CS371R: Final Exam

Dec. 11, 2023

NAME: _____

UTEID: _____

INSTRUCTIONS:

- This exam has 8 problems and 15 pages. Before beginning, check that your exam is complete.
- You have 2 hours to complete the exam.
- The exam is closed book, closed notes, and closed computer, except for a scientific calculator and the provided equation sheets.
- Mark your answers **on the exam itself**. We will not grade answers on scratch paper.
- Make sure that your answers are legible and your handwriting is dark. We will be scanning the exams and grading them using Gradescope.
- In order to maximize your chance of getting partial credit, show all of your work and intermediate results.

Final grades will be available on Canvas on or before December 14.

Thank you for a great semester! Good luck and have a good break!

1. (10 points) Assuming simple term frequency weights (no IDF factor), no length normalization, and no stop words, compute the cosine similarity of the following two simple documents:

   (a) "five thousand five hundred and fifty five dollars"

   (b) "fifty six thousand five hundred sixty five dollars"

2. (14 points) Consider the following web pages and the set of web pages that they link to:

   Page A points to pages B and C.
   Page B points to page C.

   All other pages have no outgoing links.

   Consider running the PageRank algorithm on this subgraph of pages. Assume $\alpha = 0.15$. Simulate the algorithm for three iterations. Show the page rank scores before and after normalization for each page for each iteration. Order the elements in the vectors in the sequence: A, B, C.

   (a) Show work for iteration 1 below:

(b) Show work for iteration 2 below:

(c) Show work for iteration 3 below:

3. (14 points) Consider examples described using the following three binary-valued features:

```
A:  0,1
B:  0,1
C:  0,1
```

Show the trace of a perceptron learning from the training set of [A,B,C] examples:

```
[1,0,0]: positive
[1,1,1]: positive
[0,1,1]: negative
```

Assume all of the weights and the threshold start at 0 and the learning rate is 1 and that during each epoch the examples are processed in the exact order given above. Show the weight vector (in the order [A,B,C]) and the threshold after every example presentation. The procedure should converge after only 4 epochs. **NOTE:** For the purposes of this problem, assume that the net input must be **strictly greater than** the threshold in order output a 1 (so this is slightly different from the equations in neural-net lecture slides).

If represented as a logical rule, what is the function learned?

(a) Show work for iteration 1 below:

(b) Show work for iteration 2 below:

(c) Show work for iteration 3 below:

(d) Show work for iteration 4 below:

What logical rule is learned for this concept?

4. (10 points) Consider the problem of learning to classify a name as being Food or Beverage. Assume the following training set:

   Food: "cherry pie"
   Food: "buffalo wings"
   Beverage: "cream soda"
   Beverage: "orange soda"

   Apply 3-nearest-neighbor text categorization to the name "cherry soda". Show all the similarity calculations needed to classify the name, and the final categorization. Assume simple term-frequency weights (no IDF) with cosine similarity.

   Afterwards answer the question: Would the result for this particular problem be guaranteed to be the same with 1-nearest-neighbor (assuming that we break ties randomly)? Why or why not?

5. (14 points) Assume we want to categorize science texts into the following categories: Physics, Biology, Chemistry. Consider performing naive Bayes classification with a simple model in which there is a binary feature for each significant word indicating its presence or absence in the document. The following probabilities have been estimated from analyzing a corpus of preclassified web pages:

c	Physics	Biology	Chemistry
P(c)	0.35	0.4	0.25
P(atom|c)	0.2	0.01	0.2
P(carbon|c)	0.01	0.1	0.05
P(proton|c)	0.1	0.001	0.05
P(life|c)	0.001	0.2	0.005
P(earth|c)	0.005	0.008	0.01

Assuming the probability of each evidence word is independent given the category of the text, compute the posterior probability for *each* of the possible categories for each of the following short texts. Assume the categories are disjoint and complete for this application. Note that words are first stemmed to reduce them to their base form, therefore "proton" and "protons" should be considered equivalent. Ignore any words that are not in the table.

(a) The carbon atom is the foundation of life on earth.

(b) The carbon atom contains 12 protons.

6. (8 points) Consider the following web pages and the set of web pages that they link to:

   Page A points to pages B, D, and E.
   Page B points to pages C and E.
   Page C points to pages F and G.
   Page D points to page G.
   Page G points to page E.

   Show the order in which the pages are indexed when starting at page A and using a breadth-first spider (with duplicate page detection) as implemented in the course `Spider` class. Assume links on a page are examined in the orders given above.

7. (10 points) Consider the following item ratings to be used by collaborative filtering.

Item	User1	User2	User3	User4	Active User
A	8	5	10		10
B	6	2		3	3
C	2	8	2	5	1
D	9	2	6	2	
E	1	7	3	1	
$c_{a,u}$	0.88	−0.21	1	−1	

The Pearson correlation of each of the existing users with the active user ($c_{a,u}$) is already given in the table. Compute the predicted rating for the active user for items D and E using standard significance weighting and the two most similar neighbors to make predictions using the method discussed in class.

8. (20 points) Provide short answers (1–3 sentences) for each of the following questions (1 point each):

   Why is harmonic mean rather than arithmetic mean used to combine precision and recall to get F-measure?

   What is pseudo relevance feedback?

   What is an "anytime" algorithm?

   In machine learning, why is "Greed is good" an appropriate aphorism?

In both IR and ML, why is "All models are wrong, but some are useful," an appropriate aphorism?

What is "smoothing" (e.g. Laplace estimate) with respect to probabilistic categorization and why is it important?

Why do transformer models produce better word embeddings than LSA (Latent Semantic Analysis) or Word2Vec?

What are the two factors that have contributed most to the success of deep learning?

Name and define **two** shortcomings of collaborative filtering for recommending.

What indexing approach is appropriate for efficiently retrieving similar documents when using deep-learned dense embeddings of documents and queries?

(Extra credit) Which movie was the source of the quote "Greed is good"?

(Extra credit) Sort the following scientists' last names in increasing order of their birthdates: Gerald Salton, Pierre Simon Laplace, Hans Peter Luhn, Tim Berners-Lee, and Thomas Bayes.

(Extra credit) What comedy group's sketch was the source of the term "spam" as a description of unwanted, annoying messages?