CS 276 / LING 286 Information Retrieval and Web Search
Spring 2013 Final Exam

This examination consists of 15 printed sides, 10 questions, and 100 points. The final counts for 30 percent of your final grade. Please write your answers on the exam paper in the spaces provided. You may use the back of a page if necessary. You have 3 hours to complete the exam. Examinations turned in after the end of the examination period will be either penalized or not graded at all. The exam is open book and open notes. You are allowed to use a laptop/tablet but with access to the Internet or any other communication means disabled. You are not allowed to use any programming capabilities beyond a basic calculator.

If you are taking the exam **remotely**, please send us the exam by **Friday 3:00 pm PDT**. You can either send a scanned copy at cs276-spr1213-staff@lists.stanford.edu (preferable) or fax at +1 (650) 725-1449.

Stanford University Honor Code: I attest that I have not given or received aid in this examination, and that I have done my share and taken an active part in seeing to it that others as well as myself uphold the spirit and letter of the Honor Code.

SUID (Stanford email ID): _____ Name (printed): _____

Signature: _____

Question	Score	Possible
1. True or False		10
2. Short Answers 1		10
3. Short Answers 2		10
4. Clustering		10
5. Link Analysis		10
6. SVM		10
7. Evaluation		12
8. Vector Space Models		8
9. Ranking		8
10. Index Construction		12
TOTAL		100

The standard of academic conduct for Stanford students is as follows:

1. The Honor Code is an undertaking of the students, individually and collectively: a. 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; b. that they will do their share and take an active part in seeing to it that they as well as others uphold the spirit and letter of the Honor Code.
2. 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.
3. 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.

1 True or False (10 pts)

Indicate T(rue) or F(alse).

1. Stemming increases the size of the dictionary in an inverted index	
2. Using positional postings lists, we can determine the width of the smallest window in a document containing all query terms in time linear in the lengths of the postings lists of the query terms	
3. Skip pointers speed up a disjunctive X OR Y query	
4. Using SVMs with nonlinear kernels dramatically improves performance for IR tasks	
5. When sufficient data is available, SVMs generally perform as well or better than other common classifiers	
6. K-means clustering (initialized with random centroids) is deterministic	
7. K-means requires more memory than C-HAC	
8. One can choose the number of clusters after running C-HAC	
9. C-HAC has higher running time than one iteration of K-means	
10. Recall of documents is more important for a legal domain IR system than for a modern web search engine	
11. Precision at 1 is more important for a legal domain IR system than for a modern web search engine	
12. Specialized queries with /k and wildcards are more important for a legal domain IR system than for a modern web search engine	
13. Compared to just a term's tf, multiplying by idf to give a tf-idf score always gives a bigger number	
14. For multi-term queries, the WAND algorithm will always provide a savings in retrieval cost over standard postings traversal	
15. Link analysis provides scoring that is spam-proof	
16. SVMs are only usable when the classes are linearly separable in the feature space	
17. Adding training data always results in a monotonic increase in the accuracy of a Naive Bayes classifier	
18. The main reason to take the logs of probabilities in Naive Bayes is to prevent underflow	
19. Using a variable byte code gives better compression than using gamma encoding	
20. The centroid of a set of (length normalized) unit vectors is a unit vector	

2 Short Answers 1 (10 pts)

1. [2pts] Consider a collection with one billion tokens (i.e., with 10^9 tokens). Suppose the first 1,000 of these tokens results in a vocabulary size of 1,000 terms, and the first 100,000 tokens results in a vocabulary size of 10,000 terms. Use *Heap's law* to estimate the vocabulary size of the whole collection.

2. [2pts] For a conjunctive query, is processing postings lists in increasing order of their lengths guaranteed to be optimal execution order? Explain why it is, or give an example where it isn't.

3. [2pts] In one sentence, why are pairwise ranking approaches more successful for learning to rank search results than classification or regression approaches?

4. [2pts] For learning search engine rankings, what are two problems with adopting a straightforward approach of minimizing pairwise ranking errors, which have been studied and corrected in subsequent research?

   a)

   b)

5. [2pts] In the Binary Independency Model, the retrieval status value formula for each query term i contains the fraction $(1 - r_i)/r_i$ where $r_i = (n_i - s_i)/(N - S)$. Here, n_i is the total number of documents containing the term i, of which s_i documents are relevant to the query. N is the total number of documents in the collection, of which S documents are relevant.

   It was suggested that this quantity $\log(1 - r_i)/r_i$ can be well approximated by the IDF for term $i = \log N/n_i$. What are the two assumptions necessary to get this result?

   a)

   b)

3 Short Answers 2 (10 pts)

1. [1pt] What is the largest number that can be stored in 4 bytes using unary encoding?

