Computer Science 572 Exam
Prof. Horowitz
Wednesday, February 22, 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. If *tp* represents true positive, *fp* represents false positive, *fn* represents false negative and *tn* represents true negative, define Precision and Recall

Answer:

2. The harmonic mean of precision and recall is sometimes called the *F* measure. If *R* represents Recall and *P* represents Precision, define *F*

Answer:

3. Define the three main properties of cryptographic hash functions

Answer:

4. Name one cryptographic hash function

Answer:

5. Given two sets A and B, define their Jaccard similarity

Answer:

6. If *df(t)* is the document frequency of term *t* out of *N* documents, what is the range of values *df(t)* can take on

Answer:

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

Answer:

8. One alternative to the use of document frequency is collection frequency. Define collection frequency for a term *t.*

Answer:

9. if *tf(t,d)* is the term frequency of term *t* in document *d*, out of a total of *N* documents, and *df(t)* is the document frequency of term *t*, define the *tf-idf* weight of term *t* in document *d*

Answer:

10. Given a query *q* and a set of documents *D*, and the *tf-idf(t,d)* the weighted score of term *t* in document *d* in *D*, what is the score of the query *q* with respect to document *d*

Answer:

11. State Heap's Law

Answer:

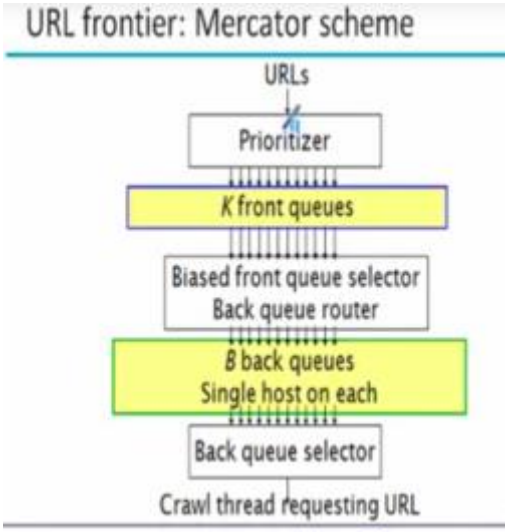
12. According to Zipf's law how often will the most frequent word occur as compared to the second most frequent word? How often will the most frequent word occur as compared to the third most frequent word?

Answer:

13. The rank/frequency data that satisfies Zipf's law is generally drawn using a log-log scale. Why?

Answer:

Below is a diagram that explains part of the Mercator crawler and it was the subject of one of the videos shown in class. Please answer the following questions about the diagram



14. Which queues control freshness, the front queues or the back queues?

Answer:

15. Which queues control politeness, the front queues or the back queues?

Answer:

16. Which queues make use of a min-heap data structure?

Answer:

17 In Google if the query is as shown in the next line
February President OR “Abe Lincoln” dies
show how Google interprets the query by fully parenthesizing the query and insert all Boolean operators

Answer:

18. Google offers a set of special operators that can be used to modify queries. Mention four:

Answer:

19. What is the technique used to speed up the merging of postings in an inverted index?

Answer:

20. Name four criteria YouTube uses to rank its search results.

Answer:

21. To come up with related videos, YouTube forms a co-visitation count for each pair of videos. What is the co-visitation count?

Answer:

22. An inverted index is often split into two parts, name them.

Answer:

23. In one sentence explain who uses ContentID and what it does.

Answer:

24. The Java program below is part of crawler4j and defines which pages to crawl. In particular crawler4j will not crawl css, js, gif, jpg, png, mp3, gz, and zip files. However there are two lines that include XXXXXXX denoting code that is missing. Fill in the missing code.

```
public class MyCrawler extends XXXXXXX {
   private final static Pattern FILTERS =
Pattern.compile(".*(\\.(XXXXXXX))$");
    @Override
    public boolean shouldVisit(Page referringPage, WebURL url) {
       String href = url.getURL().toLowerCase();
       return !FILTERS.matcher(href).matches()
            && href.startsWith("http://www.cnn.com/");
    }
```

Answer:

25. The Java program below is part your crawler4j homework exercise. However there are two lines that include XXXXXXX denoting code that is missing. Fill in the missing code.

```
public class XXXXXXX {
   public static void main(String[] args) throws Exception {
      String crawlStorageFolder = "/data/crawl";
      int numberOfCrawlers = 7;
      CrawlConfig config = new CrawlConfig();
      config.setCrawlStorageFolder(crawlStorageFolder);
      PageFetcher pageFetcher = new PageFetcher(config);
      RobotstxtConfig robotstxtConfig = new RobotstxtConfig();
      RobotstxtServer robotstxtServer = new RobotstxtServer(robotstxtConfig, pageFetcher);
      CrawlController controller = new XXXXXXX;
      controller.addSeed("http://www.cnn.com/”);
     controller.start(MyCrawler.class, numberOfCrawlers);
   }
}
```

Answer: