Computer Science 572 Exam
Prof. Horowitz
Monday, November 26, 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 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] Define Hypervisor

2. [3 pts] Google cloud offers five major sections: Compute, Storage, Stackdriver, Tools and Big Data. To set up your cluster for homework #3 you used DataProc from which section?

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

   doc1 – To be or not to be, that is the question
   doc2 –Not who but when

4. [3 psts] Given the two documents above, doc1 and doc2, provide the reducer output if an invertedIndex is run on the documents in a Hadoop cluster.

5. [3 pts] Suppose one advertiser bids $102.75 for his ad to be displayed and a second advertiser bids $101.25 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 he pay Google?

6. [3 pts] What is the difference between Google's AdWords system and Google's AdSense system?

7. [3 pts] Google offers four types of keyword matching to its advertisers. Mention two of them.

8. [3 pts] What is the difference between a taxonomy and an ontology?

9. [3 pts] A knowledgebase will typically support both forward chaining and backward chaining. Which inference technique is typically used by a search engine?

10. [4 pts] Passage scoring is the process whereby snippets returned in answer to a set of queries are ranked for their usefulness to answer a question. Three criteria were provided for ranking the snippets. Name two of them.

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)
```

11. [3 pts] What functions are defined?

12. [3 pts] What function is used to invoke (start) the program

13. [3 pts] What data set is used to initialize the dictionary?

14. [3 pts] How many levels of edits does the program investigate?

15. [3 pts] What is the cluster hypothesis for search engines

16. [3 pts] The notes mention three criteria for adequacy of clustering methods. Name one.

17. [3 pts] Define hard clustering:

18. [3 pts] In the k-means clustering algorithm, the means refers to computing the average of a set of points. In one sentence, when in the algorithm is the average of a set of points computed?

19. [3 pts] If *m* is the size of the vector, *n* is the number of vectors (items), *k* is the number of clusters, and *i* is the number of iterations, what is the computing time for the k-means algorithm?

20. [3 pts] Given the two strings: “satisfactory” and “satisfying”, what is their minimum edit distance assuming the operations (replace, delete, insert) all have a count of *1*?

21. [3 pts] In one sentence describe what the WordNet system provides.

22. [3 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 $O(N^2)$*?*

23. [3 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?

24. [4 pts] There are three criteria that define a good clustering algorithm, describe two:

25. [4 pts] Featured snippets are Google's attempt to answer the query right on the search results page. There are 3 types of featured snippets. Mention any two of them.

26. [4 pts] There are four different approaches mentioned in class for evaluating clustering algorithms. Mention any two of them.

27. [4 pts] What are long tailed keywords.

28. [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.

29. [6 pts] There are 6 different query types supported by Solr. Mention any four of them.

30. [4 pts] There are 2 approaches discussed in class to handle non-word spelling error correction. What are they?