Computer Science 572 Exam
Prof. Horowitz
Monday, November 27, 2017, 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. Each question is worth 2 ½ points.
4. **Place your answer immediately below the question. Limit answers to ONE SENTENCE unless more is requested.**

1. What is the name of the Google system that permits users to place Google ads on their website?

2. If an advertiser has selected "Broad Match" for the keywords '**white house**' will the query '**house is white**' be matched.

3. What formula does Google use to determine the order of ads to be placed on a search result page?

4. If an advertiser bids $5 and his ad (AD1) has a click through rate of 0.05 and a second advertiser bids $10 and his ad (AD2) has a click through rate of 0.02, which ad is listed first? AD1 or AD2 ?

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

   doc1 – Fourscore and seven years ago
   doc2 – Double double toil and trouble

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

7. 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?

8. Name three types of spelling errors

9. 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?

10. 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*?

11. 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*?

12. Given 50 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?

13. Given three documents whose vector descriptions are: (7,2,3,10); (4,5,6,7); (7,2,6,4), what is the centroid of this cluster?

14. For the k-means algorithm, there are four parameters that affect the computing time. What are they??

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

16. There are three criteria that define a good clustering algorithm, describe one:

17. Give an example of top-down and bottom-up clustering algorithms respectively.

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

19. What are the methods/operations used by Levenshtein algorithm for generating the candidates during spelling correction?

20. Given the two strings: “kitten” and “sitting”, what is their minimum edit distance assuming all operations have a count of *1*?

21. Is a web page automatically copyrighted or must the publisher explicitly add the word "copyright" on the page?

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

23. Which process comes first: clustering or classification?

24. Which HTML tag field is used by Google as the default for snippet?

25. 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?

26. What type of graph is provided to NetworkX?

27. Lucene combines to formulas for ranking search results. What are those?

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

29. What is the default port of Apache Solr on localhost?

30. Name one way to choose a name for a cluster

31. Define the Purity Index for evaluating the result of a cluster algorithm.

32. If *TP* assigns two similar documents to the same cluster; *TN* assigns two dissimilar documents to different clusters; *FP* assigns two dissimilar documents to the same cluster and *FN* assigns two similar documents to different clusters, then how does one define the Rand Index?

33. What is the default response format in Solr (versions>=7.0.0), XML or JSON?

34. What is the library internally used by Tika to parse HTML?

35. Write a command to index only the html, xml data in the folder "./data/" in the core "IR_EXAM_2"

36. 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 12 ½ 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 line)

        //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));
} }
```