2. [2pts] Many systems, e.g., Lucene, provide a separate function to index a set of documents in bulk, instead of just one document? What do you think is an advantage of such a function?

3. [3pts] Assume that postings lists are *gap encoded* using γ codes. Using this encoding, suppose that the postings list for the term **information** is the bit sequence:

   1111 1111 1011 1100 1101 0011 1110 0000 0

   and the postings list for the term **retrieval** is the bit sequence:

   1111 1111 1100 0000 0011 1011 1101 111

   What docids match the following query:

   information AND NOT **retrieval**

4. [4pts] Suppose the vocabulary for your inverted index consists of the following 6 terms:

   elite
   elope
   ellipse
   eloquent
   eligible
   elongate

   Assume that the dictionary data structure used for this index stores the actual terms using *dictionary-as-a-string* storage with *front coding* and a *block size* of 3. Show the resulting storage of the above vocabulary of 6 terms. Use the special symbols * and $\diamondsuit$ as used in the discussion on front coding in Chapter 5.

4 Clustering (10 pts)

1. [4pts] Calculate purity and Rand Index of the following two clusterings. D_i's are documents and C_i's are classes. (Purity of a clustering is an average of purity of individual clusters.) The true labels of the documents are:
$\{(D_1 : C_1), (D_2 : C_2), (D_3 : C_1), (D_4 : C_1), (D_5 : C_2)\}$

Clustering 1: Purity:
Cluster 1: D_1
Cluster 2: D_2 Rand Index:
Cluster 3: D_3
Cluster 4: D_4
Cluster 5: D_5

Clustering 2: Purity:
Cluster 1: D_1, D_2, D_3, D_4
Cluster 2: D_5 Rand Index:

2. [2pts] Is purity a good evaluation measure by itself? In 1–2 sentences write why or why not.

3. [4pts] Each iteration of K-means can be run using the Map-Reduce framework. Write down in 1-2 sentences what would be the (key, value) pairs in the map and the reduce step.

   Map step:

   Reduce step:

5 Link Analysis (10 pts)

Given a collection of text, consider a Markov Chain M built as follows. We have one state for each term in the collection. The transition probability from any state t_1 to another t_2 is the fraction (in the text collection) of times that term t_1 is immediately followed by term t_2. We only have states for dictionary terms; ignore start/end-of-string markers/symbols.

Consider the text: *a rose is a rose is a rose*

1. [2pts] Draw the corresponding Markov Chain M.

2. [1pt] Is M ergodic?

3. [1pt] We now augment M with a teleportation operation, to create a new Markov Chain M_1. At each step of M_1, we follow M with probability 87%, while with probability 13% we jump to a random state. Is M_1 ergodic?

4. [3pts] Compute the stationary distribution of M_1.

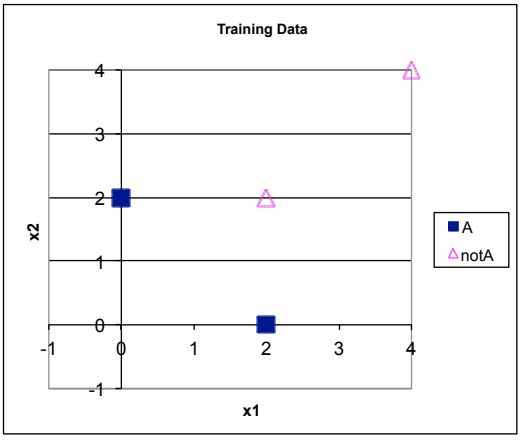
5. [1pt] Is the vector of stationary probabilities of M_1 the same as the vector of fractions of occurrences for each term in the text of part (1)?

6. [2pts] Show where the addition of a single word in the text of part (1) would change the answer in part (5) to the question: are the two vectors the same?

6 SVM (10 pts)

We are training an SVM classifier on this small training data set of 4 points:

(2,0)	Class A	(2, 2)	Class not-A
(0,2)	Class A	(4, 4)	Class not-A



Let the class A correspond to $y_i = +1$.

1. [1pt] Is this data linearly separable?

2. [1pt] What is the (geometric) margin for a hard-margin SVM on this data?

3. [2pts] Write an equation for the optimal separating hyperplane for a hard-margin SVM trained on this data. Work out the answer geometrically, based on the data points above.

4. [1pt] What is the SVM decision function (giving the class assignment)?

