Computer Science 572 Exam
Prof. Horowitz
Tuesday, February 22, 2022, 7:00pm – 8:00pm
ONLY ONE ANSWER SUBMISSION IS PERMITTED UP TO 8:30PM (PST)

Name: **Student Id Number:**

1. **This is an open book exam. You may consult any resource.**
2. **Please answer all questions.**
3. **Question points vary.**
4. **Place your answer immediately below the question or on a separate sheet of paper where it is clear what question you are answering. REMEMBER to put your name and ID on any separate sheet of paper**
5. ***WARNING: Uploading your answers at the last minute will likely fail due to server overload, UPLOAD EARLY!***

1. [3 pts] Search engines use *precision* and *recall* for evaluating how good their results are. But at least three other techniques were cited in the notes. Name two of them.

2. [3 pts] Search engines must put URLs into a canonical form. Transform the following URLs into canonical form as described in class

   http://www.betasoftware.com/%7Emyfolder
   http://www.betasoftware.com:8081/a%c2%b1b
   http://www.betasoftware.com/FOLDERX/marie%2dsmith

3. [3 pts] Crawlers must do a lot of DNS lookups. Name two steps that are taken to make this more efficient.

4. [3 pts] Suppose you have a document with *n* words and you are creating *k-shingles*. How many shingles will be produced?

5. [3 pts] Typically should white space be counted when creating shingles,

Yes or NO?

6. [3 pts] Consider the (k = 2)-shingles for each document D1, D2, D3, and D4:
D1 : [I am], [am Sam]
D2 : [Sam I], [I am]
D3 : [I do], [do not], [not like], [like green], [green eggs], [eggs and], [and ham]
D4 : [I do], [do not], [not like], [like them], [them Sam], [Sam I], [I am]

Compute the Jaccard similarity for each pair of documents to at most 2 decimal digits:
JS(D1, D2) =
JS(D1, D3) =
JS(D1, D4) =

7. [3 pts] In one of the videos from class, a result by Church and Gale was cited that is similar to Heaps Law. What is the result and remember to define your terms?

8. [3 pts] The following statement has four parts separated by vertical bars (|). Explain what each part of the statement accomplishes, where shakes.txt contains all of the works of Shakespeare

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

Part 1:
Part 2:
Part 3:
Part 4:

9. [3 pts] For sets *A* and *B* what is the formula involving A and B that calculates the size of set A union set B, or | *A U B* | ?

10. [3 pts] Suppose there are five relevant documents, R1, R2, R3, R4, and R5 and there are five non-relevant documents NR1, NR2, NR3, NR4 and NR5; suppose the search engine ranking algorithm returns the documents in the following order

R3, NR4, R2, R1, NR1, NR2, R4, R5, NR3, NR5

Compute below the ten pairs of (Recall, Precision) values associated with each of the ten results. Use two decimal places and don't round your answer up or down

	R3,	NR4,	R2,	R1,	NR1,	NR2,	R4,	R5,	NR3,	NR5
Recall										
Precision										

11. [3 pts] To test if two words are similar one of the videos suggests a few methods. Mention two.

12. [3 pts] The Lesk Algorithm was cited in one of our videos as a way to determine if two concepts are similar. In one sentence, what is the Lesk method?

13. [3 pts] In the class notes what formula is given for defining the term *Accuracy*?

14. [3 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 in a web server. Which queues implement politeness and which ones implement prioritization?

15. [3 pts] The number of content types indexed by Google is approximately (circle your answer)?

10
50
100
500

16. [1 pts] At a website where will a web crawler look to find the robots.txt file?

17. [3 pts] If *P(n)* is the frequency of occurrence of the *n-th* ranked word, then according to Zipf's Law *P(n)* it is proportional to 1 over *n* raised to a power *k*, or it is inversely proportional to the rank. To show Zipf's law as a straight line, what must the axes of the graph be?

18. [3 pts] Run the Soundex algorithm on the two terms: Mumbai and Bombay. What are the results and are the resulting numbers the same?

19. [3 pts] Give 4 examples of the morphology of the term “computer”.

20. [3 pts] In one sentence describe the difference between stemming and lemmatization.

21. [3 pts] An inverted index generally is composed of two parts, the dictionary and the postings list. 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?

22. [3 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?

