Computer Science 572 Exam
Prof. Horowitz
Tuesday, November 30, 2021, 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 40 questions, each is worth 2 ½ points.**
3. **Some questions are short, but some are long.**
4. **I DON'T EXPECT YOU TO BE ABLE TO ANSWER ALL OF THE QUESTIONS IN THE TIME ALLOTTED**
5. **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.**
6. ***WARNING: Uploading your answers at the last minute will likely fail due to server overload UPLOAD EARLY!***

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

```
Hello world
Hello this world
```

1. [2 ½ points] What are the input key-value pairs that are sent to the Mapper?
Hint : You may refer to TextInputFormat subclass of FileInputFormat (an implementation of InputFormat interface in Java)

Key:Value Pairs: (0:Hello world), (1:Hello this world)

2. [2 ½ points] What are the two environment variables that get set when we open a new SSH terminal on GCP?

export PATH=${JAVA_HOME}/bin:${PATH}
export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar

3. [2 ½ 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>
```

Provide one sentence that could be the root cause of the problem? Assume the student has defined his Mapper & Reducer classes inside the Unigram class

The reason this is happening is that the user-defined Mapper and Reducer classes have to be defined as static classes; because they are not, we are facing this error.

4. [2 ½ points] Is an implementation of the Google File System part of Hadoop, YES or NO?

NO

5. [2 ½ points] Name two things about GFS that are different from a conventional UNIX file system

1. Lacks typical per-directory data structure to list each file in the directory
   Does not support aliases (i.e. hard or sym links)
2. Namespace: lookup table that maps full pathnames to metadata
   Lookup table fits in memory (prefix compression)

6. [2 ½ points] Do the Google workstations that run the GFS make use of the Linux File System as well, YES or NO?

YES

7. [2 ½ points] An Ad Network sits between an advertiser and a publisher. What is the name of Google's Ad Network?

Google Ads

8. [2 ½ points] How is an Ad Exchange different from an Ad Network? In one sentence define an Ad Network and in another sentence an Ad Exchange, emphasizing the differences.

An *ad network* . . . The key function of an ad network is aggregation of ad space supply from publishers and matching it with advertiser demand.

An *ad exchange* . . .

An ad exchange is a technology platform that facilitates the buying and selling of media advertising inventory from "multiple ad" networks and the prices of that inventory are determined through technology-driven bidding.

9. [2 ½ 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.

A. What is the phenomenon that describes this fact Anatomy of the Long Tail

B. what is the surprising (or unusual) result.

In some cases items sold from the long tail, (i.e. those not particularly popular) can cumulatively outweigh the initial portion of the graph, an in effect produce the majority of sales.

10. [2 ½ 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

Answer:

11. [2 ½ points] If a Google advertiser bids on the phrase
"figure skates"
and uses Phrase Match, select ALL 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

Answer(s):

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.

A.
X barks - A
X eats alpo - B
X is Brown - C
X is a dog - D

A ^ B -> D
D -> C

P- Cold
Q- Tastes sweet
R- Chocolate
S- Ice Cream

P ^ Q -> S
S -> R

12. [2 ½ points] This question has two parts:
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

A. A) A rule of inference used to draw logical conclusions, which states that if p is true, and if p implies q (pq), then q is true. B) From fact#1, Rover barks and eats Alpo, so Rover is a dog (this is fact #5). From #3, since Rover is a dog, Rover is brown (#6 is formed) which is what we want

B.

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
-> According to our modus ponens rule A and B both are true, therefor D is true. And if D is true, by the forward chaining we can con say that C is also true.
((A ^ B -> D) and A ^ B) -> D
(D->C and D) -> C

13. [2 ½ points] Name one way you made use of the provided URLToHTML_newsite_news.csv file while working on HW#4?

1. If URL is missing, fetch the url from the provided mapping file.
2. To generate the EdgeList file of PageRank, we parse each file to extract urls and find the corresponding file name in the provided mapping file.

14. [2 ½ points] In HW#4, you created an external_PageRankFile that contains the Page Rank values for each of the crawled documents. Let us say that you failed to include a document in the

external_PageRankFile. What PageRank value would Solr assign to such a document not found in external_PageRankFile?

0

15. [2 ½ points] What element did you add to Solr's configuration file(s) in HW#4 so that it can access the external_PageRankFile whenever the index is reloaded?

solrconfig.xml:
<listener event="newSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader" />
<listener event="firstSearcher" class="org.apache.solr.schema.ExternalFileFieldReloader" />

managed-schema:
<fieldType name="external" keyField="id" defVal="0" class="solr.ExternalFileField"/>
<field name="pageRankFile" type="external" stored="false" indexed="false"/>

16. [2 ½ points] In HW#4, you made sure that Solr always queries from a particular field by default. Which field did you choose to be the default for searching?

Answer: <str name="df">_text_</str>

17. [2 ½ points] What are the 3 types of elements extracted from the html files using jsoup in HW#4?

Answer: below is in the format of element - CSS selector:

1. media - [src]
2. links - a[href]
3. imports - link[href]

18. [2 ½ points] What are the five operations supported by Solr?

Answer: query, index, delete, commit, optimize

19. [2 ½ points] Mention any 3 fields present in the response json returned for each indexed file in the search results from Solr but not displayed on the webpage for HW#4.

Answer:

1.
2.
3.

['article_published_time', 'twitter_card', 'stream_content_type', 'og_site_name', 'description', 'twitter_creator', 'twitter_image', 'twitter_image_alt', 'twitter_site', 'dc_title', 'content_encoding', 'article_content_tier', 'content_type', 'stream_size', 'x_parsed_by', 'og_type', 'article_section', 'twitter_title', 'fb_profile_id', 'og_title', 'resourcename', 'fb_pages', 'article_author', 'fb_app_id', 'viewport', 'twitter_description', 'brightspot_contentid', 'content_language', 'article_opinion', 'version']

20. [2 ½ 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.

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));
} }
```

