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.

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 $\vec{d}$

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

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* $O(N^2)$?

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?

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

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.

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

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?

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

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.

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?

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?

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

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

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

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

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

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*?

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?

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?

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

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

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?

24. [4 pts] Define: breadcrumbs

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?