5. [2pts] For the algebraic formulation, we impose the standard constraint that $\mathbf{w}^T\mathbf{x}_i + b \geq 1$ if $y_i = 1$ and $\mathbf{w}^T\mathbf{x}_i + b \leq -1$ if $y_i = -1$, and then proceed to minimize $\|w\|$. What weight vector $\mathbf{w}$ and corresponding b achieves this?

6. [1pt] What is the name of the quantity calculated by $2/\|\mathbf{w}\|$ and what is its value here?

7. [2pts] Consider now a soft-margin SVM, where the value of C is not too high, so that one but only one point will be moved with a slack variable, as far as needed to maximize the margin (concretely, $C = 0.1$ will achieve this effect given this data, but you don't need to use this value to answer the question). What will the SVM decision function be now, and what is the margin?

7 Evaluation (12 pts)

1. [2pts] In the class, we discussed several ways to evaluate ranking systems, for e.g., precision, recall, NDCG etc. However, for all these metrics, we need the relevance values for the results. Two possible methods to collect relevance values, as mentioned in class, are: a) click feedback from users, b) expert judgements (like what we did for the search ratings task for PA3). Write one advantage and one disadvantage of each of these methods.

   a) Click feedback:

   b) Expert judgments:

2. [1pt] An alternative method to evaluate ranking systems is to do a pairwise comparison of two ranking systems instead of evaluating an individual system, i.e. if you have two ranking systems A and B, instead of evaluating them independently, you can perform the same query on both systems and compare the results. Write why this might give better results than evaluating individual systems independently.

3. [3pts] One such method, which uses click feedback to compare two ranking systems, is balanced interleaving. Given a query q, it combines the results, $A = (a_1, a_2, \ldots)$ of the first system and $B = (b_1, b_2, \ldots)$ of the second system, into one consolidated ranking, I as shown in Algorithm 1. Note that this is a simple merge algorithm which breaks ties using a random coin toss (done once at the beginning).

Algorithm 1 Balanced Interleaving

Input: Rankings $A = (a1; a2; \ldots)$ and $B = (b1; b2; \ldots)$

$I := ()$; //output ranking list, initialized to empty list
$k_a := 1; k_b := 1$; //counters for the lists A and B, respectively
$AFirst :=$ randomly select 0 or 1 //decide which ranking gets priority
while ($k_a <= |A|$ **and** $k_b <= |B|$) **do** //if not at end of A or B
    if ($k_a < k_b$ **or** ($k_a == k_b$ **and** $AFirst == 1$)) **then**
        if $A[k_a] \notin I$ **then** I.append($A[k_a]$) //append a result from A
        $k_a := k_a + 1$
    else
        if $B[k_b] \notin I$ **then** I.append($B[k_b]$) //append a result from B
        $k_b := k_b + 1$
    end if
end while
Output: Interleaved ranking $I = (i_1; i_2; \ldots)$

If the two input ranking lists are:
A: (a; b; c; d; g; h)
B: (b; e; a; f; g; h)
what is the output ranking I (if A wins the toss, i.e., $AFirst = 1$)?

4. [3pts] When the user enters a query q in the search engine, the back end merges the results of the two systems A and B using Algorithm 1, and the user is presented with the output list I. The user then clicks on one or more of the results, and these clicks are used to figure out which ranking, A or B, is preferred by the user.

   Formally, let $(c_1; c_2; \ldots)$ be the ranks of the clicks w.r.t. I and c_{max} be the rank of the lowest clicked link, (with i_{cmax} being the corresponding result in list I), i.e., if the user clicks on the 1^{st}, 3^{rd} and 7^{th} link, $c_{max} = 7$. Let, $k = \min\{j : (i_{cmax} = a_j) \vee (i_{cmax} = b_j)\}$; we find how many clicked results occur in the top k results for each list.

   For example, if the lowest link clicked by the user on I corresponds to the 5^{th} result in A and 7^{th} result in B, then k = 5 and we'll look at the number of clicked results present in the top 5 results of A and B. To derive a preference between A and B, we simply compare the number of clicks in the top k results of A and B, i.e., if more results from A are clicked, then A is preferred.

   In the example lists given in part 3 above, which system, A or B, is preferred if the user clicks on the 1^{st} and 4^{th} link in the output I? Explain your answer.

5. [3pts] Now, consider a bot which clicks on the results presented by the search engine uniformly at random, i.e., each link is equally likely to be clicked by the bot. Write an example of input lists A, B so that such a random clicker generates a strict preference for one of the lists in expectation. Is this a problem? Why or why not?

