Joshua Mitchell
CS 7312 Assignment 4

1.1 (5 points) Draw the inverted index that would be built for the following document collection. (See Figure 1.3, Page 6 of the textbook, for an example.)

Doc 1 new home sales top forecasts

Doc 2 home sales rise in july

Doc 3 increase in home sales in july

Doc 4 july new home sales rise

Value	Key
new	1, 4
home	1, 2, 3, 4
sales	1, 2, 3, 4
top	1
forecasts	1
rise	2, 4
in	2, 3
july	2, 3, 4
increase	3

1.2 (5 points) Consider the following fragment of a positional index with the format:

word: document: <position, position, . . .>; document: < position, . . .>
. . .
Gates: 1: <3>; 2: <6>; 3: <2,17>; 4: <1>;
IBM: 4: <3>; 7: <14>;
Microsoft: 1: <1>; 2: <1,21>; 3: <3>; 5: <16,22,51>;

The /k operator, word1 /k word2 finds occurrences of word1 within k words of word2

Describe the set of documents that satisfy the query Gates /2 Microsoft.

Documents 1 and 3 satisfy this query.
Gates is within 2 words of Microsoft in document 1, and Gates is within 1 word of Microsoft in document 3.

(20 points) Computing ranking scores in a search engine with the lnc.ltc weighting scheme. Let the query be "good student" and the document be "good bad student good bad instructor". Fill out the empty columns in the following table and then compute the cosine similarity between the query vector and the document vector. In the table, df denotes document frequency, idf denotes inverse document frequency (i.e., idft = log10N/dft), tf denotes term frequency, log tf denotes the tf weight based on log-frequency weighting as shown in slides (i.e., 1+log10tft,d for tft,d> 0 and 0 otherwise), q is the query vector, q' is the length-normalized q, d is the document vector, and d' is the length-normalized d. Assume N = 10,000,000.

The cosine similarity between q and d is the dot product of q' and d', which is:

			query				document			
terms	df	idf	tf	log tf	q	q'	tf	log tf	d	d'
bad	1000	4	0	0	0	0	2	1.3	1.3	0.561
good	10000	3	1	1	3	0.794	2	1.3	1.3	0.561
instructor	10	6	0	0	0	0	1	1	1	0.431
student	50000	2.3	1	1	2.3	0.608	1	1	1	0.431

Dot product of q' and d': 0.707

3.1 (10 points) An IR system returns 8 relevant documents, and 10 nonrelevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system for this search, what is its recall? what is the balanced F measure?

The precision is 0.44, the recall is 0.4, and the balanced F measure is 0.421.

3.2 (10 points) Consider an information need for which there are 4 relevant documents in the collection. Compare two systems that run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):

System 1: R N R N N N N N R R

System 2: N R N N R R R N N N

a. What is the MAP of each system? Which has a higher MAP?
b. What is the R-precision of each system? Does it rank the systems the same as MAP?

a.
System 1 MAP: $(1 + 0 + 2/3 + 0 + 0 + 0 + 0 + 0 + 3/9 + 4/10) / 4 = 0.6$
System 2 MAP: $(0 + 1/2 + 0 + 0 + 2/5 + 3/6 + 4/7 + 0 + 0 + 0) / 4 = 0.493$
System 1 has the higher MAP.

b.
The R precision of System 1 is 0.5, and the R precision of System 2 is 0.25.
The ranking order of each system is the same (System 1 comes first).

3.3 (20 points) The following list of R's and N's represents relevant (R) and nonrelevant (N) documents in a ranked list of 20 documents in response to a query from a collection of 10,000 documents. The leftmost item is the top ranked search result. This list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection.

R R N N N N N N R N R N N N R N N N N R

a. What is the precision of the system on the top 20?
b. What is the F1 (balanced F measure) on the top 20?
c. What is the uninterpolated precision of the system at 25% recall?
d. What is the interpolated precision at 33% recall?

e. Assume that these 20 documents are the complete result set of the system. What is the MAP for the query?

Assume, now, instead, that the system returned the entire 10,000 documents in a ranked list, and these are the first 20 results returned.

f. What is the largest possible MAP that this system could have?
g. What is the smallest possible MAP that this system could have?

a. Precision: $6/20 = 0.3$
b. F1: $(2 * 0.3 * 0.75) / (0.3 + 0.75) = 0.429$
c. Well, the precision values at 25% recall are: 2/2, 2/3, 2/4, 2/5, 2/6, 2/7, and 2/8. The average of these values is 0.491
d.
e. System MAP: $(1 + 1 + 0 + 0 + 0 + 0 + 0 + 0 + 3/9 + 0 + 4/11 + 0 + 0 + 0 + 5/15 + 0 + 0 + 0 + 0 + 6/20) / 6 = 0.555$

f. Largest MAP: $(1 + 1 + 0 + 0 + 0 + 0 + 0 + 0 + 3/9 + 0 + 4/11 + 0 + 0 + 0 + 5/15 + 0 + 0 + 0 + 0 + 6/20 + 7/21 + 8/22) / 8 = 0.503$
g. Smallest MAP: $(1 + 1 + 0 + 0 + 0 + 0 + 0 + 0 + 3/9 + 0 + 4/11 + 0 + 0 + 0 + 5/15 + 0 + 0 + 0 + 0 + 6/20 + 0 + ... + 0 + 7/9999 + 8/10000) / 8 = 0.416$

4.1 (15 points) Consider a web graph with three nodes 1, 2 and 3. The links are as follows: 1 -> 2, 3 -> 2, 2 -> 1, 2 -> 3. Write down the transition probability matrices for the random surfer's walk with teleporting, for the following three values of the teleport probability:
(a) a = 0; (b) a = 0.5 and (c) a = 1.

$$a: \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{bmatrix} \quad b: \begin{bmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{bmatrix} \quad c: \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

4.2 (15 points) For the web graph shown below,
compute PageRank, hub and authority scores for each of the three pages.

PageRank: Assume that at each step of the PageRank random walk,
we teleport to a random page with probability 0.1, with a uniform
distribution over which particular page we teleport to.

Hubs/Authorities: Normalize the hub (authority) scores so that the
maximum hub (authority) score is 1.

$$P = \begin{bmatrix} 1/30 & 29/60 & 29/60 \\ 1/30 & 1/30 & 14/15 \\ 1/30 & 14/15 & 1/30 \end{bmatrix}$$

The above is the normalized random walk matrix for the web graph. It turns out that $(1/3, 1/3, 1/3)^T$ is an eigenvector with an eigenvalue of 1.

(I picked (1/3, 1/3, 1/3) initially like the example in the slides, and it converged immediately.)

Thus, the PageRank of each vertex is 1/3.

For HITS, the initial h(x) and a(x) scores as well as the updates are as follows:

Vertex	h(x)	a(x)
1	2	0
2	1	2
3	1	2

$\Rightarrow$

Vertex	h(x)	a(x)
1	1	0
2	1/2	1
3	1/2	1

$\Rightarrow$

Vertex	h(x)	a(x)
1	1	0
2	1/2	3/4
3	1/2	3/4

So it appears that vertex q1 is the best hub, and q2/q3 are the best authorities.