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

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

USC, doc1),
Viterbi, doc1),
School, doc1),
of, doc1),
Engineering, doc1),
Andrew, doc2),
Viterbi, doc2),
invented, d2
invented, doc2),
the, doc2
the, doc2),
Viterbi, doc2
Viterbi, doc2),
algorithm, doc2
(algorithm, doc2)

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 x 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

USC, doc1),
Viterbi, doc1),
School, doc1),
of, doc1),
Engineering, doc1),
Andrew, doc2),
Viterbi, doc
invented, d2
the, doc2
Viterbi, doc2
algorithm, doc2

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

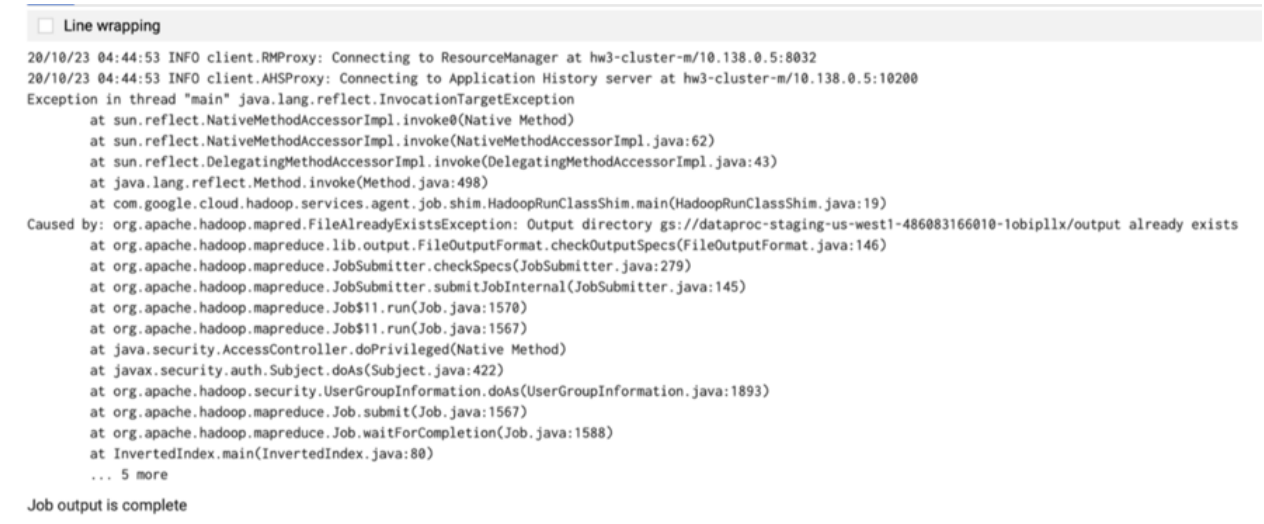
31. [4 points]. Given the two documents above, doc1 and doc2, provide the reducer output if an invertedIndex is run on the documents in a Hadoop cluster.

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/

Computer Science 572 Exam
Prof. Horowitz
Monday, April 22, 2019, 8:00am – 9:00am

Name: **Student Id Number:**

1. This is a closed book exam.
2. Please answer all questions.
3. There are a total of 40 questions.
4. **Place your answer immediately below the question. Limit answers to ONE SENTENCE unless more is requested.**

```
Mapper
class WordCountMapper extends Mapper<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
{
       String line = value.toString();
       StringTokenizer tokenizer = new StringTokenizer(line);
       while (tokenizer.hasMoreTokens())
       {
         XXXXXXXXXXXXXXXXXXXXX   word.set(tokenizer.nextToken());
         XXXXXXXXXXXXXXXXXXXXX   context.write(word, one);
       }
     }
}
```

```
Reducer

class WordCountReducer extends Reducer<Text, IntWritable, Text, IntWritable>
{
  public void reduce(Text key, Iterable<IntWritable> values, Context context)
       throws IOException, InterruptedException
  {
       int sum = 0;
       for (IntWritable value : values)
       {
         XXXXXXXXXXXXXXXXXXXXX   sum += value.get();
       }
       context.write(key, new IntWritable(sum));
  }
}
```

1. [2 1/2 pts] Above is the mapper and reducer classes for a WordCount program to be run on Hadoop. Two of the lines in Mapper are missing, denoted by XXXXXXXXXXXXXXXXXXXXX. Provide the missing lines.

2. [2 1/2 pts] Above is the mapper and reducer classes for a WordCount program to be run on Hadoop. One of the lines in Reducer is missing, denoted by XXXXXXXXXXXXXXXXXXXXX. Provide the missing line.

3. [2 1/2 pts] 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 :

    Repeated from Spring '21.

4. [2 1/2 pts] 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:

    Repeated from Spring '21.

5. [2 1/2 pts] Suppose x and y have initial values and there are two threads of computation as shown.

    x = 1; y = 0;
    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. For example you might start with: ***execute thread 2, first statement***

    Repeated from Spring '21.

