# Computer Science 572 Exam(1)

## Prof. Horowitz

### Monday, March 2, 2020, 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. [4 pts] When a search engine crawls the web and visits a website for the first time, what is the first file the crawler should look for?

2. [4 pts] For the file mentioned above, where on the server is the file located?

3. What is the order of magnitude of the number of websites that exist today

a. 1,000,000. b. 10,000,000. c. 100,000,000. d. 1,000,000,000 e. 10,000,000,000. f. 100,000,000,000

4. If *tp* represents true positive, *fp* represents false positive, *fn* represents false negative and *tn* represents true negative, define Precision and Recall

5. [4 pts] In the formula for Discounted Cumulative Gain, how are documents appearing lower in a search result list penalized?

6. [4 pts] Approximately how many different file types does Google index?
   a.10 b.30 c.100 d.150 e. 200

7. [4 pts] If *k* and *c* are constants, *x* and *y* variables, state Zipf's Law both as a formula and one sentence explanation

8. [4 pts] In one sentence, what is the purpose of the Soundex Algorithm?

9. [4 pts] In one sentence what is the purpose of Porter's Algorithm?

10. [4 pts] What data structure was suggested as the best way of determining if a URL has been seen before by a search engine?

11. [4 pts] What are the names of the Google and Yahoo! web crawlers?

12. [4 pts] Three methods for distributed crawling were mentioned. Name all three and for each method, in one sentence each, describe how it is supposed to work.

13. [4 pts] The class notes define 4 properties that a distance measure must satisfy. For points *x, y,* what are they?

14. [4  pts] What is de-duplication and give two examples of why it needs to be done.

15. [4  pts] True or False, Google, Yahoo, and Bing record all user clicks, both on ads and on organic search results.

16. [4  pts] YouTube uses an 11 digit identifier for its uploaded videos. In the video we saw explaining the process, what was the base that YouTube uses to come up with the identifier.

17. [4  pts] What is the purpose of YouTube's ContentID system

18. [4 pts] YouTube's recommendation system uses a co-visitation graph. Define the co-visitation graph by saying what the nodes A, B, C, . . . and edges <A,B>, <A,C>, etc. of the graph represent.

19. [4 pts] Google offers a variety of special operators that can be used to narrow a search. Describe how these operators work in Google search: filetype, site, allinanchor.

20. [4 pts] Suppose there are only two web pages, each with only one link that points to the other web page. What will be the PageRank of each page?

21. [4  pts] As a website grows and adds more pages with more links to web pages outside of the website, how is the total PageRank of the website affected?

Consider the following table containing the URLs for the top five search results from two different search engines.

| Google | Assigned Search Engine |
|---|---|
| "https://www.howacarworks.com/cooling-systems/how-to-replace-a-car-thermostat" | “https://www.dummies.com/home-garden/car-repair/heating-cooling-system/how-to-replace-your-vehicles-thermostat/” |
| "https://shop.advanceautoparts.com/r/car-projects/how-to-replace-a-vehicle-thermostat" | ”https://www.familyhandyman.com/automotive/car-maintenance/how-to-change-coolant/” |
| "https://www.dummies.com/home-garden/car-repair/heating-cooling-system/how-to-replace-your-vehicles-thermostat/" | “https://www.wikihow.com/Replace-a-Thermostat” |
| "https://axleaddict.com/auto-repair/How-to-Replace-a-Thermostat" | “https://www.carid.com/articles/when-is-it-time-to-replace-my-engine-thermostat.html” |
| "https://www.readersdigest.ca/cars/maintenance/replace-thermostat-instructions/" | “https://www.howacarworks.com/cooling-systems/how-to-replace-a-car-thermostat” |

22. [4 pts] What is the spearman coefficient for the above mentioned results?
Formula for spearman coefficient :

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}.$$

1) -1.45
2) -19

3) 0

4) 2

23. [4 pts] What is the percentage overlap for the above mentioned results?
    1) 40%
    2) 60%
    3) 20%
    4) 80%

24. [4 pts] As per HW1, What will be the value of rho (spearman coefficient) when there are no overlapping results between Google and your assigned search engine?
    1) 1

    2) -1

    3) 0

    4) 0.5

25. [4 pts] In crawler4j the controller class has several parameters that were set as part of the exercise. Name three of them: