CSE 494: Information Retrieval, Mining and Integration on the Internet

Midterm. 18th Oct 2011 (Instructor: Subbarao Kambhampati)
In-class Duration: Duration of the class 1hr 15min (75min)

Total points: 65

Name: Solutions Student ID: Rao.

There are 10 pages, including the front page, in this exam.

Closed book and notes; you are allowed one 8.5x11 sheet (both sides) of whatever information you want to remember.

Must be answered on this document, in the space provided (*answers on separate ruled sheets etc won't be accepted)*. If you need more space, you may use the backs of the sheets (but then put a note so I won't miss them).
[You must SHOW YOUR WORK to get partial credit]

Qn I Vector Similarity/Bag-of-words/indexing/tolerant dictionaries [15]	
Qn II correlation analysis/Latent Semantic Indexing [17]	
Qn III PageRank/Authorities-Hubs [14]	
Qn IV Short Answer [19]	

PLEASE LOOK AT THE ENTIRE PAPER ONCE. EASY QUESTIONS MAY BE LURKING ALL OVER THE PLACE.

Qn I.[15] **[Vector Similarity/Bag of Words/indexing/tolerant dictionaries]**

A document corpus C consists of the following three documents:

D1: "new york times"
D2: "new york post"
D3: "los angeles times"

a. [4] Assuming that term frequencies are normalized by the maximum frequency in a given document, calculate the TF-IDF weighted term vector for the document D2. Assume that the words in the vectors are ordered alphabetically. (use logarithms to base 2 in computing IDFs. Here are some useful logs: log 3 = 1.585; log 1.5= .585.

$w_{new} = tf \times idf = \frac{1}{1} \times \log_2 3/2 = 0.585$

$w_{post} = 1/1 \times \log_2 3 = 1.585$

$w_{york} = 1/1 \times \log_2 3/2 = 0.585$

$D_2 = \langle 0, 0, 0.585, 1.585, 0, 0.585 \rangle$

b. [3] Given the following query: Q:"*new times*". Calculate the cosine similarity between Q and the document D2. **Assume that the query vector doesn't consider IDF weights (i.e., has just term frequency weights).**

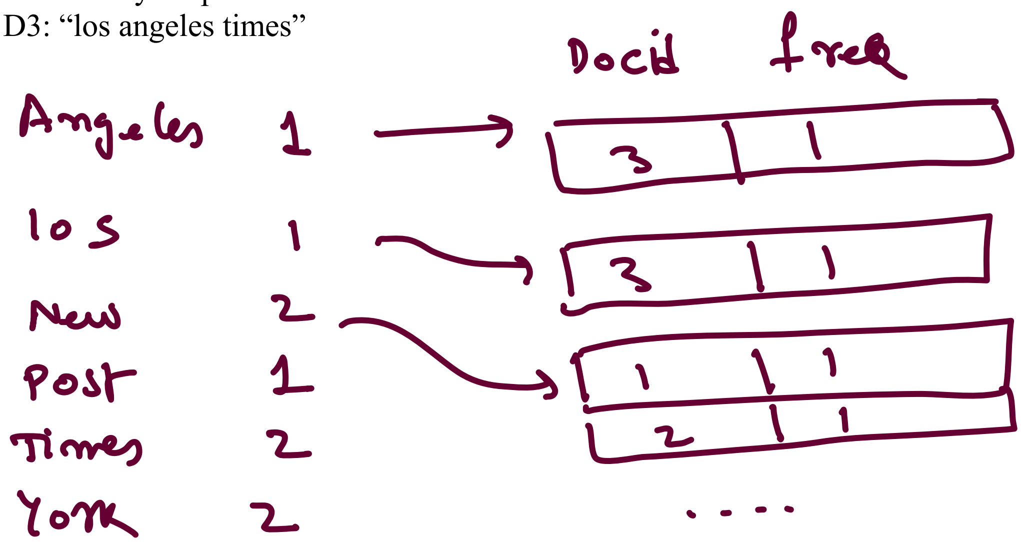
$Q = \langle 0, 0, 1, 0, 1, 0 \rangle$

$$sim = \frac{0.585 \times 1}{\sqrt{2}\sqrt{0.585^2 + 0.585^2 + 1.585^2}}$$

$= 0.23$

c.[4] Show graphically how the inverted index for the corpus C looks like. (Recall that C has the three documents:

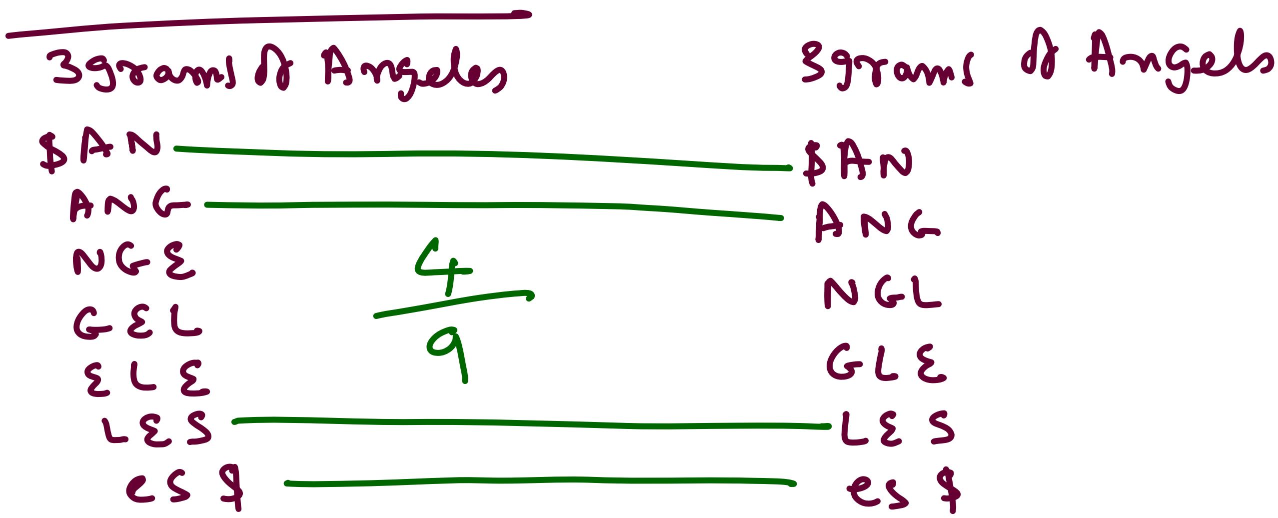
D1: “new york times”
D2: “new york post”
D3: “los angeles times”



d.[4] A user (mis)types the word “angles”. Assuming we are using Damerau-Levenshtein distance, what is the closest word to this word in the lexicon? What is its distance? What is that same word’s “similarity” to “angles” if we use 3-gram Jaccard Similarity?

Closest word is Angeles

D-L distance = 1
(insert e at loc 4)



Qn II [17] **[Scalar clustering & Latent Semantic Indexing]** Continuing our obsession with the corpus C from the previous question, suppose we were to take the D-T matrix of the corpus, but with the terms represented by just their frequency (and no normalization, no IDF).

In this case D-T for this corpus would be

0	0	1	0	1	1
0	0	1	1	0	1
1	1	0	0	1	0

where the rows correspond to the documents D1,D2 and D3 and columns correspond to terms “angeles”, “los”, “new”, “post”, “times” and “york” (basically terms alphabetically arranged).

We want to compute the association clusters, and so compute the T-T matrix as D-T’ * D-T. This 6x6 matrix is:

1	1	0	0	1	0
1	1	0	0	1	0
0	0	2	1	1	2
0	0	1	1	0	1
1	1	1	0	2	1
0	0	2	1	1	2

$$A_{uv} = \frac{a_{uv}}{a_{uu} + a_{vv} - a_{uv}}$$

[4] What is the *normalized* association cluster for the word “new”? According to this calculation, which is the word that is most correlated with it?

New = [0 0 1 1/2 1/3 1]

$$A_{35} = \frac{1}{2+2-1} = 1/3$$

$$A_{36} = \frac{2}{2+2-2} = 1 \qquad A_{34} = \frac{1}{2+1-1} = 1/2$$

[2]If you continue and do scalar clustering on these words, give an example of word correlation you are likely to see increase, and explain why (an intuitive explanation is enough).

we are likely to increase correlation of 'POST' and 'TIMES' as they both occur together with new & York but in diff docs

Doing SVD (LSI) analysis on D-T gives the following three matrices as D-F, F-F and T-F:

-0.7071	0.0000	0.7071
-0.6325	-0.4472	-0.6325
-0.3162	0.8944	-0.3162

2.2882	0	0	0	0	0
0	1.7321	0	0	0	0
0	0	0.8740	0	0	0

-0.1382	0.5164	-0.3618	0.1772	-0.7378	-0.0874
-0.1382	0.5164	-0.3618	-0.6436	0.3979	0.1039
-0.5854	-0.2582	0.0854	-0.3057	-0.1044	-0.6921
-0.2764	-0.2582	-0.7236	0.4664	0.3399	-0.0164
-0.4472	0.5164	0.4472	0.4664	0.3399	-0.0164
-0.5854	-0.2582	0.0854	-0.1607	-0.2355	0.7085

Where D-T = D-F * F-F * T-F^T (Equation 1)

In the exam I asked D-F x D-F' (instead of D-T x D-T') for DF x DF', the eigen values are all 1 since D-F x DF' will be I (D-F is an orthonormal matrix)

b. [1] What is the primary eigen value of D-T * D-T^T ?

$(2.2882)^2$

Suppose we decide to reduce the dimensionality of the data to just 2 dimensions

c. [2] What is the fraction of data variance that we have lost by this decision?

$$\text{loss} = 1 - \frac{2.2882^2 + 1.732^2}{2.288^2 + 1.732^2 + 0.874^2} \approx 8.4\%$$

e. [2+6] What is the vector-space similarity between D1 and D2 in the original and reduced LSI space? (For LSI space, consider just the 2-dimensional one)

original $D_1 = [0\ 0\ 1\ 0\ 1\ 1]$

$[0\ 0\ 1\ 1\ 0\ 1]$

$$\text{Sim}(D_1, D_2) = \frac{1.1 + 1.1}{\sqrt{3}\sqrt{3}} = 2/3 = 0.6\bar{6}$$

factored space

$$D\text{-}f \times f\text{-}f = \begin{bmatrix} -0.7071 & 0 \\ -0.6325 & -0.4472 \end{bmatrix} \times \begin{bmatrix} 2.2882 & 0 \\ 0 & 1.7321 \end{bmatrix}$$

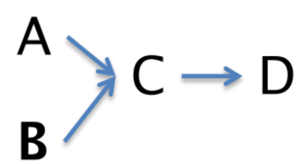
So D_1 in factored space $= [-1.6179 \quad 0]$

$D_2 = [-1.4472 \quad -0.7745]$

$$sim(D_1, D_2) = \frac{(-1.6179)(-1.4472)}{\sqrt{1.6179^2}\sqrt{1.4472^2 + 0.7745^2}} \approx \boxed{0.88}$$

Qn IV [6+6+2] Consider a webgraph with just the following links:

Page A points to page C
Page B points to page C
Page C points to page D



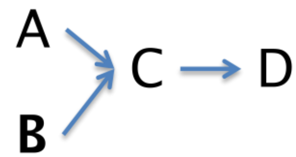
IV.A. If you run Authorities/Hubs computation on this graph, starting with uniform authority and hub values, what are the A and H values after one iteration on each? <u>Show the appropriate matrices.</u>