6. [2 1/2 pts] Google has two programs for advertisers, one that places ads next to search engine results and one that places ads on a website. Name both of the programs.

   Google Ads, Google Adsense

7. [2 1/2 pts] Two improper techniques used to enhance a web page's ranking in search results are cloaking and page jacking. Using one sentence each, define them both.

   When the web server detects a request from a crawler, it returns a different page than the page it returns from a user request, the page is mistakenly indexed - CLOAKING
   Pagejacking is the process of illegally copying legitimate website content (usually, in the form of source code) to another website designed to replicate the original website. A pagejacker's intention is to illegally direct traffic from the original site to cloned Web pages.

8. [2 1/2 pts] Suppose one advertiser bids $1.00 for his ad to be displayed and a second advertiser bids $0.50 for his ad to be displayed and all other factors affecting ads are identical. If the first advertiser's ad is clicked on, how much does he pay Google?

   $0.51

9. [2 1/2 pts] Briefly describe the difference between a broad match and an exact match in the context of AdWords.

   Broad Match- Ads may also show for expanded matches, including synonyms and plurals.
   Broad matches are often less targeted than exact.

   Exact Match- The search query must exactly match your keyword

10. [2 1/2 pts] When viewed as a graph, a knowledge graph is what sort of graph? Use conventional graph terms. We are expecting at least two graph properties.

    directed labeled multigraph, where the nodes are entities and the edges relations

    Properties-
    1. A multigraph is a graph which is permitted to have multiple edges that have the same end nodes.
    2. Two vertices may be connected by more than one edge

11. [2 ½ pts] Given the statements P and Q, what is the modus ponens rule?

    "P implies Q. P is true. Therefore Q must also be true."

12. [2 1/2 pts] An ontology supports classes and subclasses. Is WordNet an ontology, yes or no?

    YES

Synset: In WordNet, similar words are grouped into a set known as a Synset (short for Synonym-set).

Hypernyms: a word or phrase that is a more general than the given word.(A broad or superordinate)

Hyponyms: a word or phrase that is a more specific than the given word.(More specific)

Meronym: A meronym denotes a constituent part of, or a member of something

13. [2 1/2 pts] WordNet uses the following terms: synset, hypernym, hyponym and meronym. In one sentence define one of them.

14. [2 1/2 pts] How are Wikipedia, WikiData and WikiMedia related. Using one sentence for each, describe each one.

Wikipedia: a multilingual open online encyclopedia written and maintained by a community of volunteers

WikiData: a sister project of WIkipedia and is an effort to convert the Wikipedia data into a knowledgebase

WikiMedia: the foundation/company which has project Wikipedia, WikiData

15. [2 1/2 pts] Several heuristic techniques were presented for speeding up the computation of the ranked results. Mention two of them.

Repeated from Spring '21.

16. [2 1/2 pts] Write out the 3-grams for the phrase below: (ignore the quotes)
"Fourscore and seven years ago our fathers brought forth a nation"
How many 3-grams are there?

Fourscore and seven, and seven years, seven years ago, years ago are, ago are fathers, are fathers brought, fathers brought forth, brought forth a, forth a nation

9

17. [2 1/2 pts] Lucene builds an inverted index from documents it parses. Is the inverted index positional?

YES

18. [2 1/2 pts] Is the NetworkX graph used in the PageRank algorithm directed or undirected?

Directed

19. [2 1/2 pts] There are 6 different query types supported by Solr. Mention any four of them.

1. Single and multi-term queries
2. +, -, AND, OR, NOT operators are supported
3. Range queries on date or numeric fields,
4. Boost queries:
5. Fuzzy search : is a search for words that are similar in spelling
6. Proximity Search : with a sloppy phrase query. The closer together the two terms appear, higher the score.

20. [2 1/2 pts] Given two strings, one of length m and the other of length n, what is the computing time of the Levenshtein algorithm when applied to these two strings?

O(mn)

21. [2 1/2 pts] In the Levenshtein algorithm, given two strings $X[1 . . m]$ and $Y[1 . . n]$ what is the definition of $D(i, j)$, the Levenshtein distance function in terms of X and Y?

Given two strings: X of length n and Y of length m, define d(i, j) as
– the minimum edit distance between X[1..i] and Y[1..j]
• i.e. the first i characters of X and the first j characters of Y
– Then the minimum edit distance between X and Y is thus D(n,m)

d(i , j) = min { d(i-1,j) + 1, d(i, j-1) + 1, d(i-1, j-1) + 1 if chars not equal, 0 otherwise}

22. [2 1/2 pts] Given the assumptions of the previous question what are the values of *D(i, 0)* for i = 1 , . . , m and what are the values of *D(0 , j)* for j = 1 , . . , n?

D(i,0) = i & D(0,j) = j

23. [2 1/2 pts] Given the two strings: "SIMPLIFY" and "AMPLIFIES", what is their minimum edit distance assuming the operations (replace, delete, insert) all have a count of 1?

