Computer Science 572 Exam
Prof. Horowitz
Wednesday, October 7, 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 38 questions. Question points vary.
4. **Place your answer immediately below the question. Limit answers to ONE SENTENCE.**

1. [1 pts] The total number of websites is approximately (circle your answer):

500 million
1 billion
1.8 billion
10 billion

2. [1 pts] The total number of active websites are approximately (circle your answer):

10 million
200 Million
1 Billion
2 Billion

3. [1 pts] Which TLD has the largest number of registrations?

4. [1 pts] The number of content types indexed by Google is approximately (circle your answer)?

10
50
100
500

5. [2 pts] Google has suggested that the total number of unique URLs is at least (circle your answer)

100 Million
1 Billion
100 Billion
1 Trillion

30 Trillion

6. [2 pts] Let P stand for precision and R stand for recall. Assuming no other parameters, what is the definition of the combined measure F?

7. [2 pts] Define the Harmonic Mean of n numbers $x_1, \ldots, x_n$

8. [2 pts] How is the time required by DNS resolution resolved in a web crawler?

9. [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?

10. [2 pts] At a website where will a web crawler look to find the robots.txt file?

11. [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)*

12. [2 pts] In one sentence describe the purpose of the Soundex Algorithm?

13. [2 pts] Define the term-document incidence matrix

14. [4 pts] Write out the 3-grams for the phrase:
"Fourscore and seven years ago our fathers brought forth a nation"

15. [2 pts] What data structure was suggested as the best way of determining if a URL has been seen before by a search engine?

16. [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.

17. [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.

18. [2 pts] What is the purpose of YouTube's ContentID system

You are building a new inverted index system for your web search engine company Acme La Vista that is a focused vertical search engine for Italian Food restaurants. Your inverted index includes 3 documents (*D1…D3*) fetched by your crawler, which are decomposed into the following words:

D1	Italian luigis Fetuccini pasta Heineken
D2	Dona Marie Italian sausage tortellini
D3	Fetuccini butter salt oregano garlic

19. [6 pts] Derive the term dictionary ("index file") and the postings file for the above 3 documents.

20. [2 pts] David Filo and Jerry Yang are (circle the best answer):

- founders of Google

- creators of spreadsheets

- founders of Yahoo

21. [2 pts] Google first appeared on the web in (circle the best answer):

- 1998

- 2004

- 2010

22. [2 pts] Which content type is NOT indexed by Google? (circle the best answer):

- swf

- xlsx

- rtf

- svg

- all of the above are indexed

23. [2 pts] A cryptographic hash function of file X has three main properties:

1. it is easy to compute

2. it is difficult to find a file that has the same hash value,

What is the third property?

24. [2 pts] In one sentence define the different between stemming and lemmatization

25. [2 pts] In the class we have often mentioned Google, Yahoo, and Bing as the major web search engines. However, others have also been mentioned. Name three:

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”

26. [4 pts] What is the spearman coefficient for the above mentioned results? ***Look carefully at the URLs***. The formula for spearman coefficient :

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

1) -30.5

2) -24

3) 0

4) -11.5

27. [4 pts] What is the percentage overlap for the above-mentioned results?

1) 50%

2) 33.33%

3) 16.67%

4) 66.67%

28. [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 you come across in HW2 (Web Crawling):

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

29. [4 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?

```
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)$
```

30. [2 pts] In which method should the above regular expression be placed in order to filter out URLs that are NOT to be crawled?

31. [2 pts] Explain in brief the difference between the shouldVisit() and visit() ?

32. [2 pts] In crawler4j the controller class has several parameters that were set as part of the exercise. Name three of them:

33. [2 pts] What is a seed URL?

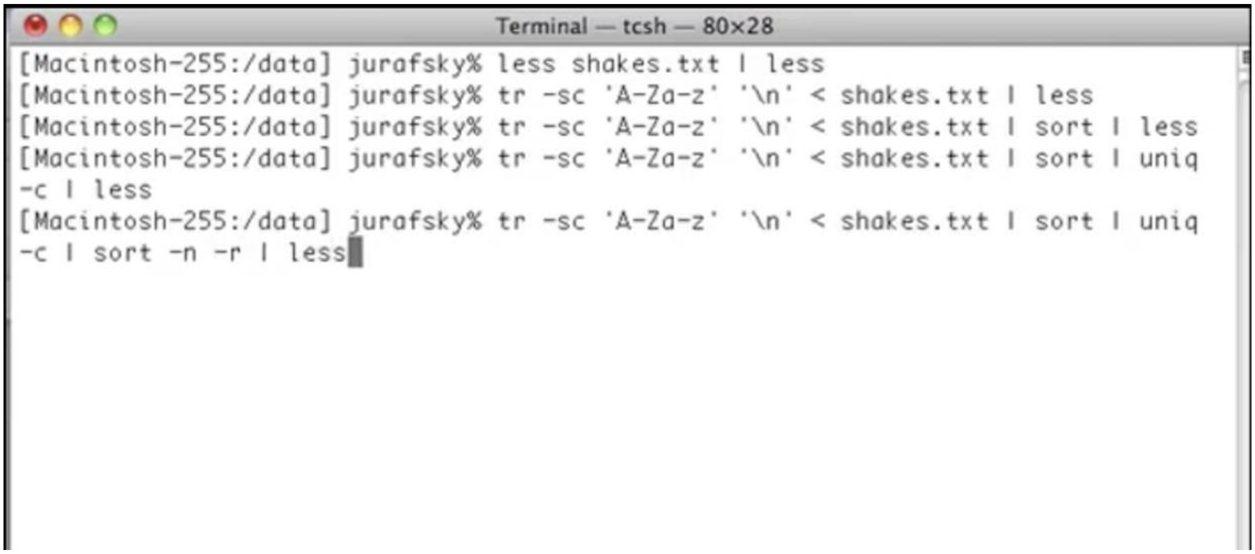
34. [2 pts] What is the return type of the shouldVisit() and visit() that are overridden?

35. [8 pts] In the formula below, using one sentence for each, define the four terms (everything except the 1 and log $_{10}$)

$$\mathrm{w}_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \times \log_{10}(N / \mathrm{df}_t)$$

36. [8 pts] In the Word Tokenization video by Jurafsky he is working with a text file containing all of the words from Shakespeare’s plays. To obtain a list of the words and their frequencies he uses the following UNIX/MacOS operators:

```
tr -sc ‘A-Za-z’‘\n’
sort
uniq
less
```



```
[Macintosh-255:/data] jurafsky% less shakes.txt | less
[Macintosh-255:/data] jurafsky% tr -sc 'A-Za-z' '\n' < shakes.txt | less
[Macintosh-255:/data] jurafsky% tr -sc 'A-Za-z' '\n' < shakes.txt | sort | less
[Macintosh-255:/data] jurafsky% tr -sc 'A-Za-z' '\n' < shakes.txt | sort | uniq
-c | less
[Macintosh-255:/data] jurafsky% tr -sc 'A-Za-z' '\n' < shakes.txt | sort | uniq
-c | sort -n -r | less
```

Please explain the operators above:

37. [4 pts] In the video *Word Similarity and Thesaurus Methods* Jurafsky mentions the terms *hypernym hierarchy* and *glosses.* Using one sentence for each, define the terms:

38. [2 pts] According to the "Characterizing the Web" lecture, approximately how many internet users are there worldwide (circle your answer)?

500 million
1 Billion
3.8 Billion
6 Billion