Computer Science 572 Exam
Prof. Horowitz
Thursday, April 28, 2022, 7:00pm – 8:30pm

Name: **Student Id Number:**

INSTRUCTIONS

1. **This is an open book exam. You may consult any resource. The exam is intentionally long!**
2. **There are 34 questions, point value may vary.**
3. **Some questions are short, but some are long.**
4. **Place your answer immediately below the question or if necessary on a separate piece of paper with your name, ID, question numbers and your answers.**
5. ***WARNING: By student request, test answer submissions after 8:30pm will automatically be penalized 10%. So please don't attempt to upload your answers at the last minute. UPLOAD EARLY!***

In homework #3, suppose you have a text file whose name is hello.txt with the following data

```
Data Science
Data is Science
```

1. [3 points] What are the input key-value pairs that are sent to the Mapper function?
Hint : You may refer to TextInputFormat subclass of FileInputFormat (an implementation of InputFormat interface in Java)

2. [2 points] What are the two environment variables that get set when we open a new ssh terminal on GCP?

3. [3 points] After performing the Map-Reduce operation in GCP, and before merging the output files, the files generated were in the format "part-r-xxxxx". "r" here stands for reducer output. What does "xxxxx" represent?

4. [2 points] Name two things about GFS that are different from a conventional UNIX file system

5. [3 points] Do the Google workstations that run the GFS make use of the Linux File System as well, YES or NO?

6. [2 points] If the WordMapper class and WordReducer class are nested within an outer class (whose name is, say, WordIndexer), what reserved keywords and/ or access specifiers should the Mapper and Reducer class have?

7. [2 points] In one sentence define an Ad Network and in another sentence an Ad Exchange, emphasizing the differences.

8. [3 points] In general do keywords that reference rare or generally unusual items have higher or lower cost-per-click prices.

9. [3 points] Suppose you set a maximum CPC of US$10.00. If 250 people see your ad and 20 of them click on your ad, how much money will you owe Google?
a. exactly $200.00
b. likely less than $200.00
c. likely more than $200.00

Given the following set of facts stored in a knowledgebase

1. If X barks and X eats Alpo, then X is a dog.
2. If Y is cold and Y tastes sweet, then Y is ice cream.
3. If X is a dog, then X is brown.
4. If Y is ice cream, then it is chocolate.

10. [2 points] This question has two parts, A and B:
A. State the modus ponens rule
B. Citing one or more of the four facts above, show how to use modus ponens to establish that if Rover barks and Rover eats Alpo then Rover is brown

11. [3 points]. In Solr what file contains the configuration for data dictionary?

12. [3 points]. In Solr what file contains definitions of the field types and fields of a document?

13. [3 points] What element did we add to Solr's configuration file(s) in HW#4 so that it can access the external_PageRankFile whenever the index is reloaded?

14. [3 points] A student comes across the following error after submitting his Map-Reduce job to GCP

Error : java.lang.RuntimeException : java.lang.NoSuchMethodException : InvertedIndexJob$IDMapper<init>

What could be the root cause? Assume the student has defined his Mapper & Reducer classes inside the Unigram class

15. [3 points] What is the role of Request Handler? Circle all correct answer(s)

a. Search query request is processed by Request Handler.
b. Index update request is processed by Request Handler.
c. Request Handler parses the queries that we pass to Solr and verifies the queries for syntactical errors
d. Request Handler generates the formatted output for the user queries

16. [3 points] Writing "their" when you mean "they're" is what kind of spelling error?

Below is a portion of the program for computing the Levenshtein distance between two given character strings s and t and below that is a table used by Levenshtein to determine the minimum edit distance between the two strings. Answer the questions below

```
function LevenshteinDistance(char s[1..m], char t[1..n]):
//for all i and j, d[i,j] will hold the Levenshtein distance between
//the first i characters of s and the first j characters of t
. . . . . .
for j from 1 to n:
      for i from 1 to m:
            if s[i] = t[j] then substitutionCost := 0 else
substitutionCost := 1
            d[i,j] := min (d[i-1,j] + 1,             // deletion
                      d[i,j-1] + 1,                  //insertion
                     d[i-1, j-1] + substitutionCost  //substitution
                    )
return d[m,n]
```

17. [3 points] For the Levenshtein algorithm above
A. What is the cost for deletion?
B. What is the cost for insertion?
C. What is the cost for substitution?

18. [6 points] Using the program for Levenshtein above, fill in the correct values in the table below:

	#	**S**	**P**	**U**	**R**	**S**
#	0	1	2	3	4	5
P	1					
A	2					
C	3					
E	4					
R	5					
S	6					

19. [3 points] Snippets are computed at indexing time, TRUE or FALSE?

20. [3 points] A single snippet is computed for each web page, TRUE or FALSE?

21. [3 points] In generating a snippet, one Google factor is the distance from the start of the web page text, TRUE or FALSE?

22. [3 points] Name two properties that are typically used to determine if a clustering algorithm works well.

23. [3 points] What is the range of values for:
a. cosine similarity,
b. Pearson coefficient,
c. Jaccard Similarity

24. [3 points] During the process of Question/Answering, a part-of-speech recognizer identifies question keywords to be examined by a search engine. Three techniques for increasing the keyword set were mentioned. Name the three and give an example of each:

25. [4 points] For the following words and relations provide an example in the column labeled **Example**

Word	**Relation**	**Example**
Breakfast	Hypernym	
Meal	Hyponym	
Table	Meronym	
Dinner	Holonym	
Leader	Antonym	

26. [3 points] Below are three lines that represent the start of the definition of the WordNet hyponym Synset for the term "bass" (an adult male singer with the lowest voice).

Bass, basso (an adult male singer with the lowest voice)
- ⇨ Singer, vocalist, vocalizer,
  - ⇨ Musician, instrumentalist, player

Below are the nine remaining terms making up the complete Synset. Put them in their proper order:

```
whole, person, object, performer, living thing, entity, physical entity
```

27. [2 points] Given a set of relevant documents D_R and non-relevant documents D_{NR}, the Rocchio algorithm suggests finding a query that maximizes the differences between the documents. Does the algorithm compute the differences between
A. All of the D_R and D_{NR} documents
B. Just the closest document in D_R to the closest document in D_{NR}
C. the centroid of D_R and the centroid of D_{NR}

28. [3 points] Given two classes of documents, C1 and C2, where each document is described by a vector of dimension 2:
C1: (1,6), (2,7), (3,8)
C2: (9,7), (11,9), (8, 10),
What are the centroids of C1 and C2

29. [3 points] Name two other ways to compute distance for normalized vectors other than Euclidean distance:

30. [3 points] Voroni regions are:
A. Circles
B. Squares
C. Rectangles
D. Polygons

	BRAHMA	HIGASHI	MANGO	FORNAIO	ZAO	MING	RAMONA	STRAITS	HOMMA
ALICE		1	-1	1				-1	
BOB		1				-1		-1	
CINDY				1	-1			-1	
DAVE	-1			-1	1	1			1
ESTIE				-1	1	1		1	
FRED	-1						-1		

31. [3 points] Above is a utility matrix for people and restaurants. what is the similarity between
A. Dave and Estie

B. Dave and Cindy

32. [3 points] Name three criteria that Google checks for when doing image similarity match.

33. [3 points] Search engines provide for image matching. One technique used is to examine in detail the color histogram of the query image and match that up with images in the search engine database. Name two other techniques that are used to match images?

34. [3 points] In order to create ImageNet, millions of pictures were categorizes according to WordNet synsets. In one sentence, how was this accomplished?