5

24. [2 1/2 pts] Which HTML tag field is used by Google as the default for creating a snippet?

<meta> Meta-description

25. [2 1/2 pts] There are two special types of snippets used by Google. What is their names?

Featured snippets, Rich snippets

26. [2 1/2 pts] The schema.org website defines a technology that is used by Google, Yahoo and Bing. In one or two sentences what is that technology?

Rich snippet technology:Two other formalisms for creating rich snippets have been suggested:

• RDFa (Resource Description Framework – in Attributes) http://en.wikipedia.org/wiki/RDFa
• Microformat Encoding http://en.wikipedia.org/wiki/Microformat

27. [2 1/2 pts] Define breadcrumbs.

A breadcrumb trail on a page indicates the page's position in the site hierarchy. A user can navigate all the way up in the site hierarchy, one level at a time, by starting from the last breadcrumb in the breadcrumb trail.

28. [2 1/2 pts] To implement rich snippets two technologies are offered, microformats and microdata. In one sentence how does microformat work and give a one line example?

Microformats use only existing HTML, e.g. the class attribute in HTML tags (often or) to assign brief and descriptive names to entities and their properties.

29. [2 1/2 pts] Define: Dendrogram

A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.

30. [2 1/2 pts] What is the difference between hard clustering and software clustering?

Hard clustering: Each document belongs to exactly one cluster.
Soft clustering: A document can belong to more than one cluster.

31. [2 1/2 pts] Mention one possible criterion for determining when the k-means algorithm can terminate.

For termination conditions there are several possibilities, e.g.,
– After a fixed number of iterations
– When the document partition is unchanged
– When the centroid positions don't change

32. [2 1/2 pts] Is the Agglomerative Clustering Algorithm top-down or bottom up?

bottom up

33. [2 1/2 pts] Is the Divisive Clustering algorithm top-down, bottom-up, or both?

top-down

34. [2 1/2 pts] For the k-means algorithm, if *M* is the size of a document vector, *N* is the number of vectors, *K* is the number of clusters, and *I* is the number of iterations, what is the worst-case computing time of the algorithm?

O(IKMN)

35. [2 1/2 pts] What set of points does K-means clustering use to identify a cluster?

centroids: mean of all the points in a cluster

36. [2 1/2 pts] The k-means++ algorithm uses a different method than the k-means algorithm for choosing the initial clusters. What is that method?

Pick the most distant (from each other) points as cluster centers.

37. [2 1/2 pts] The mean reciprocal rank is a statistical measure for evaluating a process that produces a list of responses to a query. If |Q| represents the number of queries and rank(i) represents the rank of the correct result for the ith query, then define the Mean Reciprocal Rank or MRR

Mean Reciprocal Rank:
– For each query return a ranked list of M candidate answers
– Its score is 1/Rank of the first right answer (or 0 if no answers are correct)\
– Take the mean over all |Q| queries

MRR =Σ(1/ranki) / |Q|

38. In determining an answer to a question, it was suggested that n-grams be used. What is the definition of the weight of an n-gram?

occurrence count, each weighted by "reliability"(weight) of rewrite that fetched the document

39. [2 1/2 pts] We looked at two algorithms for classifying documents into groups. What are they called?

KNN, Rocchio

40. [2 1/2 pts] In one sentence define the contiguity hypothesis

Documents in the same class form a contiguous region of space.

Computer Science 572 Exam
Prof. Horowitz
Monday, November 26, 2018, 8:00am – 8:50am

Name: **Student Id Number:**

1. This is a closed book exam.
2. Please answer all questions.
3. There are a total of 30 questions. Question points may vary.
4. **Place your answer immediately below the question. Limit answers to ONE SENTENCE unless more is requested.**

1. [3 pts] Given the following two threads and initial values for x = 6 and y = 0, what are two different possible ending values for x and y? For each possible ending values provide the order of execution that supports your results.

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

Thread 2
```
void bar( ) {
  y++;
  x++;
}
```

Thread 2,1 = 8,8 &
Thread 1,2,1,2 = 8,7

2. [3 pts] Does correlation necessarily imply causation. Yes or No?

NO

3. [3 pts] In the Google cloud, once your cluster has been set up what command is used to connect to the master machine?

SSH

4. [3 pts] Suppose one advertiser bids $1.50 for his ad to be displayed and a second advertiser bids $1.25 for his ad to be displayed and all other factors affecting ads are identical. If the first advertiser's ad is clicked on how much does the advertiser pay Google?

