CS371R: Final Exam

Dec. 14, 2019

NAME: ____

UTEID: ____

INSTRUCTIONS:

- This exam has 11 problems and 19 pages. Before beginning, check that your exam is complete.
- You have 3 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 18.

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

1. (8 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. (8 points) Draw a basic inverted index constructed for the following corpus. For each word, draw a linked list of tuples of the document name and term frequency.

   Here is an example of how to draw an entry in the inverted index:

   word → (document1, term frequency1) → (document2, term frequency2)

   - Doc1: “school of business”
   - Doc2: “school of social work”
   - Doc3: “school of public affairs”

   You are given the following list of stop words:

   - at
   - in
   - of
   - on

   Perform stop word removal and order tokens in the inverted index alphabetically. You need to include TF information, but not IDF nor document-vector lengths.

3. (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.

4. (8 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. 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. (9 points) Consider the following web pages and the set of web pages that they link to:

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

   All other pages have no outgoing links.

   Consider running the HITS (Hubs and Authorities) algorithm on this subgraph of pages. Simulate the algorithm for three iterations. Show the authority and hub scores before and after normalization for each page for each iteration. Order the elements in the vectors in the sequence: A, B, C, D, E.

   (a) Show work for iteration 1 below:

(b) Show work for iteration 2 below:

(c) Show work for iteration 3 below:

6. (9 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:

7. (10 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.

8. (9 points) Consider the problem of clustering the following documents using K-means with $K = 2$ and cosine similarity.

   Doc1: go Longhorns go
   Doc2: go Texas
   Doc3: Texas Longhorns
   Doc4: Longhorns Longhorns

   Assume Doc1 and Doc3 are chosen as the initial seeds. Assume simple term-frequency weights (no IDF, no length normalization). Show all similarity calculations needed to cluster the documents, centroid computations for each iteration, and the final clustering. The algorithm should converge after only 2 iterations.

9. (8 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.

10. (6 points) Write a regular expression in PERL syntax for matching a time and date expression such as “12:07 PM 9/27/61” and “7AM 12/2/15”, specifically a word boundary, followed by a one or two digit hour, followed optionally by a colon and two digits for the minutes, followed by an optional space, then AM or PM, followed by some non-zero amount of whitespace, then a one or two digit month, slash, a one or two digit day, slash, a two-digit year, ending in a word boundary.

11. (17 points) Provide short answers (1–3 sentences) for each of the following questions (1 point each):

What is the purpose of the "log" function in the calculation of IDF?

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

What is pseudo relevance feedback?

Why is Zipf's law important to the significant performance advantage achieved by using an inverted index?

What is the **disadvantage** of using breadth-first search for spidering compared to using depth-first search?

What is an “anytime” algorithm?

What are **three** “anytime” algorithms that we discussed in class?

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,” also an appropriate aphorism?

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

When using the language model approach with Naive Bayes text classification to solve the ad-hoc document retrieval problem, why is appropriate smoothing even more important than it normally is in most applications of text categorization?

What is the primary advantage of K-means clustering over hierarchical agglomerative clustering?

What is semi-supervised learning for text categorization?

Why do ELMO and BERT produce better word embeddings than Word2Vec?

How does compositional semantics in semantic parsing help resolve lexical and syntactic ambiguity?

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

Briefly describe **two** different approaches to combining collaborative filtering and content-based recommending.

(Extra credit) The 2019 Turing Award was given to Hinton, Bengio, and LeCun for their contributions to what area of computing?

(Extra credit) The founder of the web, Tim Berners-Lee, got his only college degree from Oxford in what subject?

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

(Extra credit) What was the original focus of Symantec (now NortonLifeLock Inc.) before they got into antivirus software?