23. [3 pts] In one sentence explain the purpose of YouTube’s ContentID system.

24. [3 pts] A cryptographic hash function of file X has four main properties. One property is that it is easy to compute. What are the other three properties?

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"

25. [3 pts] From the four options below, what is the correct Spearman Coefficient for the above mentioned results? Circle the correct answer below. Look carefully at the URLs. The formula for the Spearman Coefficient is:

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

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

26. [2 pts] What is the percentage overlap for the above-mentioned results?

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

27. [2 pts] As per HW1, what will be the value of the Spearman Coefficient when there are no overlapping results between Google and your assigned search engine? Circle the correct answer.

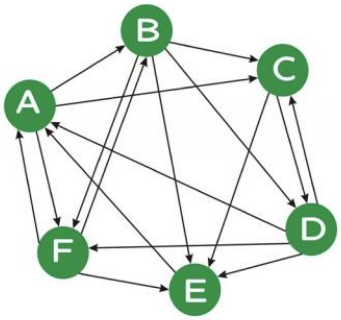
1) 1
2) -1
3) 0
4) 0.5

28. [3 pts] In HW2, in which method should a regular expression be placed in order to filter out URLs that are NOT to be crawled?

29. [3 pts] Explain in brief the difference between the `shouldVisit()` and `visit()` ?

30. [3 pts] In crawler4j the controller class has several parameters that were set as part of the exercise. Name three of them:

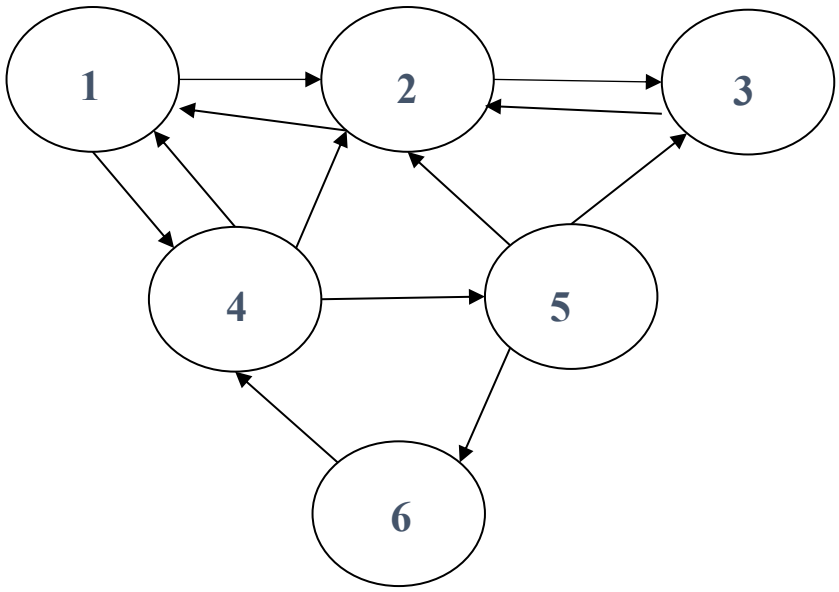
31. [3 pts] What are the return types of the `shouldVisit()` and `visit()` that are overridden?



32. [3 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.

Place your answer here

33. [4 pts] Below is a directed graph with six nodes. Applying the simple PageRank algorithm (iteration 0) we assume that initially all nodes have a PageRank of 1/6. To compute iteration 1 we start with Node1 which has two outgoing links, one incoming link from Node2 and one incoming link from Node 4. Below compute the page ranks of Nodes 1, 2, 3, 4, 5 and 6 for iteration 1.



Nodes	Iteration 0	Iteration 1
1	1/6	
2	1/6	
3	1/6	
4	1/6	
5	1/6	
6	1/6	

34. [4 pts] The query below has three terms AND'd together, say T1, T2 and T3; Using the document frequency of each term as shown below, recommend a query processing order for the query

```
(grapes OR plums) AND (strawberries OR oranges) AND (bananas OR pears)
```

Where T1 is (grapes OR plums), T2 is (strawberries OR oranges) and T3 is (bananas OR pears)

TERM	FREQUENCY
Grapes	213456
Plums	67111
Strawberries	107345
Oranges	271321
Bananas	36555
Pears	316234