$1.26

5. [3 pts] Given two documents below, doc1 and doc2, provide the mapper output if an invertedIndex is run on the documents in a Hadoop cluster.

   doc1 – seven years ago our fathers did
   doc2 – three years ago our sisters did

   seven doc1
   years doc1
   ago doc1
   our doc1
   fathers doc1
   did doc1
   three doc2
   years doc2
   ago doc2
   our doc2
   sisters doc2
   did doc2

6. [3 pts] Given the two documents above, doc1 and doc2, provide the reducer output if an invertedIndex is run on the documents in a Hadoop cluster.

   seven doc1:1
   years doc1:1 doc2:1
   ago doc1:1 doc2:1
   our doc1:1 doc2:1
   fathers doc1:1
   did doc1:1 doc2:1
   three doc2:1
   sisters doc2:1

7. [3 pts] The class notes describe long-tailed keywords as search queries made up of three, four or more word phrases. Is the conversion rate for long-tailed keywords better or worse that it is for shorter keywords?

   better, 2.5 times than shorter

8. [3 pts] What is the name of Google's knowledgebase?

   Google's Knowledge Grpah

9. [3 pts] When viewed as a graph, an ontology is what sort of graph? Use conventional graph terms. We are expecting at least three graph properties.

labeled, directed & cyclic

10. [4 pts] The discussion of search engine optimization (SEO) identified many factors which correlate strongly with attaining a high ranking on the search engine result page. Mention four of them.

1. Content factors • Content relevance • Word count
2. User signals • Click Through Rate (CTR) • Bounce rate
3. Technical factors • Presence of H1/H2 • Use of HTTPS
4. User experience • Number of internal/external links
5. Social signals • Facebook total • Tweets
6. Backlinks- No. of backlinks, No.of DoFollow backlinks

11. [3 pts] Define Mean Reciprocal Rank scoring

Repeated from Spring '19.

Below is the Norvig spelling corrector program written in Python and presented in class. Please answer the questions that follow the program.

```
import re, collections
def words(text): return re.findall('[a-z]+', text.lower())
def train(features):
    model = collections.defaultdict(lambda: 1)
    for f in features:
        model[f] += 1
    return model
NWORDS = train(words(file('big.txt').read()))
alphabet = 'abcdefghijklmnopqrstuvwxyz'
def edits1(word):
   splits     = [(word[:i], word[i:]) for i in range(len(word) + 1)]
   deletes    = [a + b[1:] for a, b in splits if b]
   transposes = [a + b[1] + b[0] + b[2:] for a, b in splits if
len(b)>1]
   replaces   = [a + c + b[1:] for a, b in splits for c in alphabet if
b]
   inserts    = [a + c + b     for a, b in splits for c in alphabet]
   return set(deletes + transposes + replaces + inserts)
def known_edits2(word):
    return set(e2 for e1 in edits1(word) for e2 in edits1(e1) if e2 in
NWORDS)
def known(words): return set(w for w in words if w in NWORDS)
def correct(word):
    candidates = known([word]) or known(edits1(word)) or
known_edits2(word) or [word]
    return max(candidates, key=NWORDS.get)
```

12. [3 pts] What functions are defined?

words(text)
train(features)
edits1(word)
known_edits2
known
correct

13. [3 pts] What function is used to invoke (start) the program

words

14. [3 pts] What edit operations are included in the program?

inserts
deletes
transposes
replaces

15. [3 pts] How many levels of edits does the program investigate?

2

16. [3 pts] For a clustering to be considered good (or successful) what is the similarity property to be satisfied for its intra-class elements and inter-class elements?

intra class-high
inter class-low

17. [3 pts] Define soft clustering

Repeated from Spring '19.

18. [4 pts] In the k-means clustering algorithm there are several possible criteria for termination. mention two.

Repeated from Spring '19.

19. [4 pts] Given n documents each expressed as an m-element vector, what is the computing time for the Agglomerative Clustering algorithm assuming priority queues are used?

O(mnlogn)

20. [4 pts] Given the two strings: “information” and “interrogation”, what is their minimum edit distance assuming the operations (replace, delete, insert) all have a count of *1*?

5

21. [3 pts] In class we discussed that there are three distinct phases for a question/answering system. Name them.

Question Processing
Passage Retrieval
Answer Processing

22. [4 pts] The terms hyperonymy and hyponymy are used with respect to WordNet. Define one of the two terms and give an example

Repeated from Spring '19.

Hyponym: plant -> tree (specialization)
Hypernym: apple -> fruit (generalization)

23. [3 pts] Given 75 documents divided into three clusters where cluster 1 has 10 related documents, cluster 2 has 5 related documents and cluster 3 has 10 related documents, what is the Purity Index of this clustering?

1/3, NOT IN PORTION

24. [3 pts] For the k-means algorithm, is the centroid necessarily a document in the set of documents?

NO

25. [3 pts] Is the graph provided to NetworkX directed or undirected?

directed

26. [4 pts] This semester we examined two algorithms for clustering and two algorithms for classification. Name all four.

Clustering-
K-means
Heirarchical

Classification-
KNN
Rocchio

27. [3 pts] There are three types of spelling errors. Non-word errors, Typographical errors and Cognitive errors. Define cognitive errors and give an example.

Cognitive errors (homophones, sounds alike)
– piece - peace,
– too - two,
– your - you're

28. [4 pts] There are six different strategies to speed up indexed retrieval mentioned in class. Mention any three of them.