8 Vector Space Models (8 pts)

Consider the following retrieval scheme:
Eg. for a query ***rising interest rates***

- *Run the query as a phrase query*
- *If* $< K$ *docs contain the phrase* **rising interest rates**, *run two phrase queries* **rising interest** *and* **interest rates**
- *If we still have* $< K$ *docs, run the vector space query* **rising interest rates**. *Rank matching docs by vector space scoring*

Consider the following set of documents:

D1 : Do you like green eggs and ham

D2 : I do not like them Sam I am

D3 : I do not like green eggs and ham

D4 : I do not like eggs

D5 : Why are they green

Let the query be ***green eggs*** and suppose we eventually want to show the user 4 documents. That is, $K = 4$.

1. [2pts] Run the phrase query ***green eggs***. What are the documents that get retrieved in this step ?

2. [2pts] What are the scores of the documents you retrieved in part (1) for the phrase query using cosine similarity on a bigram vector space ?

3. [4pts] If you did not have K documents in the first step, continue scoring according to the technique. What are the final 4 documents that you return and their scores) ?

9 Ranking (8 pts)

BM25 is a measure of how relevant a document is for a query. Below is the formula for Okapi BM25:

$$RSV^{BM25} = \sum_{i \in q} (\log \frac{N}{df_i}) \frac{(k_1 + 1) tf_i}{k_1((1-b) + b\frac{dl}{avdl}) + tf_i}$$

where q is the query.

1. In this part of the problem, we will consider how BM25 performs on the query "employee benefits". Below are some statistics for our corpus.

Number of webpages	1000
Number of webpages containing "employee"	10
Number of webpages containing "benefits"	100
Average document length	78

There are only two documents relevant to "employee benefits". Below are their statistics.

	benefits.pdf	welcome.pdf
Document Length	26	39
Frequency of "employee"	3	4
Frequency of "benefits"	3	2

Let $k_1 = 1$ and $b = 0.6$. Calculate the RSV^{BM25} for each document for the query "employee benefits".

(a) [1pt] $RSV_{benefits.pdf}$:

(b) [1pt] $RSV_{welcome.pdf}$:

2. Suppose our query has only one term, briefly answer in 1 or 2 sentences for each of the following questions on the effects of varying different parameters to the RSV^{BM25}. Assuming all other parameters are fixed:

(a) [2pts] What is the effect on RSV^{BM25} of varying b from 0 to 1?

(b) [2pts] What is the effect on RSV^{BM25} when tf_i goes to infinity?

(c) [2pts] How would you manipulate the parameters k_1 and b so that the RSV^{BM25} approximates TF-IDF? Justify your answer.

10 Index Construction (12pts)

Consider constructing a nonpositional index. Due to hardware constraints and the nature of each document collection, several indexing algorithms have been proposed, e.g., the blocked sort-based indexing (BSBI) algorithm, the single-pass in-memory indexing (SPIMI) algorithm, and dynamic indexing. In this exam question, we will test your understanding of various indexing algorithms.

1. Briefly answer in **one sentence** each of the below questions:

(a) [1pt] What hardware constraint does BSBI address?

(b) [1pt] What is the advantage of using termIDs instead of terms in BSBI?

(c) [1pt] What limitation of BSBI does SPIMI address?

(d) [1pt] When do we need to use dynamic indexing?

2. [3pts] In the text book (IIR Figure 4.4), only part of the SPIMI algorithm, the inversion procedure for each block, is described, not the entire algorithm.

Your task is to complete the SPIMI algorithm (**in no more than 10 lines of code**), similar to the BSBI algorithm as detailed in IIR Figure 4.2 or below.

```
BSBIndexConstruction()
1  n ← 0
2  while (all documents have not been processed)
3  do n ← n + 1
4     block ← ParseNextBlock()
5     BSBI-Invert(block)
6     WriteBlockToDisk(block, f_n)
7  MergeBlocks(f_1, ..., f_n; f_merged)
```

For your convenience, you can:

- Call the function *SPIMI-Invert(token_stream)*
- Call any functions used in the BSBIndexConstruction algorithm above.
- View all documents in a collection as a single *token_stream* given by the function *Get-Collection-Stream()*.
- Query the method *Has_Next(token_stream)* which will return *False* if the end of the collection has been reached.

(continues on next page)

3. [5pts] Now, apply the SPIMI algorithm to the following collection:

d1: bsbi use term id
d2: sort term id doc id
d3: spimi use term
d4: no term id sort

Assume that main memory can only hold two documents at a time, i.e., the SPIMI algorithm will write to disk each time after two documents, a block, have been processed.

Write out the content of each block **just before merging** and the result **after merging** in the following format:

Block 1:
bsbi → 1
...
term → 1, 2