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad A^T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$a_1 = A^T h_0 = \begin{bmatrix} 0 \\ 0 \\ 2 \\ 1 \end{bmatrix} \quad \text{Normalize} \quad \begin{bmatrix} 0 \\ 0 \\ 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}$$

$$h_1 = A\,a_1 = \begin{bmatrix} 2/\sqrt{5} \\ 2/\sqrt{5} \\ 1/\sqrt{5} \\ 0 \end{bmatrix} \quad \text{Normalize} \quad \begin{bmatrix} 2/3 \\ 2/3 \\ 1/3 \\ 0 \end{bmatrix}$$

IV.B. If you run the PageRank algorithm on this graph, with a uniform reset matrix, and the reset probability 0.15, what is the page rank after one iteration?



$$M = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad Z = \begin{bmatrix} 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 1/4 \end{bmatrix}$$

$$\text{Reset Matrix } K = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

$$M^* = 0.85\,(M+Z) + 0.15\,K = \begin{bmatrix} 0.0375 & 0.0375 & 0.0375 & 0.25 \\ 0.0375 & 0.0375 & 0.0375 & 0.25 \\ 0.8875 & 0.8875 & 0.0375 & 0.25 \\ 0.0375 & 0.0375 & 0.8875 & 0.25 \end{bmatrix}$$

$$R_1 = M^* \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 0.090625 \\ 0.090625 \\ 0.515625 \\ 0.303125 \end{bmatrix}$$

sum will be 1

IV.C. If we now find out that left to themselves, users tend to land on D twice as often as any other page. How can this information be used to improve pagerank computation? (You don't need to recompute anything—just explain what part will change and how).

use K to be

$$\begin{bmatrix} 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 \\ 0.4 & 0.4 & 0.4 & 0.4 \end{bmatrix}$$

← prob of going to D is double

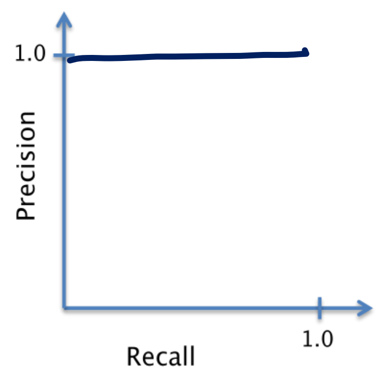
Qn IV.[19] Short Answer Questions

[2] An IR system returned 10 ranked documents for a given query. According to the gold standard labeling, there are 5 relevant documents for this query. The only relevant documents returned by the IR system are in 2nd, 3rd, 4th and 8th positions. What is the precision and recall of the IR system as computed at the 10th ranked document?

Recall = 4/5

Precision = 4/10

[2] What does the precision-recall curve of a perfect IR system look like? Sketch it on the graph to the right.



[1+2] Your friend has one of those magnetic word sets. Every day, he arranges all of these words into a new sentence (he takes pride in always using every word). (i) According to the vector space bag-of-words model, what is the similarity between the maximally dissimilar sentences your friend can make? (ii) Can you think of a technique we discussed in class that can better discriminate among the sentences?

Similarity is always 1

Can use either K-grams or Proximity Search to take word order into account

[2] One way link analysis on web graphs can be made efficient is if the matrices are "sparse". The adjusted stochastic matrix used in pagerank computation is not however sparse. So how do algorithms for computing pagerank handle this efficiently?

Rather than use M^*, we just use the link matrix (which is sparse) and compute entries of M^* as needed

[2] Describe one way in which the Google prototype described in the Brin & Page paper differs from the standard web search engine design as discussed in the class.

- doesnt use idf
- uses 2 barrels

. .

[2] Briefly describe how indexing is made efficient by the use of Map-Reduce architecture

easiest is to draw the Map-Reduce figure for indexing ✓

[2] List one advantage and one disadvantage of link analysis at the query time as compared to global link analysis.

- doing at Query time makes importance Query sensitive
- but the additional time for importance calculation is part of Query time!

[2] What is *eigen gap* and how is it useful in predicting the stability of A/H computation?

Eigengap = $|\lambda_1 - \lambda_2|$

the larger it is, the more stable A/H is wrt random changes

[2] Write a non-trivial deep detail that you understood and wished I asked a question about it on the exam.