Scores:

1)	5)
2)	6)
3)	7)
4)	

COMP 4321 Search Engine for Web and Enterprise Data
Final Examination, Fall 2013
October 31, 2013
Time Allowed: 65 mins

Name: ______________________________ **Student ID:** ________________________

Note: Answer all questions in the space provided. Answers must be precise and to the point.

1. **[30]** Circle True or False in the following questions (totally 30 sub-questions):

[✓ T] F The term project requires you to do stemming on the keywords.

[✓ T] / F The project requires you to crawl a web site in breadth-first order.

T / [✓ F] The project requires you to keep the parent-child relationship between two pages for implementing PageRank.

[✓ T] / F The project requires you to implement the cosine similarity function.

T / [✓ F] The project requires you to implement the peak-and-plateau stemming method to stem the keywords before inserting them into the index.

T [✓ F] Google’s Adwords program charges the advertisers based on the number of impressions of the ads.

T / [✓ F] The project requires you to use cookies to remember the users of your search engine.

[✓ T] F In the evaluation done by Blair, the precision was too low but recall high enough.

[✓ T] F The system evaluated by Blair was based on the vector space model.

T [✓ F] Term frequency and inverse document frequency are the only two parameters we can use to determine term weights.

[✓ T] F Similarity between two queries can be computed with an appropriate similarity measure, as in the case of the similarity between a query and a document.

[✓ T] F Similarity between two documents can be computed with an appropriate similarity measure, as in the case of the similarity between a query and a document.

T [✓ F] In the tfxidf weighting method, a term is most important if it appears in every document.

[✓ T] F Given a query, when terms that do not match any of the query terms are added to a document, the inner product similarity of the document does not change.

T [✓ F] Given a query, when terms that do not match any of the query terms are added to a document, the cosine similarity of the document does not change.

T [✓ F] Given a query, when terms that do not match any of the query terms are added to a document, the Jaccard similarity of the document does not change.

T **F** In the vector space model, documents containing all of the query terms will always be ranked higher than those containing only some of the query terms.

T **F** A phrase must be broken down into individual words and represented as individual words in the document vector.

T **F** Terms are assumed to be independent in the document collection.

T **F** Query term weights are always binary.

T F A page with high PageRank means it has a higher chance to be visited by a user.

T **F** A page with high Page Rank means it is more relevant to the query.

T **F** HyPursuit uses links to estimate the similarity between two documents

T F WISE uses links to adjust the similarity between the query and the pages.

T F A spider is hard to make reliable because it must deal with different brands and different versions of web servers

T F A spider should be able to automatically extract hyperlinks without human intervention

T **F** Google was started by two young professors at Stanford University

T F Although most search engines use indexes to speed up search, pattern matching is still needed for other post-search functions

T F The overhead of Karp Rabin pattern matching method includes the computation of hash function and verification upon a match of the hash codes.

T **F** Karp Rabin is suitable for short patterns because the computation of hash codes is less expensive

2. **[2]** When the web has become too large, too many pages, too many users, what problems does it cause to search engines?
1 User interests are too diversified to capture in a few query terms
2 Take time to process so many pages
3 Too many relevant, high quality pages; it is hard to determine what to present to the user
4 Too many irrelevant, low-quality pages; it is hard to eliminate the junk

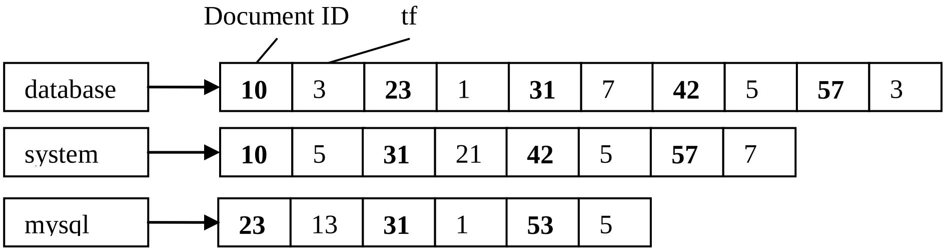
- 1, 2 and 3 only
- 3 and 4 only
- 2, 3 and 4 only
- **1, 2 and 4 only**
- All of the above

3. **[2]** The damping factor in Google PageRank formula has the following effects:
1 Normalize the PageRank value to the range from 0 to 1
2 To account for pages having no incoming links
3 To eliminate the impact of rank sinks
4 Speed up convergence

- 1, 3 and 4 only

- 2, 3 and 4 only
- 2 and 4 only
- 2 and 3 only
- All of the above

4. **[6]** The following index is a complete index of a collection, compute the (tf/tf_max)xidf weights of ‘database’ in document 10 and document 42.



