

# CS371R Information Retrieval and Web Search: Final Exam

Dec. 13, 2011

NAME: \_\_\_\_\_

This exam has 12 problems and 17 pages. Before beginning, be sure your exam is complete.

In order to maximize your chance of getting partial credit, show all of your work and intermediate results.

Final grades will be available online on or before Dec. 16.

Good luck and have an enjoyable winter break!

1. (6 points) Assume that the total number of documents in a corpus is 20,000 and that the following words occur in the following number of documents:

- “and” occurs in 19,500 documents
- “at” occurs in 18,000 documents
- “Austin” occurs in 100 documents
- “of” occurs in 18,000 documents
- “Texas” occurs in 500 documents
- “the” occurs in 19,500 documents
- “state” occurs in 15,000 documents
- “University” occurs in 10,000 documents

You are given the following list of stop words:

- “and”
- “at”
- “of”
- “the”

Assuming that term frequencies are normalized by the maximum frequency in the given document, calculate the TF-IDF weighted term vector for the following simple document:

“University of Texas at Austin and the state of Texas”

Perform stop word removal and order tokens in the vector alphabetically.

2. (8 points) Assuming Zipf's law with a corpus independent constant  $A = 0.1$ , what is the fraction of words that appear more than 3 times in any fixed corpus? (Tip: Since half of the words appear *only* once, the fraction should obviously be less than one half.)

3. (8 points) Assume in response to the results of the query “austin bats,” that using relevance feedback, the user rates the following two documents as *relevant*:

- “bats austin night wings”
- “mexican free tailed bats”

and the following document as *irrelevant*:

- “austin ice bats fan club”

Assuming simple term-frequency weights (no normalization, no IDF), show the revised query vector computed using the “Ide regular” method. Order tokens in the vectors alphabetically. Assume  $\alpha = \beta = \gamma = 1$ .

4. (6 points) Consider the following web pages and the set of web pages that they link to:

Page A points to pages B and D.

Page B points to pages C, F, and G.

Page C points to page D.

Page D points to page H.

Page G points to pages E and H.

Page H points to page C.

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.

5. (8 points) Consider the following pages and the set of web pages that they link to:

Page A points to page E.

Page B points to pages D and E.

Page C points to pages D and E.

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 for each page twice for each iteration, both before and after normalization, order the elements in the vectors in the sequence: A, B, C, D, E.

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

Page A points to page C.

Page B points to page C.

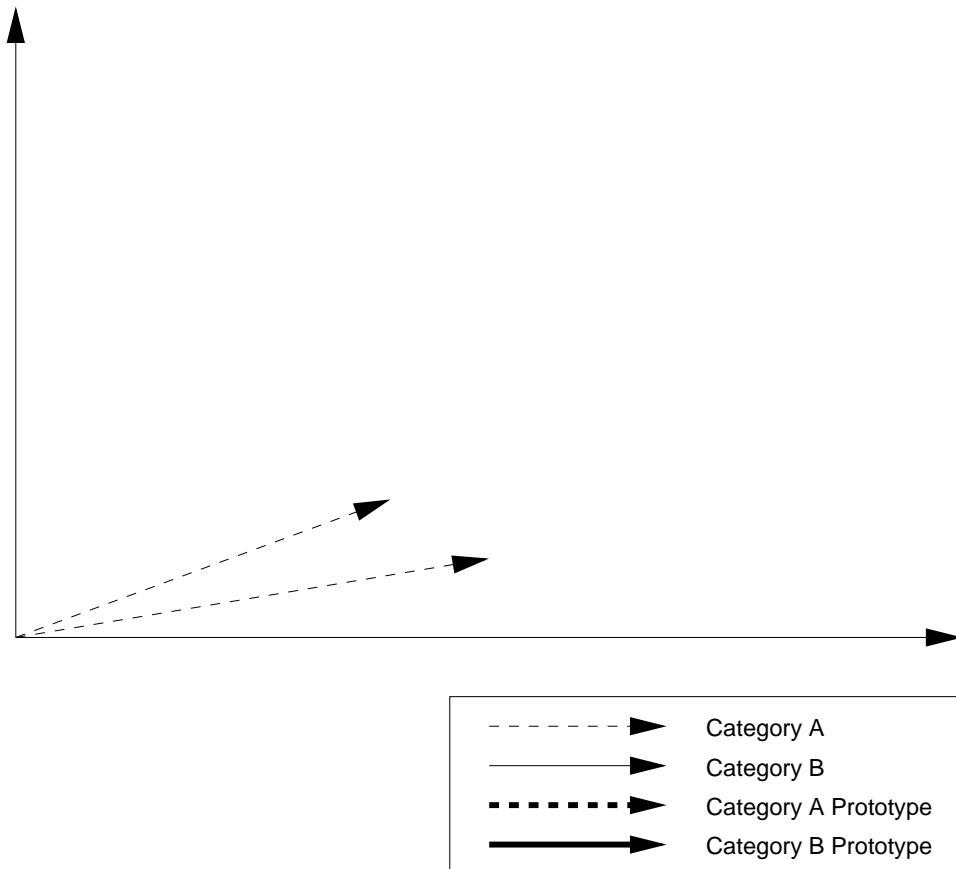
Page C points to page D.

