Stanford University Computer Science Department
CS347 Spring 2001 Mid-term (Total of 30 points.)

This exam is open book, open notes. You have 70 minutes.

Print your name:_____

The Honor Code is an undertaking of the students, individually and collectively:

1. that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;
2. that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.

The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.
While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.
I acknowledge and accept the Honor Code.

Signed:_____

Problem	Points	Maximum
1		4
2		5
3		5
4		5
5		6
6		5
Total		

Question 1: (4 points)

Shown below is a portion of the positional index in the format

term: doc1: position1,position2 …; doc2: position1, position2 …; etc.

angels:	*2*:36,174,252,651;	*4*:12,22,102,432;	*7*:17;
fools:	*2*:1,17,74,222;	*4*:8,78,108,458;	7:3,13,23,193;
fear:	*2*:87,704,722,901;	*4*:13,43,113,433;	*7*:18,328,528;
in:	*2*:3,37,76,444,851;	*4*:10,20,110,470,500;	*7*:5,15,25,195;
rush:	*2*:2,66,194,321,702;	*4*:9,69,149,429,569;	*7*:4,14,404;
to:	*2*:47,86,234,999;	*4*:14,24,774,944;	*7*:199,319,599,709;
tread:	*2*:57,94,333;	*4*:15,35,155;	*7*:20,320;
where:	*2*:67,124,393,1001;	*4*:11,41,101,421,431;	*7*:16,36,736;

Which document(s) if any meet each of the following queries, where each expression within quotes is a phrase query?

(i) ***"fools rush in"***

(ii) ***"fools rush in"*** *AND* ***"angels fear to tread"***

Question 2: (1+1+3 points)

(i) How many postings entries are generated in the bigram index for the following text (underlined portion only): *Far from the madding crowds ignoble strife* ?

(ii) For this bigram index, how would the wild-card query ***madd*ing*** be expressed as an *AND* query?

(iii) Consider a corpus with 10 million documents, with an average of 1000 words each. The average word length across the corpus is 4 characters, not counting spaces or punctuation. Estimate the number of trigram occurrences for this corpus.

Question 3: (5 points)

Recall the estimate of the total size of the postings entries (45Mbytes using γ codes) from Lecture 1 using Zipf's law. Using the same parameters (1 million documents, 500,000 terms), re-compute this estimate if we were to omit from indexing the 1% of the most frequently occurring terms.

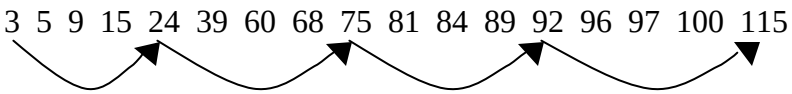
Question 4: (5 points)

Mark each of the following assertions as True or False.

	Assertion	T/F
(i)	The optimal order for query processing in an *AND* query is always realized by starting with the term occurring in the fewest documents.	
(ii)	The γ code for 17 is 111000001.	
(iii)	The base of the logarithm used in the tf×idf formula makes no difference to the cosine distance between two documents (provided they both use the same base).	
(iv)	If we were to take a document and double its length by repeating every occurrence of every word, then the normalized tf×idf values for all terms in this document remain unchanged.	
(v)	The optimal order for query processing in an *AND* query is always realized by starting with the term occurring in the fewest documents.	

Question 5: (6 points)

(i) Consider the following postings list augmented with skip pointers at a uniform skip size of 4.



The entries in the list are NOT gap encoded; they directly represent document IDs. At some stage in processing an AND query, we need to merge the entries in this postings list with the entries in the candidate list (3,5,89,95,97,99,100,101). We define the following four operations:

- **Compare(A,B):** compare entry A in postings list with entry B in the candidate list
- **Output(X):** output value X as a result of the merge process
- **LookAheadTo(X):** peek ahead to take a look at the target X of a skip pointer
- **SkipTo(X):** follow a skip pointer to reach entry X

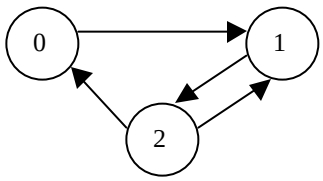
a) In performing this merge, list the instances of Output, LookAheadTo, and SkipTo operations in the order in which they would be executed. (Note: Do not list instances of Compare).

b) How many Compare operations would be executed? (Note: Do not list the actual instances).

(ii) Suppose we wish to gap-encode postings list containing skip pointers. One approach is to store all entries in the list as gaps, except for targets of the skip pointers which will be stored as absolute values. For example, in the postings list of Part (i), 24, 75, 92, and 115 will be stored as is, whereas the remaining entries will be gap-encoded, yielding (3 2 4 6 24 15 21 .... 4 1 3 115). Suppose (7 8 9 39 28 60 130 70 10 215) represents a gap-encoded postings list with skip pointers at **skip size 3**. What will be the result of merging this postings list with the candidate list (9 24 127 135 210)?

Question 6: (5 points)

(i) Represent the following simplified graph of the web as a Markov chain by providing the corresponding transition probability matrix. Assume teleportation to a random page (including the start page) occurs with **50%** probability.



(ii) Using the initial probability vector [0 1 0], carry forward the Markov chain 1 time step. (I.e., give the probability vector for time t = 1)