 572 Exam
Prof. Horowitz
Wednesday, October 4, 2017, 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 25 questions. Each question is worth 4 points.
4. **Place your answer immediately below the question. Limit answers to ONE SENTENCE.**

1. Computer experts tell us we are now in the Mobile/Internet Computing Technology cycle. The class notes mention four previous technology cycles. Name two of them.

2. The number of websites, to an order of magnitude is: 10 million, 100 million, 1 billion, 10 billion or 1 trillion?

3. In two sentences define the deep web and the dark web?

4. True or false, by default Google maintains a user's entire query history?

5. Define the algorithm for computing the harmonic mean of *n* numbers

6. Does *idf* (inverse document frequency) have an effect on ranking for one term queries?

7. IF *REL(i)* is the relevance of the *i*th result for *p* total results, define in a formula the Discounted Cumulative Gain, *DCG* =

Zipf's Law and Heap's Law are special cases of a power law, which has the form
$y = K*X^C$

8. For Zipf's Law what does *x* and *y* represent and what is the value of *C*?

9. For Heap's Law what does *x* and *y* represent and what is a typical value of *C*?

10. A distance measure, *D(x,y)*, must satisfy 4 properties for points *x* and *y*. What are they?

11. The data that satisfies Zipf's law is generally drawn using a log-log scale. Why?

12. Wikipedia gives four rules for normalizing URLs. Name them:

13. Name one technique used to speed up the merging of postings in an inverted index?

14. Consider the two sets ***A = {0, 1, 2, 5, 6}*** and ***B = {0, 2, 3, 5, 7, 9}.*** What is the Jaccard Similarity of A and B? What is the Jaccard distance of A and B?

15. In Google if the query is as shown in the next line

    President OR “Abraham Lincoln” died

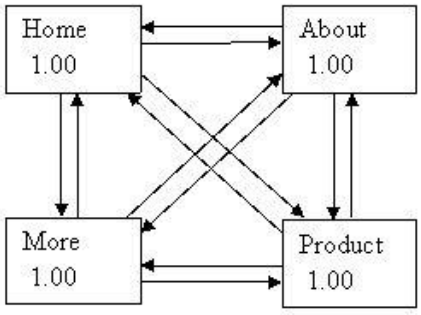
    Show how Google interprets the query by fully parenthesizing the query and insert all implied Boolean operators.

16. What is the purpose of the SoundEx algorithm?

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

18. Shown below are four web pages each containing links to every other web page. Each page initially has a PageRank of 1. Suppose a new page is added that points to Home and a link to the Product page is added that points to another web page at another website. Do the PageRank values of Home, About, More and Product go up, down or stay the same? Circle the appropriate answer Up or Down or The_Same for each of the 4 pages:

Home:	Up	Down	The_Same
About:	Up	Down	The_Same
More:	Up	Down	The_Same
Product:	Up	Down	The_Same



19. Are the final PageRank values for Home, About, More and Product the same or different?

20. The HITS algorithm divides the resulting set of web pages into a bipartite graph with two types of nodes (web pages). What are they?

21. The mean reciprocal rank is a statistical measure for evaluating any process that produces a list of possible responses to a sample of queries, ordered by probability of correctness. Given a set of *N* queries *Q* where the $rank(Q_i)$ is the rank of the i-th query, define the mean reciprocal rank.

22. Define Lexicon and Tokenization:

23. Name the system YouTube uses to identify content that is uploaded by someone who does not own the copyright.

24. and 25. (This question is worth 8 points). Let A, B, C, and D be the relevant documents and let W, X, Y, and Z be the irrelevant documents. Suppose for a given search query the results are returned as:
W, A, X, Y, B, C, D, Z
Compute the recall and precision at each fixed position

Result List	RECALL	PRECISION
W		
A		
X		
Y		
B		
C		
D		
Z		