Assignment 4 (100 points)

Submit to TRACS

1. The following questions are related to chapter 1 and chapter 2.

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

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 (on either side), where k is a positive integer argument. Thus k = 1 demands that word1 be adjacent to word2.

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

Doc 1<Microsoft:1, Gates:3>, Doc 3<Gates:2, Microsoft:3>

2. The following question is related to chapter 6.

(20 points) Computing ranking scores in a search engine with the Inc.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., $\text{idf}_t = \log_{10}N/\text{df}_t$), tf denotes term frequency, log tf denotes the tf weight based on log-frequency weighting as shown in slides (i.e., $1+\log_{10}\text{tf}_{t,d}$ for $\text{tf}_{t,d} > 0$ 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.

			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**	**5.2**	**0.54**
good	10000	**3**	**1**	**1**	**3**	**0.79**	**2**	**1.3**	**4.9**	**0.51**
instructor	10	**6**	**0**	**0**	**0**	**0**	**1**	**1**	**6**	**0.62**
student	50000	**2.3**	**1**	**1**	**2.3**	**0.61**	**1**	**1**	**2.3**	**0.24**

The cosine similarity between q and d is the dot product of q’ and d’, which is: **0.55**

3. The following questions are related to chapter 8.

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?

Precision: 0.44
Recall: 0.4
F-measure: 0.42

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?

MAP-1: 0.46
MAP-2: 0.38
System 1 has the higher MAP.

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

R-1: 0.5
R-2: 0.25
System 1 has the higher R-precision which is the same ranking as MAP.

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?

Precision: 0.3

b. What is the *F1* (balanced F measure) on the top 20?

F-measure: 0.43

c. What is the uninterpolated precision of the system at 25% recall?

Precision: 1.0

d. What is the interpolated precision at 33% recall?

Precision: 0.80

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

MAP: 0.41

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?

MAP: 0.00004102

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

MAP: 0.00004100

4. The following questions are related to chapter 21.

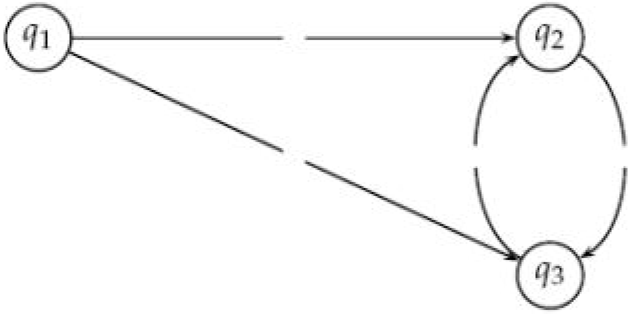
4.1 (15 points) Consider a web graph with three nodes 1, 2 and 3. The links are as follows: 1 -> 2, 3 -> 2, 2 -> 1, 2 -> 3. Write down the transition probability matrices for the random surfer's walk with teleporting, for the following three values of the teleport probability: (a) = 0; (b) = 0.5 and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

a=0	Node 1	Node 2	Node 3
Node 1	0	1	0
Node 2	0.5	0	0.5
Node 3	0	1	0

a=0.5	Node 1	Node 2	Node 3
Node 1	0.17	0.67	0.17
Node 2	0.42	0.17	0.42
Node 3	0.17	0.67	0.17

a=1	Node 1	Node 2	Node 3
Node 1	0.33	0.33	0.33
Node 2	0.33	0.33	0.33
Node 3	0.33	0.33	0.33

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.

Q1: 0.03
Q2: 0.48
Q3: 0.48

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

H1 = 1	**A1 = 0**
H2 = 0.5	**A2 = 1**
H3 = 0.5	**A3 = 1**

Submission: Type in Word (or write CLEARLY on paper and scan it) and submit electronically to TRACS.