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.