Repeated from Spring '19.

29. [4 pts] When viewed as a graph, a knowledge graph is what sort of graph? Use conventional graph terms. We are expecting at least two graph properties.

Repeated from Spring '19.

30. [5 pts] Below is the equation for maximum likelihood estimation for bigram probabilities? Explain the terms in the equation.

$$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1},\ w_i)}{count(w_{i-1})}$$

w(i-1): (i-1)th word in the document
w(i): ith word in the document

P(w(i)|w(i-1)): Probability of (i)th word occuring given (i-1)th word has already occured
count(w(i-1),w(i)): number oftimes that (i-1)th word and (i)th word occur together in given order
count(w(i-1)): frequency of (i-1)th word in the document

Computer Science 572 Exam
Prof. Horowitz
Monday, April 23, 2018, 8:00am – 9:00am

Name: **Student Id Number:**

1. This is a closed book exam.
2. Please answer all questions.
3. There are a total of 25 questions. Question points may vary.
4. **Place your answer immediately below the question. Limit answers to ONE SENTENCE unless more is requested.**

1. [4 pts] Define the contiguity hypothesis.

   Documents in the same class form a contiguous region of space.
   KNN makes sense.

2. [4 pts] Below is a formula that occurs in the class notes.

   Question 1: What does the formula define? Question 2: Define C, |C| and $\vec{d}$

$$\vec{\mu}(C) = \frac{1}{|C|}\sum_{d \in C}\vec{d}$$

   The centroid is the center of mass (or vector average) of a set of points.
   C - set of documents
   |C| - size of the set
   d - normalized vector representing document

3. [4 pts] Given N as the number of documents, is the time to train the documents according to the Rocchio method *O(N), O(N log N) or O(N^2)* ?

   O(n^2)

Watson's DeepQA system for question answering has four phases:
(1) Question Processing, (2) Candidate Answer Generation, (3) Candidate Scoring, and (4)Answer Merging and Confidence Scoring. Answer the following two questions about Watson.

4. [4 pts] In what phase does Named Entity tagging occur?

   Data Processing, Merging and Confidence Scoring

5. [4 pts] Define co-reference and in what phase does it occur?

a relationship between two words or phrases in which both refer to the same person or thing and one stands as a linguistic antecedent of the other, as the two pronouns in She taught herself but not in She taught her.

6. [4 pts] In question answering, when several passages containing the query terms are returned, there are six criteria used to rank the passages. Please name two of them.

1. The number of named entities of the right type in the passage.
2. The number of question keywords in the passage.
3. The longest exact sequence of question keywords that occurs in the passage.
4. Rank of the document from which the passage was extracted.
5. Proximity of the keywords from the original query to eachother.
6. The N-gram overlap between passage and the question.

7. [4 pts] Client applications use five fundamental operations to work with Solr using HTTP requests and responses- Name any 2.

Query , Index , Delete , Commit and Optimize

8. [4 pts] Lucene uses a Boolean and Vector space model to determine how relevant a document is to a user's query. How does the Vector Space Model score the document?

Cosine Similarity

9. [4 pts] 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 Gould School of Law
doc2 – Sam Gould enacted the criminal statute

USC :1
Gould:1
School:1
of:1
Law:1
Sam:1
Gould:1
enacted:1
the:1
criminal:1
statute:1

10. [4 pts] Given the two documents in the above question, doc1 and doc2, provide the reducer output if an invertedIndex is run on the documents in a Hadoop cluster.

USC :1
Gould:2
School:1
of:1
Law:1
Sam:1
enacted:1
the:1
criminal:1
statute:1

11. [4 pts] When using N-grams for spelling correction, if no match is found for value N, then the algorithm will step back and look for a match with the (N-1)-grams, and again if there is no match the algorithm backs up again. What is this algorithm called?

Stupid Back-off Algorithm

12. [4 pts] Given 100 documents divided into three clusters where cluster 1 has 10 related documents, cluster 2 has 5 related documents and cluster 3 has 10 related documents, what is the Purity Index of this clustering?

PI = 25/100 = 0.25

13. [4 pts] For the k-means algorithm, is the centroid necessarily a document in the set of documents? Yes or No.

No

14. [4 pts] In Solr what file contains configuration for the data dictionary?

solrconfig.xml

15. [4 pts] In Solr what file contains definitions of the field types and fields of a document?

schema.xml

16. [4 pts] What is the syntax for starting and stopping Solr?

bin/solr start
bin/solr stop

17. [4 pts] There are three criteria that define a good clustering algorithm, describe one.

1. Unlikely to be altered drastically when further objects are incorporated.
2. Stable in the sense that small errors in the description of the objects leads to small changes in the clustering.
3. The method is independent of the initial ordering of the objects.

18. [4 pts] Given the two strings: "abcde" and "azced", what is their minimum edit distance assuming the operations (replace, delete, insert) all have a count of *1*?

