Computer Science 572 Exam
Prof. Horowitz
Monday, April 26, 2021, 7:00pm – 8:30pm

Name: **Student Id Number:**

INSTRUCTIONS

1. **This is an open book exam. The exam is intentionally long!**
2. **Please answer all questions.**
3. **Question points vary. Longer questions are towards the end of the exam paper.**
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.**

1. [2 points]. Suppose x and y have initial values and there are two threads of computation as shown.

x = 0; y = 1;
Thread 1: void foo() { x = x + 1; y = x + y; }

Thread 2: void bar() { y = y + 1; x = x + y; }
Provide two sequences of execution that produce different final values for x and y.

1) x=x+1 , y = y+1, x=x+y, y=x+y
2) x=x+1 ,y=x+y, y = y+1, x=x+y

T1 S1 : x = 2
T1 S2 : y = 2
T2 S1 : y = 3
T2 S2 : x = 5
X = 5, y = 3

T2 S1 : y = 1
T2 S2 : x = 2
T1 S1 : x = 3
T1 S2 : y = 4
X = 3, y = 4

2. [2 points]. As a website grows and adds more pages with more links to web pages outside of the website, how is the total PageRank of the website affected?

NOT IN PORTION

3. [2 points]. The HITS Algorithm developed by Jon Kleinberg identifies two types of web pages that have special significance. Name them and in one sentence define them.

NOT IN PORTION

4. [2 points]. Several heuristic techniques were presented for speeding up the computation of the ranked results. Mention two of them:

1. Consider Only Query Terms with High-idf Scores
2. Consider Only Docs Containing Several Query Terms
3. Introduce Champion Lists Heuristic
4. Introduce an Authority Measure
5. Reorganize the Inverted List
6. High and Low Lists Heuristic

5. [2 points]. It is well-known that online retail sites such as Amazon are able to offer (sell) far more books than a bookseller who operates a brick-and-mortar store. What is the phenomenon that describes this fact and what is the surprising (or unusual) result ?

Repeated from Fall '21.

6. [2 points]. True or False, in a Question/Answering system, an NER will take a sentence and return its parts of speech including nouns, verbs, adjectives and adverbs?

TRUE

Consider the names of two countries, Norway and France and their capital cities Oslo and Paris respectively; below is a set of five terms from WordNet, which are not necessarily in their correct order, that is used to create a tree classification of the two countries. Answer the questions below:

Region
Country
District
Location
Entity

7. [2 points]. What is the root of the tree?

Repeated from Fall '21.

8. [2 points]. The HasPart relation of WordNet could be applied to Norway and France. What leaf nodes might be attached to the HasPart relation?

Repeated from Fall '21.

9. [2 points]. Starting with the root of the tree, what is the correct order for the five terms above?

? -> ? -> ? -> ? -> ?

Repeated from Fall '21.

10. [2 points]. Suppose you set a maximum CPC of US$2.50. If 500 people see your ad and 20 of them click on your ad, how much money will you owe Google?
a. exactly $50.00
b. likely less than $50.00
c. likely more than $50.00

Repeated from Fall '21.

11. [2 points]. When Google must decide how to order the ads for a given query phrase, what formula does it use?

CTR * Bid Amount

12. [2 points]. If a Google advertiser bids on the phrase
`"figure skates"`
and uses Phrase Match, select the input queries below that would be a match for the above phrase

a. figure shoes
b. figure skates
c. figure skates for sale
d. figure nike skates
e. best figure skates

Repeated from Fall '21.

13. [2 points]. During the process of snippet generation there is an effort to identify the important (salient) words. But rather than using TF-IDF to compute the weight of a word another rule was suggested.
a. State the name of the rule, and
b. Give the formula for the rule

Repeated from Fall '21.

14. [2 points]. The acronym TLDR was used in class. In one sentence say exactly what TLDR stands for and explain what it means with respect to the SERP and the fold.

TOO LONG DIDN'T READ
Above the fold refers to a search engine results page ranking on the first page that is visible without having to scroll down.

15. [4 points]. a. Does schema.org define markup for a web page? b. Can schema.org definitions be used in email messages as well as web pages?

Repeated from Fall '21.

