Computer Science 572 Exam
Prof. Horowitz
Monday, March 8, 2020, 7:00pm – 8:00pm

Name: **Student Id Number:**

1. This is an open book exam.
2. Please answer all questions.
3. There are a total of 32 questions. Question points vary.
4. **Place your answer immediately below or to the right of the question. Limit answers to ONE SENTENCE.**
5. **Note: there is no re-grading permitted**

1. [2 pts] If P stands for precision and R for recall
a. What is the definition of the F measure in terms of P and R?
b. A beta parameter for the F measure was presented. Is there a limit to the range of beta, yes or no?
c. Can the value of beta be negative?

2. [2 pts] In the class notes Accuracy is defined as

(tp + tn) / (tp + fp + fn + tn)

Give an example where accuracy is very high, but tp is very low?

3. [2 pts] How is the time required by DNS resolution reduced in a web crawler?

4. [2 pts] In our textbook and in a video shown in class, two sets of queues (*back queues* and *front queues*) were used to implement politeness and prioritization. Which ones implement politeness and which ones implement prioritization?

5. [2 pts] The number of content types indexed by Google is approximately (circle your answer)?
10
50
100
500

6. [2 pts] At a website where will a web crawler look to find the robots.txt file?

7. [2 pts] If *P(n)* is the frequency of occurrence of the *n*-th ranked word, then according to Zipf's Law what can you say about *P(n)*

8. [2 pts] We have seen the concepts of *term document frequency* and *inverse document frequency*. Define *collection frequency* and explain why it is not used by referring to the following example:

Word	Collection frequency	Document frequency
Insurance	10440	3997
Try	10422	8760

9. [2 pts] Define the traditional formula for *tf-idf* weighting as shown in the class notes and the videos

10. [2 pts] Define the formula for scoring a document ***d*** against a query ***q*** using document frequency as the weighting factor.

11. [2 pts] Run the Soundex algorithm on the two terms: Beijing and Peking.
Are the two a match? Yes or No?

12. [4 pts] Give 4 examples of the morphology of the term *computer*

13. [2 pts] Taking the words *painting*, *painted*, and *paints* and storing them in the lexicon as *paint* is an example of what information retrieval principle?

14. [2 pts] In one sentence describe the difference between stemming and lemmatization

15. [6 pts] While handling the works of Shakespeare as contained in the file shakes.txt the following UNIX command was issued:

```
tr  -sc 'A-Za-z'  '\n' < shakes.txt   | sort  |  uniq
```

For each of the UNIX commands explain exactly what they are doing in the above statement.

16. [2 pts] An inverted index generally is composed of two parts, the dictionary (which contains the vocabulary) and the postings list which contains for each vocabulary word the list of documents that contain the word. We looked at two techniques for phrase matching: a. bi-word indexing and b. positional indexing. Which technique expands the dictionary and which technique expands the postings list?

17. [2 pts] What data structure was suggested as the best way of determining if a URL has been seen before by a search engine?

18. [2 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.

19. [2 pts] In one sentence explain the purpose of YouTube's ContentID system

20. [2 pts] A cryptographic hash function of file X has four main properties:
1. it is easy to compute
2. it is difficult to find a file that has the same hash value,
3. a small change to the text yields a totally different hash value
What is the fourth property?

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

Google	Assigned Search Engine
"https://www.livescience.com/33991-difference-fruits-vegetables.html"	"https://www.questionsanswered.net/article/how-measure-differential-pressure?ad=dirN&qo=serpIndex&o=740012"
"https://www.healthline.com/nutrition/fruits-vs-vegetables"	"https://www.britannica.com/story/is-a-tomato-a-fruit-or-a-vegetable"
"https://fruitsandveggies.org/expert-advice/difference-fruit-vegetable-2/"	"https://www.healthline.com/health/food-safety-fruits-vegetables"
"https://www.bhg.com/gardening/vegetable/difference-between-fruits-vegetables/"	"https://www.verywellfit.com/getting-more-fruits-and-vegetables-in-your-diet-2506856"
"https://recipes.howstuffworks.com/difference-between-fruit-and-vegetable.htm"	"https://www.healthline.com/nutrition/fruits-vs-vegetables"
"https://www.britannica.com/story/is-a-tomato-a-fruit-or-a-vegetable"	"https://www.livescience.com/33991/difference-fruits-vegetables.html"