3 is the minimum no of edits

19. [4 pts] Suppose one advertiser bids $1.00 for his ad to be displayed and a second advertiser bids $0.50 for his ad to be displayed and all other factors affecting ads are identical. If the first advertiser's ad is clicked on how much does he pay Google?

$0.51

20. [4 pts] We discussed clustering and classification. One is an example of supervised learning and the other is an example of unsupervised learning. Which one is supervised and which one is unsupervised?

Clustering - unsupervised
classification - supervised

21. [4 pts] Is the graph provided to NetworkX directed or undirected?

directed

22. [4 pts] This semester we examined two algorithms for clustering and two algorithms for classification. Name all four.

Clustering
K-Means
Hierarchical Clustering Algorithms

Classification
Rocchio
K Nearest Neighbour

23. [4 pts] Microsoft, Google and Yahoo agreed on a formalism for including rich snippets in web pages. What website contains the specification of this formalism? What is the name of the formalism?

Website:- schema.org
Name of Formalism:- microFormat

24. [4 pts] Define: breadcrumbs

a breadcrumb trail on a page indicates the page's position in the site hierarchy. A user can navigate all the way up in the site hierarchy, one level at a time, by starting from the last breadcrumb in the breadcrumb trail.
For example, Books › Authors › Ann Leckie › Ancillary Justice

25. [4 pts] In our discussion of knowledgebases we discussed the need for instances, classes and a taxonomic hierarchy. Wikipedia includes many instances. Does it also include classes and a taxonomic hierarchy?

classes - Yes
taxonomic hierarchy - No

Computer Science 572 Exam
Prof. Horowitz
Tuesday, November 29, 2016, 9:00pm – 10:30am

Name: **Student Id Number:**

1. This is a closed book exam.
2. Please answer all questions.
3. There are a total of 50 questions. Each question is worth 2 points.
4. **Place your answer immediately below the question. Limit answers to ONE SENTENCE.**

1. What is the second largest web search engine in terms of revenue?

Alibaba

2. This semester we have seen at least a dozen short videos by two professors from Stanford. Name one of them

Precision and Recall,Jurafsky and Manning

3. Google recently announced that it is indexing how many pages?

Now - 130 Trillion

4. True or False, Google retains a user's entire query history?

True

5. Which content type is NOT indexed by Google? (circle the best answer):
- swf
- xlsx
- rtf
- svg
- None of the above as all of them are indexed

swf

6. The definitions of Precision and Recall both have "the number of relevant items retrieved" as their numerator. What is the denominator for Precision?

Total number of retrieved documents

7. The definitions of Precision and Recall both have "the number of relevant items retrieved" as their numerator. What is the denominator for Recall?

Total number of relevant documents

8. In the formula for Discounted Cumulative Gain, how are documents appearing lower in a search result list penalized?

penalty -the graded relevance value is reduced logarithmically proportional to the position of the result.

9. In one sentence define “cloaking”

Cloaking is when the web server detects a request from a crawler, it returns a different page than the page it returns from a user request
– The page is mistakenly indexed

10. This question has two parts. A study of how to design a web page crawler to locate the best quality pages was done by Cho and Garcia-Molina. What measure of quality did they use? Secondly, what algorithm did they determine would produce the highest quality pages in the shortest time? measure - Avg freshness

Algoritmn - Uniform policy: re-visiting all pages in the collection with the same frequency, regardless of their rates of change.

11. True or False: Cho and Garcia-Molina showed that in order to optimize the freshness of a web-crawl, we should crawl pages as fast as possible.

False

12. A cryptographic hash function of a file has three main properties:
it is easy to compute
it is difficult to find a file that has the same hash value,
and what is the third property?

It is extremely computationally difficult to calculate an alphanumeric text that has a given hash.

13. Given two sets A and B, define the Jaccard similarity.

• JS(A, B) = size(A intersection B) / size(A union B)

14. In one sentence what is the English description of tf-idf?

higher the occurrence, lower the importance

15. In one sentence define Heaps Law.

Heap's law describes the number of distinct words (V) in a set of documents as a function of the doc length (n). If V is the size of the vocabulary and n is the number of words: V= K*n^B

16. State Zipf's Law

Zipfs law states that the frequency of any word is inversely proportional to its rank in the frequency table. An equation of the form y = kx^c is called a power law. Zipf's law is a power law with c = –1

17. In class you saw the Block Sort-Based Indexing algorithm. What was the algorithm attempting to minimize? minimize the number of random disk seeks during sorting

18. An inverted index is often split into two parts. Name them.

Index file and Postings List

19. Suppose there are only two web pages, each with only one link that points to the other web page. What will be the PageRank of each page?

Each will have a score of one as PR algorithm over large iterations will converge to one.

20. 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?

Decreases

21. The HITS Algorithm developed by Jon Kleinberg identifies two types of web pages that have special significance. What are these two types of web pages?