16. [5 points]. Each of the five items below (A-E) state possible properties of Google snippets. Indicate which properties are true and which are false about snippets.

Google snippets may

A. be as large as 300 characters long FALSE
B. include tables TRUE
C. include lists of items TRUE
D. include videos TRUE
E. Google may use the meta-description to create the snippet TRUE

17. [4 points]. Name two properties that are typically used to determine if a clustering algorithm works well.

Repeated from Fall '21.

18. [3 points]. What is the range of values for:
a. cosine similarity,
b. Pearson coefficient,
c. Jaccard Similarity

Repeated from Fall '21.

19. [3 points]. From our question/answering lecture what does the term wh-word mean?

Repeated from Fall '21.

20. [3 points]. How is the failure of a Map worker handled in the MapReduce framework?

The compute node of a Map worker fails:
– This is detected by the Master and all Map tasks that were assigned are re-done
– The Master sets the status of each Map task to idle and re-schedules them when a worker becomes available
– The Master informs each Reduce task of the location of its new input

21. [3 points]. In HW3, What are the two environment variables we set when we open a new SSH terminal on GCP.

Repeated from Fall '21.

22. [3 points]. In Solr what file contains the configuration for data dictionary?

solrconfig.xml

23. [3 points]. In Solr what file contains definitions of the field types and fields of a document?

managed-schema

24. [3 points]. Spelling correction programs make use of a "confusion matrix". Given an $n \times n$ confusion matrix, what values are represented by the rows and columns and what value is placed in the i^{th} row, j^{th} column?

each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or viceversa)

ith row, jth column- how many times jth column was confused for ith row

25. [2 points] In one sentence indicate how you made use of the provided URLToHTML_newsite_news.csv file while working on HW#4.

Repeated from Fall '21.

26. [2 points]. In HW#4, we created an external_PageRankFile that contains the Page Rank values for each of the crawled documents. Let us say that we forgot to include a document in the external_PageRankFile. What PageRank value would Solr assign to such a document not found in external_PageRankFile?

Repeated from Fall '21.

27. [2 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?

Repeated from Fall '21.

28. [2 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?

numbers denote which reducer wrote it out.

29. [2 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

Repeated from Fall '21.

30. [4 points]. Given two documents below, doc1 and doc2, provide the mapper output if an invertedIndex is run on the documents in a Hadoop cluster.

doc1 - USC Viterbi School of Engineering
doc2 - Andrew Viterbi invented the Viterbi algorithm

Mapper Output :
USC doc1
Viterbi doc1
School doc1
of doc1
Engineering doc1
Andrew doc2
Viterbi doc2
invented doc2
the doc2
Viterbi doc2
algorithm doc2

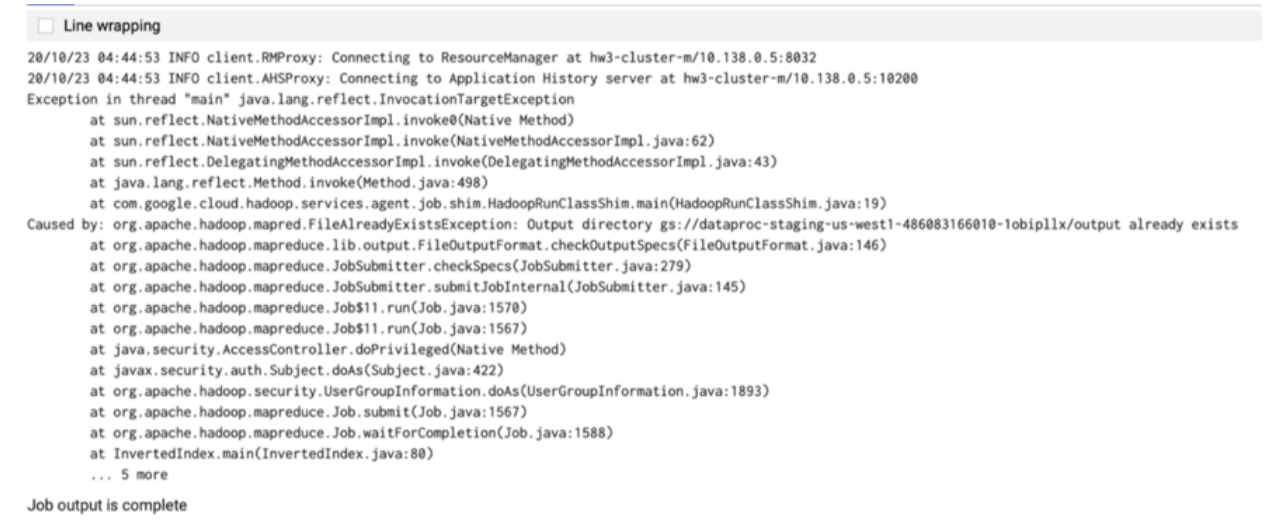
31. [4 points]. Given the two documents above, doc1 and doc2, provide the reducer output if an invertedIndex is run on the documents in a Hadoop cluster.

