CS54701 Spring 2016 Assignment 2, due 6am EST 24 February, 2016

Prof. Chris Clifton

This homework is set up as a practice exam. While you are not limited on time for the assignment, this should give you an idea of the style and time requirements of questions you might see on the exam (although the specific topics covered may be different.)

During the actual exam, you must:

Turn Off Your Cell Phone. Use of any electronic device during the test is prohibited. *(This doesn't apply to the homework assignment.)* While the exam is not open book (since many people have the book in electronic form, not using an electronic device and open book don't go together), you are allowed a note sheet: up to two sheets of 8.5x11 or A4 paper, single-sided (or one sheet double-sided).

Time will be tight on the actual exam. If you spend more than the recommended time on any question, **go on to the next one**. If you can't answer it in the recommended time, you are either giving too much detail or the question is material you don't know well. You can skip one or two parts and still demonstrate what I believe to be an A-level understanding of the material.

Note: It is okay to abbreviate in your answers, as long as the abbreviations are unambiguous and reasonably obvious.

1 Boolean Retrieval (10 minutes, 4 points)

Give an example of an information need where features from Boolean Retrieval would be useful. That is, a standard vector-space model approach such as TF-IDF weighting and cosine similarity would not give the desired results, but using some form of boolean query would likely give substantially better results. Describe what specific boolean query capabilities/features are needed.

I would suggest you give a non-boolean query example and a boolean query example, and use this to explain why the boolean would do a better job for the particular information need.

2 TF-IDF weighting (10 minutes, 4 points)

Given the following document term frequencies:

	DocA	*DocB*	*DocC*
CS54701	3	0	1
Final	2	0	4
May	1	5	2

Compute the TFIDF (ltc with natural logarithm) weight for *DocA only.*

A. Give the method you are using to determine weights. You may do this by giving the formula used.

B. Complete the table for *DocA* only.

	DocA	*DocB*	*DocC*
CS54701		x	x
Final		x	x
May		x	x

3 Language Models (15 minutes, 9 points)

Given the following document term frequencies:

	DocA	*DocB*	*DocC*
CS54701	3	0	1
Final	2	0	4
May	1	5	2

Compute a generative unigram language model for *DocC only.*

A. Complete the table for *DocC* only.

	DocA	*DocB*	*DocC*
CS54701	x	x	
Final	x	x	
May	x	x	

B. Explain how you would use this language model to rank the documents in response to a query.

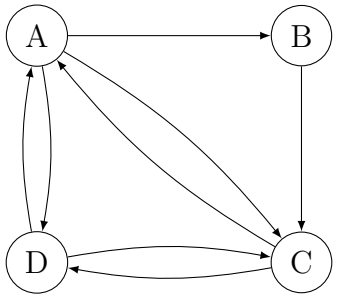
C. Is the following statement true or false? Explain your answer: A probabilistic retrieval model gives each document a score that is the probability that the document meets the user's information need.

4 Latent Semantic Indexing (15 minutes, 5 points)

One purported advantage of Latent Semantic Indexing is that it can retrieve documents that address the same concept as a query, even if the words do not match. Explain briefly how this can happen. While you may find it useful to give an example matrix, I do not expect you to do an actual singular-value decomposition; an example that shows how this could be happen is fine, even if the matrices don't actually multiply out to something close to the original term-document matrix.

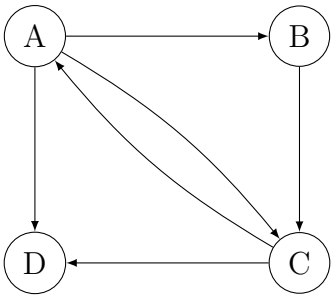
5 Link Analysis/PageRank (15 minutes, 6 points)

A. Construct the stochastic matrix for the following graph:



B. Show the first three iterations of the eigenvector computation.

C. Consider a modified graph in which the edges (d, a) and (d, c) are removed. What is the problem with this graph and how can it be fixed?


A
B
D
C