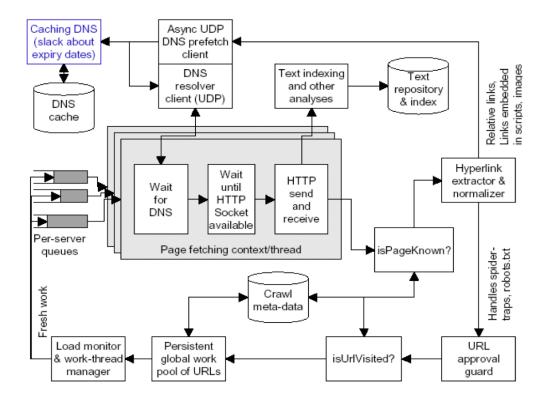
Computer Science 572 Exam Prof. Horowitz

Monday, February 26, 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 25 questions. Each question is worth 4 points.
- 4. Place your answer immediately below the question. Limit answers to ONE SENTENCE.
- 1. Define Precision and Recall.
- 2. The harmonic mean of precision and recall is sometimes called the *F* measure. If *R* represents Recall and *P* represents Precision, define *F*.
- 3. An Information Retrieval system returns 8 relevant documents and ten non-relevant documents. There are a total of 20 relevant documents in the collection. What is the precision and recall of the system on this search?
- 4. For a set of Q queries where the average precision of the i^{th} query is AvgP(i), define the Mean Average Precision function.
- 5. Suppose a search engine keeps track of how many people click on the top five search results for a specific query. If the number of clicks are 49, 36, 16, 42, and 4 respectively for results 1, 2, 3, 4 and 5, is it fair to conclude that the first result was the best one and why?
- 6. The class notes mention three strategies for coordination of distributed crawlers. What are the three strategies?



- 7. Above is an architectural diagram of the Mercator crawler. What element of the diagram controls freshness and politeness?
- 8. Above is an architectural diagram of the Mercator crawler. How many parallel threads are shown?
- 9. Crawlers need to check a new URL against the list of URLs already seen. Before making the check the URLs must be normalized. There are four steps required to normalize a URL. Name two of them.
- 10. The class notes define 4 properties that a distance measure must satisfy. What are they?
- 11. The class notes mention five specific distance measures. Name two of them.
- 12. Search engines must convert URLs to normal form so they can check if a new URL has already been seen. What other textual items/entities need to be converted to normal forms?

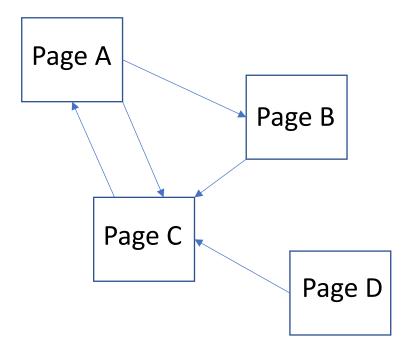
- 13. What is the Tika library used for?
- 14. Which is the more careful approach to transforming multiple forms of a word into a common base, Stemming or Lemmatization?
- 15. Using Porter's stemmer program, the words Universal, university and universe all stem to universe; is this an example of over-stemming or under-stemming?
- 16. Is the technique of using skip pointers helpful for handling OR queries, AND queries, or both OR and AND queries?
- 17 What is the meaning of hypernym?
- 18. Give an example of a hypernym and its related term.
- 19. The Resnik measure of similarity of two words c_1 and c_2 is defined by the following formula

$$sim_{resnik}(c_1,c_2) = -log P(LCS(c_1,c_2))$$

What is the meaning of P(LCS (c_1, c_2))?

- 20. Two words A and B are similar if their glosses contain similar words. What is the meaning of "glosses"?
- 21. If A and B are sets and |A| is the number of elements of A, and |B| is the number of elements of B, define the number of elements in A union B.
- 22. YouTube's recommendation system uses a co-visitation graph. For videos v1 and v2 define the co-visitation.

23. Below are four web pages A, B, C and D with connecting links. What page will have the highest PageRank? What page will have the smallest PageRank?



- 24. Can URLs that begin with https be used as initial seeds? Yes or No. If yes, what feature of crawler4j is needed to handle the https scheme?
- 25. In crawler4j, which function has to be modified to determine whether a URL must be visited or not?