database	**10**	3	**23**	1	**31**	7	**42**	5	**57**	3
system	**10**	5	**31**	21	**42**	5	**57**	7		
mysql	**23**	13	**31**	1	**53**	5				

Sol:

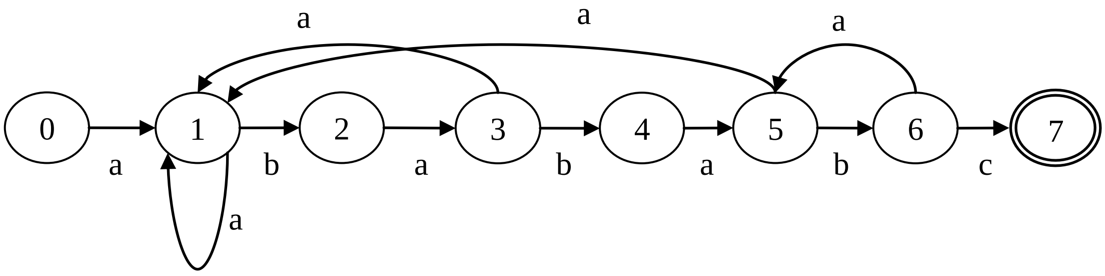
	Doc 10	Doc42
tf	3	5
idf	Log(6/5)	Log(6/5)
tf_max	5	5

5. [20] Given the pattern **abababc** (i) obtain the shift table for the basic KMP, (ii) obtain the improved KMP that avoids cascade mismatch, and (iii) a FSA matching the pattern.

Basic KMP	
	Shifts
a	1
b	1
a	2
b	2
a	2
b	2
c	2

Improved KMP	
	Shifts
a	1
b	1
a	3
b	3
a	5
b	5
c	2

FSA:



6. **[25]** The following table shows the top 10 results of query q. ‘0’ means irrelevant, ‘1’ means relevant, and ‘2’ means very relevant. There are no more relevant or very relevant documents beyond the 10th document. Compute (i) Mean Average Precision, assuming that relevant and very relevant document are relevant documents and that only one query is given, and (ii) nDCG at 10. Give the intermediate steps to show you understand the computation.

Result(q):	*D*1	*D*2	*D*3	*D*4	*D*5	*D*6	*D*7	*D*8	*D*9	*D*10
Relevance	1	0	1	2	0	0	1	0	0	2

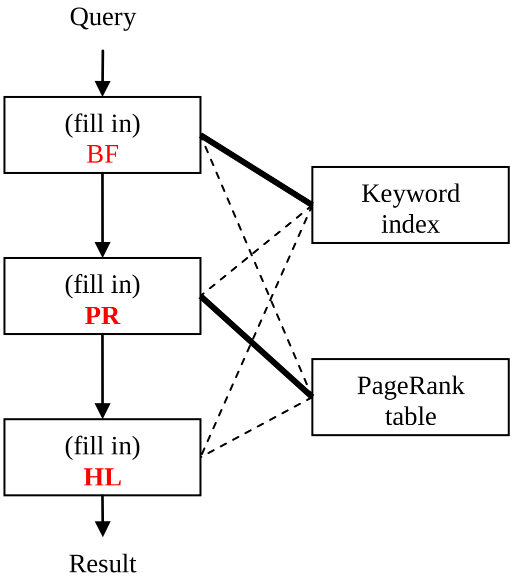
Average Precision (q) = [1*1+(2/3)*1+(3/4)*2+(4/7)*1+(5/10)*2]/5=0.948
MAP = AP(q)/1=0.948

DCG_{10} = 1 + (1/log3 + 2/log4 + 1/log7 + 2/log10)
$IDCG_{10}$ = 2 + (2/log2 + 1/log3 + 1/log4 + 1/log5)
$NDCG_{10}$ = DCG_{10} / $IDCG_{10}$

7. **[15]** The following diagram illustrates the building block of a Google-like search engine. PageRank table is a table containing the PR values of all web pages in the search engine. The dash lines indicate the candidate relations between the functional blocks. Fill in the three unlabelled boxes with the labels given on the left of the figure and draw SOLID LINES overlaid on the dash lines to identify the correct relations between the boxes and the index and PageRank table..

Fill in the boxes with the following labels:

BF: Identify documents meeting the query’s Boolean conditions
HL: Highlight matching keywords in the result
PR: Order pages by their PageRank values


Query
(fill in)
BF
(fill in)
PR
(fill in)
HL
Keyword index
PageRank table
Result