Hubs and Authorities

22. The HITS algorithm forms what type of graph when ranking pages?

bipartite graph

23. In the Google AdWords system what does CTR stand for?

Click-Through Rate

24. In one sentence what is Google AdSense?

AdSense from Google is a service for placing Google ads on third party web pages

25. What is a "tracking pixel"?

Tracking pixels are small, typically transparent images on a web page that have special names which permit the loading of the web page to be tracked by a web server.

26. When Google must decide how to order the ads for a given query phrase, what formula does it use?

Ad Rank= Bid X Click Probability

27. Suppose the Pepsi Cola company wants to bid on the words Coca Cola whenever they are entered as a query, so a Pepsi Cola ad will appear. Is this legal?

No

28. Briefly describe the difference between a broad match and an exact match in the context of AdWords.

-Broad Match ads may also show for expanded matches, including synonyms and plurals.Its matches are often less targeted than exact or phrase matches
-The search query must exactly match your keyword

29. Lucene/Solr uses two methods for ranking results. What are they?

Vector Space Model and the
Boolean model

30. Lucene builds an inverted index from documents it parses. Is the inverted index positional?

Yes

31. What is the Soundex Algorithm?

Soundex is a phonetic algorithm for indexing names by their sound when pronounced in English. The basic aim is for names with the same pronunciation to be encoded to the same string so that matching can occur despite minor differences in spelling

32. Google places an implicit Boolean operation between the terms of a query. What is it?

implicit AND

33. Define "edit distance".

the distance between two strings is the smallest number of insertions and deletions of single characters that will convert one string into the other

34. The LevenShtein algorithm assigns what weights to the operations of insertion, deletion, and substitution? +1, +1, +1/0

35. What data structure is very helpful when used to catch spelling errors as the user types?

a prefix tree (sometimes called a trie)

36. What is the difference between hard clustering and soft clustering?

Hard clustering: Each document belongs to exactly one cluster
Soft clustering: A document can belong to more than one cluster.

37. Given two vectors A = (A_1, . . . , A_N) and B = (B_1, . . . ,B_N) representing two documents, define their cosine similarity

A.B/(||A||.||B||)

38. The k-means++ algorithm uses a different method than the k-means algorithm for choosing the initial clusters. What is that method?

pick the most distant (from each other) points as cluster centers

39. Mention one possible criteria for determining when the k-means algorithm can terminate.

When the centroid positions don't change

40. Is the Aglomerative clustering algorithm top-down or bottom up?

bottom-up

41. What set of points does K-means clustering use to identify a cluster?

Minimize Sum Squared Error

42. WordNet uses the following terms: synset, hypernym, hyponym and meronym. In one sentence define one of them.

hypernym - broad or superordinate

43. In relation to capturing clicks on the search result pages, Google and Bing differ in what way?

Text

44. Name the four types of protection for intellectual property

copyright,patents,trademarks,trade secrets

45. Can a web page author claim his page is copyrighted if he forgets to insert a "Copyright" statement on the page?

Yes

46. We looked at two algorithms for classifying documents into groups. What are they called?

k-means and agglomerative hierarchical clustering

47. In one sentence define the contiguity hypothesis

Documents in the same class form a contiguous region of space

48. True or False, in vector space classification, it doesn't matter if one uses Euclidean distance or cosine similarity?

False

49. Which vector space classification method uses centroids to define the boundaries of regions?

k-means

50. Of the two vector classification algorithms discussed in class, which one decomposes the set of documents into Voroni cells?

k-nearest neighbors

Computer Science 572 Exam
Prof. Horowitz
Monday, April 25, 2016, 10:00pm – 11:45am

Name: **Student Id Number:**

1. This is a closed book exam.
2. Please answer all questions.
3. There are a total of 50 questions. Each question is worth 2 points.
4. **Place your answer immediately below or to the right of the question. Limit answers to ONE SENTENCE.**

1. what is the title of the textbook for this class? Introduction to Information Retrieval

2. Name one of the authors of the textbook Christopher D. Manning

3. AltaVista, Lycos and InfoSeek are (circle the best answer):
- apps for an iPhone
- early web search engines
- information retrieval systems

4. David Filo and Jerry Yang are (circle the best answer):
- founders of Google
- creators of spreadsheets
- founders of Yahoo

5. Google first appeared on the web in (circle the best answer):
- 1998
- 2004
- 2010

6. Google retains a user's entire query history? (circle the best answer):
- True
- False

7. What is the second largest web search engine in terms of revenue?

8. Approximately how many websites are there? (circle the best answer):
- 1 million
- 10 million
- 100 million
- 1 billion
- 10 billion
- 100 billion

9. Which content type is NOT indexed by Google? (circle the best answer):
- swf
- xlsx
- rtf
- svg
- None of the above as all of them are indexed

10. The definitions of Precision and Recall both have "the number of relevant items retrieved" as their numerator. What is the denominator for Precision?

TP+FP

11. The definitions of Precision and Recall both have "the number of relevant items retrieved" as their numerator. What is the denominator for Recall?