Answer:

1. 1 - Mapper
2. 2 - StringTokenizer tokenizer = new StringTokenizer(line);
3. 3 - tokenizer.hasMoreTokens();
4. 4 - Reducer
5. 5 - reduce
6. 6 - value.get()

21. [2 ½ points] Spelling detection and correction always requires a dictionary; name two other techniques for doing spelling error *correction*

Answer: 1. edit distance algorithms or
2. n-gram matching to come up with the correction

22. [2 ½ points] Writing “their” when you mean “they're” is what kind of spelling error?

Answer: cognitive error

23. [2 ½ points] In spelling correction an algorithm was given for handling n-grams. What is its name?

Answer: The Stupid Back-off Algorithm

24. [2 ½ points] What is the range of values for:
a. cosine similarity,
b. Pearson coefficient,
c. Jaccard Similarity

Answer:
a. [0,1]
b. [-1,1]
c. [0,1] , in terms of %: 0 to 100%

25. [2 ½ points] From our question/answering lecture the term wh-word refers to four items. What are they?

Answer: who, what, when, where

26. [2 ½ points] Snippets are computed at indexing time, TRUE or FALSE?

Answer: FALSE, at query time

27. [2 ½ points] A single snippet is computed for each web page, TRUE or FALSE?

Answer: FALSE

28. [2 ½ points] How does a featured snippet significantly differ from a snippet

Answer: Featured snippets are Google's attempt to answer the query right on the search results page. It was introduced in 2016. Google wants to give the user an immediate answer so they don't have to search the actual results. Featured snippets show up above the #1 ranked spot, and typically appear above the fold.The difference between Featured snippet and a snippet is that the former is aimed to solving a user query. They present fact/information from the webiste

29. [2 ½ points] Snippets exclusively are extracted from the web page body, TRUE or FALSE?

Answer: FALSE

30. [2 ½ points] In generating a snippet, one Google factor is the distance from the start of the web page text, TRUE or FALSE?

Answer: TRUE

31. [2 ½ points] Rich snippets differ from snippets in what significant way?

Answer: Rich Snippets give users a convenient summary information about their search results at a glance.
The mechanism calls for embedding structured data in web pages with the objective of displaying the structured data to a user in a visually outstanding way.
In snippets, we shows important information about a page in a concise manner.

32. [2 ½ points] What does PAA in Google's search engine results stand for?

Answer: People Also Ask

33. [2 ½ 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

Answer:

A. a. Topic signature based content selection

B. b. Choose words that are informative by Log-likelihood ration(LLR) or by appearing tin the query
weight(wi) = 1 if -2log λ (wi) > 10
1 if wi is in question
0 otherwise

34. [2 ½ points] This question has two parts:
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?

Answers: A. YES

B. YES

35. [2 ½ points] Name two properties that are typically used to determine if a clustering algorithm works well.

Answer:

1. 1. intra-class similarity - should be high
2. inter-class similarity - should be low

2. 3. Inertia
4. Dunn Index

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

36. [2 ½ points] What is the root of the tree?

Answer: Entity

37. [2 ½ points] The HasPart relation of WordNet could be applied to Norway and France. What leaf nodes might be attached to the HasPart relation?

Answer: Norway -> hasPart ->Oslo, France -> hasPart ->Paris

38. [2 ½ points] Starting with the root of the tree, what is the correct order for the five terms above?

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

Answer: Entity -> Location->Region->District-> Country

	Brahma Bull	Higashi West	Mango	Il Fornaio	Zao	Ming's	Ramona's	Straits	Homma's
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		

39. [2 ½ points] Given the utility matrix above, which pair of individuals have the greatest similarity? The least similarity?

Answer:
A. Greatest similarity: Dave & Estie
B. Least similarity: Cindi & Estie

40. [2 ½ points] Given a list of people and their opinions about different restaurants, if you develop a system that computes for all rated restaurants the number of positive minus the number of negatives reviews and recommends the restaurant with the highest number, is this an example of a content-based system or a collaborative-based system?

Answer: Collaborative Based System