# 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++;
}
```

2. [3 pts] Does correlation necessarily imply causation. Yes or 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?

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?

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

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.

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?

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

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.

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.

11. [3 pts] Define Mean Reciprocal Rank scoring

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?

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

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

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

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?

17. [3 pts] Define soft clustering

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

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?

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*?

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

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

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?

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

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

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

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.

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

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.

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

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