Reducer Output:

USC doc1:1
Viterbi doc1:1 doc2:1
School doc1:1
of doc1:1
Engineering doc1:1
Andrew doc2:1
invented doc2:1
the doc2:1
algorithm doc2:1

32. [2 points]. If you run a map reduce job in GCP and receive the following error log below.
a. Why does this exception occur?
b. How is this to be fixed?

a. output file already exists
b. either delete the file or rename it in the code



Line wrapping

```
20/10/23 04:44:53 INFO client.RMProxy: Connecting to ResourceManager at hw3-cluster-m/10.138.0.5:8032
20/10/23 04:44:53 INFO client.AHSProxy: Connecting to Application History server at hw3-cluster-m/10.138.0.5:10200
Exception in thread "main" java.lang.reflect.InvocationTargetException
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at com.google.cloud.hadoop.services.agent.job.shim.HadoopRunClassShim.main(HadoopRunClassShim.java:19)
Caused by: org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory gs://dataproc-staging-us-west1-486083166010-1obipllx/output already exists
        at org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:146)
        at org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:279)
        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:145)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1570)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1567)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1893)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1567)
        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1588)
        at InvertedIndex.main(InvertedIndex.java:80)
        ... 5 more

Job output is complete
```

33. [6 points]. Below is a set of code from your homework #4 where certain lines have been removed. Removed lines are numbered ①, ②, ③, ④, ⑤, ⑥.

Fill in the missing code. (This question counts for XXX points)

Note: All numbered areas take a single statement only. Do not concern yourself with the completeness of the code, just fill in with the most suitable code in the given context.

```
Class WordCountMapper extends _①_ <LongWritable, Text, Text,
IntWritable>
{
```

```
  private final static IntWritable one = new IntWritable (1);
  private Text word = new Text ();

  public void map (LongWritable key, Text value, Context context)
            throws IOException, InterruptedException
     {
            //Reading input one line at a time and tokenizing

            String line = value.toString ();

             ___②___ (create tokenizer object from string above)

           //iterating through all the tokens available,

            while(___③___)
            {
            //NO CODE REQUIRED HERE
            }
Class WordCountReducer extends __④__ <Text, IntWritable, Text,
IntWritable>
{

 public void __⑤__ (Text key, Iterable<IntWritable> values, Context
context)
            throws IOException, InterruptedException
      {
            int sum = 0;

            //Iterates through all the available values with a key

            for (IntWritable value: values)
            {
                  sum += ___⑥___;   // Get the value from object
            }
           context.write(key, new IntWritable(sum));
} }
```

Repeated from Fall '21.

34. [10 points]. The Levenshtein Edit Distance Algorithm can be defined as follows:

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Where

- ‘a’ stands for string1,

- ‘b’ stands for string2,
- ‘i’ is the terminal character position of string1
- ‘j’ is the terminal character position of string2
- ‘a_i’ refers to the character of string a at position i
- ‘b_j’ refers to the character of string b at position j
- lev $_{a,b}(i,j)$ is the distance between the first i characters of a and the first j characters of b

Given the two strings “HONDA” and “HYUNDAI”, use the Levenshtein Algorithm to show how it fills up the following table and computes the minimum edit distance between those two terms.

	#	H	Y	U	N	D	A	I
#								
H								
O								
N								
D								
A								

https://www.let.rug.nl/~kleiweg/lev/

35. [2 points] What is the minimum Levenshtein distance between the two strings mentioned above ?

3 https://planetcalc.com/1721/