Computer Science 572 Exam
Prof. Horowitz
Tuesday, October 5, 2020, 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.**
2. **Please answer all questions.**
3. **There are a total of 34 questions. 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. Limit answers to ONE SENTENCE.**
5. **Note: there is no re-grading permitted**

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.

Average Precision, Mean-Average Precision, F-measure, Cumulative Gain, Discounted Cumulative Gain, Normalized DCG, Query Logs

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.trellis.com
   http://www.trellis.com/%7Emyfolder
   http://www.trellis.com:8081/a%c2%b1b

   http://www.trellis.com
   http://www.trellis.com/-myfolder tilda in place of '-'
   http://www.trellis.com/a%C2%B1b

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

   DNS caching, Pre-Fetching client & many DNS resolvers

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

   nCk

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

   Yes or NO?

   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) = 0.33
   JS(D1, D3) = 0
   JS(D1, D4) = 0.125

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 3 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 | tr 'A-Z' 'a-z' | grep 'ing$'
```

   Part 1: takes every non-alphabetic character from 'shakes.txt' and translates it into a newline character
   Part 2: translates all uppercase letters to lowercase letters
   Part 3: finds any line that contains the regular expression 'ing$'

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

   |A U B|

10. [3 pts] In one of the videos synonymy was defined as a binary relation, but similarity was defined differently. What definition for similarity was used?

    Two words are more similar if they share more features of meaning. Similarity is a proper relation between senses.

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

Path-based Similarity (simpath(c1,c2) & wordsim(w1,w2)), Information Content Similarity (Resnick Method, Dekang Lin Method, JiangConrath Method), Lesk Similarity

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?

    A thesauras based measure that looks at glosses.

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

The fraction of classifications that are correct-
(tp + tn) / (tp + fp + fn + tn)

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?

Back Queues - Politeness
Front Queues - Prioritization/Freshness

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?

root directory

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?

P(n) = k*((n)^-1) - so, inversely proportional to the rank
Axes:- y-axis: P(n) & x-axis: n

18. [3 pts] Define the traditional formula for *tf-idf* weighting as shown in the class notes and the videos. Define each term in the formula.

wij = tfij .idfi = (1 + log tfij)* log2 (N/ dfi)

tfij - term frequency (fij / max{fij} where fij = frequency of term i in document j)
ifdi - inverse document frequency of term i (log2 (N/ dfi) where N: total number of documents)

19. [3 pts] Run the Soundex algorithm on the two terms: Beijing and Peking. Give the algorithm output and state if the two terms are a match?

Beijing - B252
Peking - P252

The two terms aren't a match.

20. [3 pts] Give 4 examples of the morphology of the term computer

compute, computes, computing, computed

21. [3 pts] In one sentence describe the difference between stemming and lemmatization

Stemming may or may not result in an actual word whereas lemmatization results in an actual language word.

22. [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?

Bi-word indexing expands the dictionary. Postional indexing expands the postings list.

23. [3 pts] What data structure was suggested as the best way of determining if a URL has been seen before by a search engine?

Trie

24. [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.

64

25. [3 pts] In one sentence explain the purpose of YouTube's ContentID system

YouTube's solution was to create a fingerprint database of copyrighted content, called Content ID.

26. [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?

The cryptographic hash function has four main properties:
1. It is extremely easy (i.e. fast) to calculate a hash for any given data.
2. It is extremely computationally difficult to calculate an alphanumeric text that has a given hash.
3. A small change to the text yields a totally different hash value.
4. It is extremely unlikely that two slightly different messages will have the same hash.

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

Google	Assigned Search Engine
https://www.britannica.com/place/England	https://historyofengland.typepad.com/
https://en.wikipedia.org/wiki/History_of_England	https://en.wikipedia.org/wiki/History_of_England
https://www.english-heritage.org.uk/learn/story-of-england/	https://thehistoryofengland.co.uk/
https://en.wikipedia.org/wiki/England	https://www.eupedia.com/england/english_history.shtml
https://thehistoryofengland.co.uk/	https://www.youtube.com/watch?v=73SSODWU2fU
https://oll.libertyfund.org/title/todd-the-history-of-england-6-vols	https://www.britannica.com/place/England

27. [3 pts] From the four options below, which is the correct Spearman Coefficient for the above mentioned results? Look carefully at the URLs. The formula for Spearman Coefficient:

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

a) -28
b) -12.5
c) -6.25 ✓
d) -5

28. [3 pts] What is the percentage overlap for the above-mentioned results?

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

29. [3 pts] As per HW1, what will be the value of rho (Spearman Coefficient) when there is only 1 overlapping result between Google and the assigned search engine which at the same rank in both sets of results?

a) 1 ✓
b) -1
c) 0
d) 0.5

30. [3 pts] In which method should a regular expression be placed in order to filter out URLs that are NOT to be crawled?

shouldVisit()

31. [3 pts] Explain in brief the difference between the `shouldVisit()` and `visit()` ?

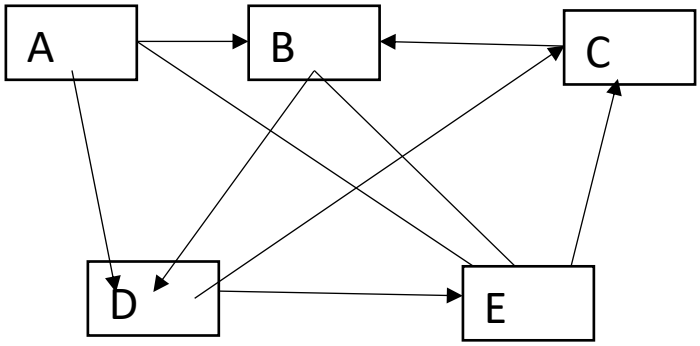
shouldVisit() - You should implement this function to specify whether the given url should be crawled or not (based on your crawling logic).
visit() - This function is called when a page is fetched and ready to be processed by your program.

32. [3 pts] In crawler4j the controller class has several parameters that were set as part of the exercise. Name three of them:

crawlStorageFolder, numberOfCrawlers, MaxPagesToFetch

33. [3 pts] What are the return types of the `shouldVisit()` and `visit()` that are overridden?

shouldVisit()- boolean
visit()- no return type



34. [3 pts] Above is a directed graph with five nodes representing web pages. Draw the initial 5 x 5 PageRank matrix whose rows and columns represent the nodes A, B, C, D, and E, and where the (i, j)th value represents the amount of PageRank initially assigned to the node.