Computer Science 572 Exam
Prof. Horowitz
Wednesday, October 4, 2017, 8:00am – 8:50am

Name: Siddhesh Rajiv Karekar **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.
   Mainframe computing, mini computing, personal computing, desktop internet computing
2. The number of websites, to an order of magnitude is: 10 million, 100 million, 1 billion, 10 billion or 1 trillion?
   ~1.7 billion
3. In two sentences define the deep web and the dark web?
   Deep web - can't be indexed (e.g. database, robots.txt), dark web - require specific software/authorization to access
4. True or false, by default Google maintains a user's entire query history?
   True
5. Define the algorithm for computing the harmonic mean of *n* numbers
   HM = n / (sum of reciprocals of the n numbers)
6. Does *idf* (inverse document frequency) have an effect on ranking for one term queries?
   No. the idf is constant for a term across a collection, so since each document's relevance is determined only by a single product (tf * idf) for that term t, it can be ignored.
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* = REL(1) + SUM i=2,p (REL(i) / log2(i))

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*?
   x: rank in frequency table, y: frequency, C: -1
9. For Heap's Law what does *x* and *y* represent and what is a typical value of *C*?
   x: document length, number of words; y: number of distinct words/vocabulary, C (beta) typically 0.4 - 0.6
10. A distance measure, *D(x,y)*, must satisfy 4 properties for points *x* and *y*. What are they?
    1. no negative distances, 2. distance = 0 iff x = y, 3. symmetry, 4. triangle inequality
11. The data that satisfies Zipf's law is generally drawn using a log-log scale. Why?
    on a log-log scale, it appears as a straight line
12. Wikipedia gives four rules for normalizing URLs. Name them:
    1. convert scheme/host to lower case, 2. Capitalize escape sequences, %-encoded triplets, 3. Decode %-encoded octet, 4. remove port 80
13. Name one technique used to speed up the merging of postings in an inverted index?
    Skip pointers
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?
    A intersection B = {0, 2, 5} 3 elements
    A union B = {0, 1, 2, 3, 5, 6, 7, 9} 8 elements
    sim = 3/8 = 0.375
    dist = 1 - sim = 0.625

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.

    (President OR "Abraham Lincoln") AND died

16. What is the purpose of the SoundEx algorithm?

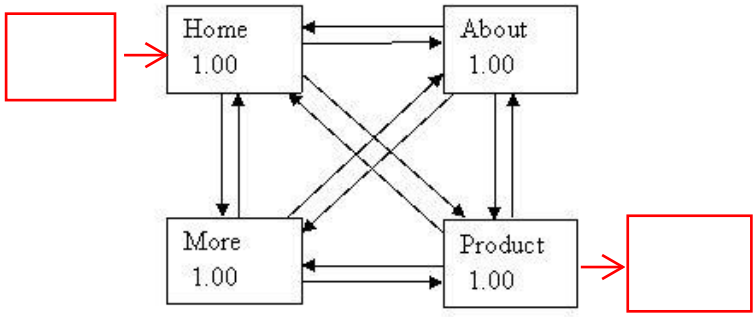
    To index together, or encode to the same representation words that are pronounced similarly in the English language

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

    A.B / (|A||B|)

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?

    All 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?

    Hubs and Authorities

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.

    MRR = sum i=1,N (1 / rank (Qi)) / N

22. Define Lexicon and Tokenization:

Lexicon: the entire database of unique words for a domain, including syntactic, semantic and morphological information
Tokenization: chopping a document unit into pieces, called tokens, and possibly throwing away certain characters

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

ContentID

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	0.25	1
A	0.25	0.5
X	0.5	0.67
Y	0.75	0.75
B	0.75	0.8
C	0.75	0.67
D	0.75	0.43
Z	1	0.5