TP+FN

12. What information is contained in the robots.txt file?

What webpages cannot be crawled

13. In what directory does one find the robots.txt file on a website, if there is one?

Root directory

14. What is the better strategy for crawling a website? (circle the best answer):
- breadth first search
- depth first search

15. Name a cryptographic hashing method: MD5, SHA-1/SHA-2

16. A cryptographic hash function of file X has three main properties:
it is easy to compute
it is difficult to find a file that has the same hash value,
and what is the third property? A slight change in text yeilds totally different hash

17. Distance measures are defined by 4 properties:
1. no distances are negative
2. $d(x,y) = 0$ *iff* $x = y$
3. $d(x,y) = d(y,x)$
what is the fourth property? D(x,y) <= D(x,z) + D(z,y)

18. Given two sets A and B, define the Jaccard similarity. n(A intersection B) / n(A union B)

19. In one sentence what is the English description of tf-idf?

higher the occurence lower is the importance

20. crawler4j was downloaded from what online source repository?

crawler4j (4.4.0) from https://jar-download.com/artifacts/edu.uci.ics/crawler4j/4.4.0/source-code

21. what formula or law is expressed this way?

$\log(y) = \log(k*x^c)$ Power Law

22. Give three examples of stopwords:

a, the, an

23. In one sentence define Heaps Law.

Heap's law describes the number of distinct words (V) in a set of documents as a function of the document length (n)

24. Define token V = Kn^b with constants K, 0 < b <1

A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit

25. Define tokenization

The task of chopping a document unit into pieces, called tokens, and possibly throwing away certain characters

26. Stemming is a method for normalizing tokens. What is the name of a famous stemming algorithm?

The Porter Stemming Algorithm

27. A phonetic algorithm identifies words that sound the same but are spelled differently. Name one. Soundex

28. In class you saw the Block Sort-Based Indexing algorithm. What was the algorithm attempting to minimize? the number of random disk seeks during sorting

29. True or False, Google does NOT permit Boolean operators in queries? False

30. Given the query

a OR "b c" d

where a, b, c, and d represent keywords being searched for, fully parenthesize the query as Google would do it. Insert any implied Boolean operators.

((a OR “b c”) and d)

31. Describe in one sentence the search results that are returned by Google if you enter the query

`filetype: pdf`

restricts your results to files ending in a specific file suffix .pdf and shows you only files created with the corresponding program

32. In one sentence define cloaking.

When the web server detects a request from a crawler, it returns a different page than the page it returns from a user request. The page is mistakenly indexed

33. in the line

<meta name=robots content="noindex,follow,noarchive">

explain the meaning of noindex, follow and noarchive

The follow tag tells search engines to not ignore that link

Using noindex is useful if you don't have root access to your server, as it allows you to control access to your site on a page-by-page

Noarchive will prevent your content from being fully cached by Google.

34. In the Google AdWords system what does CTR stand for?

Click-Through Rate

35. In one sentence what is Google AdSense?

AdSense from Google is a service for placing Google ads on third party web pages

36. Lucene/Solr uses two methods for ranking results. What are they?
Vector space model, Boolean model

37. True or false, Solr includes explicit operators for AND, OR, NOT? True

38. True or False, Solr will return results in XML format? False

39. Several heuristic techniques were presented for speeding up the computation of search results. Mention two of them:
Introduce Champion Lists Heuristic, Maintain High and Low Lists Heuristic

40. Name the three types of spelling errors: Non-word, typographical, cognitive

41. What is a popular data structure for organizing a lexicon that is especially useful to implement autocomplete or prefix matching? Trie

42. An equivalent term for the word shingles is? Words

43. What is the difference between hard clustering and soft clustering?
Hard clustering - document belongs to only one cluster, Soft Clustering - One or more clus

44. In one sentence define: Dendrogram
A tree diagram used to illustrate arrangement of clusters produced by hierarchical clustering

45. 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:
Lycos, Alta Vista, Duck-Duck Go

46. Let True Positive (TP) be defined as an invalid click that is correctly identified as invalid; let True Negative (TN) be a valid click that is correctly identified as valid; let False Positive (FP) be a valid click that is incorrectly identified as invalid; and let False Negative (FN) be an invalid click that is incorrectly identified as valid. Using the above terms define: **accuracy rate** and **error rate**: Accuracy Rate = (TP+TN)/(TP+TN+FP+FN)
Error Rate = (FP+FN)/(TP+TN+FP+FN)

47. Given a set of N queries and AVGPrec(N) the average precision of each query, what is the formula for the Mean Average Precision? Summation i= 1 to N (AVGPrec(i)) / N

48. In the formula for Discounted Cumulative Gain, how are documents appearing lower in a search result list penalized? 1/log (rank)

49. True or False, Page rank is calculated for each search query and highly relevant pages will always have a higher page rank? False

50. We looked at two algorithms for classifying documents into groups. What are they called?
k-means Clustering, Agglomerative hierarchical clustering