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 for each page twice for each iteration, both before and after normalization, order the elements in the vectors in the sequence: A, B, C, D.

7. (8 points)

- (a) Assuming a two-dimensional vector space and the use of cosine similarity as the similarity metric, graphically illustrate why the Rocchio text categorization algorithm does not guarantee a consistent hypothesis (i.e. a hypothesis which correctly classifies all of the training examples).

More specifically, in the following graph, draw two document vectors for category  $B$ . Then using the document vectors in the graph, construct the prototype vectors for both categories  $A$  and  $B$ , which must be drawn **much thicker** than the document vectors. Show that at least one of the training documents would be misclassified.



- (b) Does the 3 Nearest-Neighbor algorithm guarantee a consistent hypothesis? Why or why not?

8. (8 points) Assume we want to categorize computer-science documents into the following categories: Systems, Theory, AI. 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 by analyzing a corpus of preclassified training documents.

$c$	Systems	Theory	AI
$P(c)$	0.35	0.40	0.25
$P(\text{theorem} c)$	0.05	0.8	0.10
$P(\text{search} c)$	0.30	0.40	0.60
$P(\text{heuristic} c)$	0.05	0.01	0.50
$P(\text{disk} c)$	0.30	0.02	0.01
$P(\text{data} c)$	0.50	0.01	0.20

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. Ignore any words that are not in the table.

- (a) Data on heuristic search for theorem proving
- (b) Search for data stored on disk

9. (8 points) Consider the problem of clustering the following documents using single-link hierarchical agglomerative clustering.

- Doc1: alpha alpha gamma
- Doc2: alpha beta
- Doc3: alpha gamma
- Doc4: alpha alpha

Show all similarity calculations needed to cluster the documents, and the final cluster hierarchy. Assume simple term-frequency weights (no normalization, no IDF) with cosine similarity.

10. (8 points) Consider the following item ratings to be used by collaborative filtering.

Item	User1	User2	User3	User4	Active User
A	1	5	8	7	9
B		4	3	3	5
C	9		2	1	1
D	4	1	9	8	
E	8	8	3	1	
$c_{ij}$	-1	1	0.933	0.982	

The Pearson correlation of each of the existing users with the active user ( $c_{ij}$ ) 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.

11. (6 points) Suppose that when given the following job posting:

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Prefer 5 years or more experience with PC-Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is DOS. May go to OS-2 or UNIX in future.

an information extraction system extracted the following slot/value pairs for the job template:

- Language: PC-Based
- Platform: DOS
- Area: Voice Mail
- Desired degree: 5

Assume that the solution template consists of the following correct extractions:

- Language: C
- Platform: DOS
- Platform: OS-2
- Platform: UNIX
- Area: Voice Mail
- Required years of experience: 2
- Desired years of experience: 5

How did the information extraction system perform in terms of (a) precision, (b) recall, and (c) F-measure?

12. (18 points) Provide short answers (1-3 sentences) for each of the following questions:

Why is cosine a better similarity metric than dot product in vector-space retrieval?

Why is the vector-space model generally considered a better retrieval model than the boolean model?

What is an inverted index and why is it a critical part of an IR system?

What is thesaurus-based query expansion?

How must an inverted index be enhanced in order to support retrieval of exact phrases?

What is topic-directed spidering?

Why is  $k$  nearest neighbor typically better than plain nearest neighbor?

How is “soft” clustering different from traditional “hard” clustering?

What is the “first rater” problem in collaborative filtering?

What are two **advantages** of collaborative-filtering-based recommending compared to content-based?

The language-modeling approach to ad hoc document retrieval is based on what approach to document classification?

What is the linear interpolation approach to smoothing for language models?

What is a meta-search engine?

Give two important differences between scientific bibliographic citations and web links with regard to their relevance for judging authority?

What is the purpose of the  $E(p)$  term in the PageRank algorithm?

Define the terms “hub” and “authority” in the context of web pages.

Why is naive Bayes text categorization typically implemented using logarithms of probabilities rather than probabilities themselves?

How does Buckshot clustering use Hierarchical Agglomerative Clustering while maintaining linear time complexity in  $n$ , which is the number of training examples?

(Extra credit) What famous scientist contributed to the foundations of probability theory, formulated the “nebular hypothesis” for the formation of the solar system, and discussed the existence of God with Napoleon?

(Extra credit) The two primary donors of our new UT computer science building dropped out of undergraduate programs at what two major universities?

(Extra credit) What company awarded a million dollar prize for improving their recommender system?

(Extra credit) The winners of the prize mentioned in the previous question clinched the final victory by using what general approach to constructing an improved predictor?