21. [2 pts] From the four options below, which is the correct Spearman Coefficient for the above mentioned results? Circle your answer. ***Look carefully at the URLs***. The formula for Spearman Coefficient :

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

a) -30.5
b) -24
c) 0
d) -11.5

22. [2 pts] What is the percentage overlap for the above-mentioned results? Circle your answer.

a) 50%
b) 33.33%
c) 16.67%
d) 66.67%

23. [2 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

The following questions pertain to HW2. Consider the following URLs your crawler discovers:
https://www.nytimes.com/en/java/javase/15/install/installation-jdk-macos.css
https://www.nytimes.com/en/java/javase/15/install/installation-jdk-macos.jpeg

24. [6 pts] As per HW2, what regular expression should be used in order to allow (match) these URLs for further processing in one of the visit methods? Circle your answer.

```
1).  .*(css|js|bmp|gif|jpe?g)$
2). .^(css|js|bmp|gif|jpe?g)^
3). .^(css|js|bmp|gif|jpe?g)$
4). *.(css|js|bmp|gif|jpe?g)$
```

25. [2 pts] In which method should the above regular expression be placed in order to filter out URLs that are NOT to be crawled?

26. [2 pts] Explain in brief the difference between the `shouldVisit()` and `visit()` ?

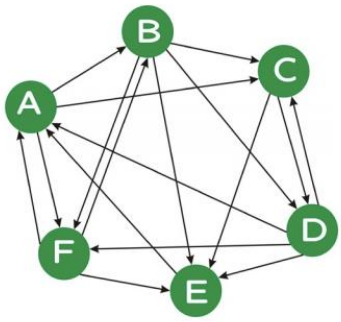
27. [6 pts] In crawler4j the controller class has several parameters that were set as part of the exercise. Name three of them:

28. [4 pts] What are the return types of the `shouldVisit()` and `visit()` that are overridden?

29. [2 pts] In the discussion of word similarity, one technique for deciding if two words are similar took this approach: look at the words included in the definition of the two words; if there is “significant” overlap in the terms used to define the two words, then we say the two words are similar. What is the name of this algorithm?

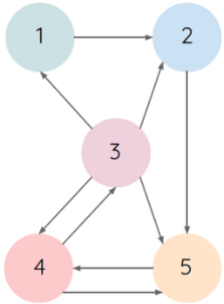
30. [6 pts] In our discussion of distributional similarity two terms are considered similar if they are surrounded by similar words. A term-document matrix is used to compute similarity where the columns are documents. Answer the three questions below:

a. In this method what are the rows of the term-document matrix,
b. What are the entries in the term-document matrix, and
c. what is the definition of similarity?



31. [10 pts] Above is a directed graph with six nodes representing web pages. Draw the initial 6 x 6 PageRank matrix whose rows and columns represent the nodes A, B, C, D, E, and F, and where the (i, j)th value represents the amount of PageRank initially assigned to the node.

32. [10 pts] Below is a directed graph with five nodes. Applying the PageRank algorithm (iteration 0) we assume that initially all nodes have a PageRank of 1/5. To compute iteration 1 we start with Node1 which has one link from Node3. Node3 has four outbound links, so we take the rank of Node3 from iteration 0 and divide it by 4, giving a rank for Node1 = 1/20. Compute the ranks of Node2, Node3, Node4 and Node5 for iteration 1



Rank of: Node2 Node3 Node4 Node5