COMP 4321 Search Engine for Web and Enterprise Data
Final Examination, Spring 2021

Date: May 17, 2021
Exam Time: 12:30pm to 3:00pm (You can work on your answers during this period)
Upload Time: 3:00pm to 3:15pm (Your upload file cannot be modified in this period)
Time Allowed: 2.5 hours

Name: _____ Student ID: _____

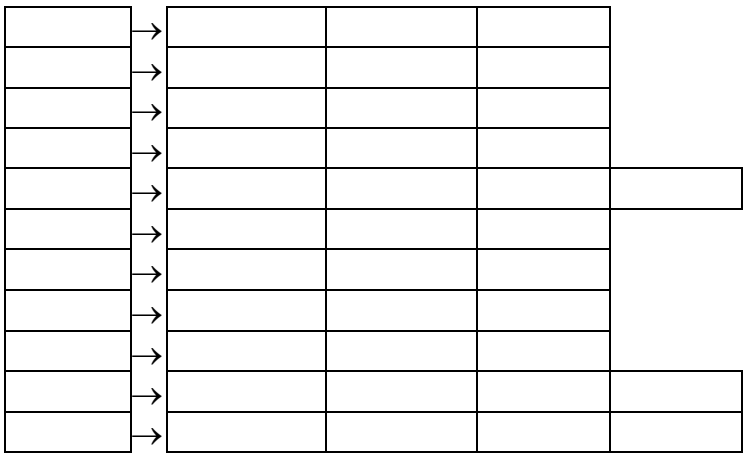
You can answer on separate pieces of paper and submit a scanned copy or photo of your answer. Or you can edit this question file to include your answer and submit a PDF file. In any case, you should **make your answer readable**.

1. [20] Given the following documents with docid D1 and D2, and preprocessing rules:

D1: The rain in Spain falls mainly on the plain.
D2: Where does the rain fall in Spain?

Preprocessing rules: (i) convert every letter to lowercase, (ii) remove special characters, (iii) convert plurals to singulars, (iv) no stop word removal.

(a) [7] Create the inverted index for the two documents and show it in the diagram below. Each postings contains the docid and positions. For example, the postings entry D1: 0,2,4 means the keyword is in D1 at word positions 0, 2 and 4. You can add or remove rows and cells to suit your answer.



(b) [3] You want to store the inverted index into a key-vale store. What would be stored in the "key" part and what would be stored in the "value" part?

(c) [2] Draw the forward index for the two documents; use any data structure you want.

(d) [8] Describe the procedure in words for computing the inner product between the two documents using the inverted and forward indexes. Document terms are weighted by term frequency (tf) only. Functions are available for you to extract data from these data structures, so you can just say "extract the tf …", "retrieve all entries matching …", etc. Your procedure should be efficient and should assume that there are many documents in the indexes. State any assumptions you have made.

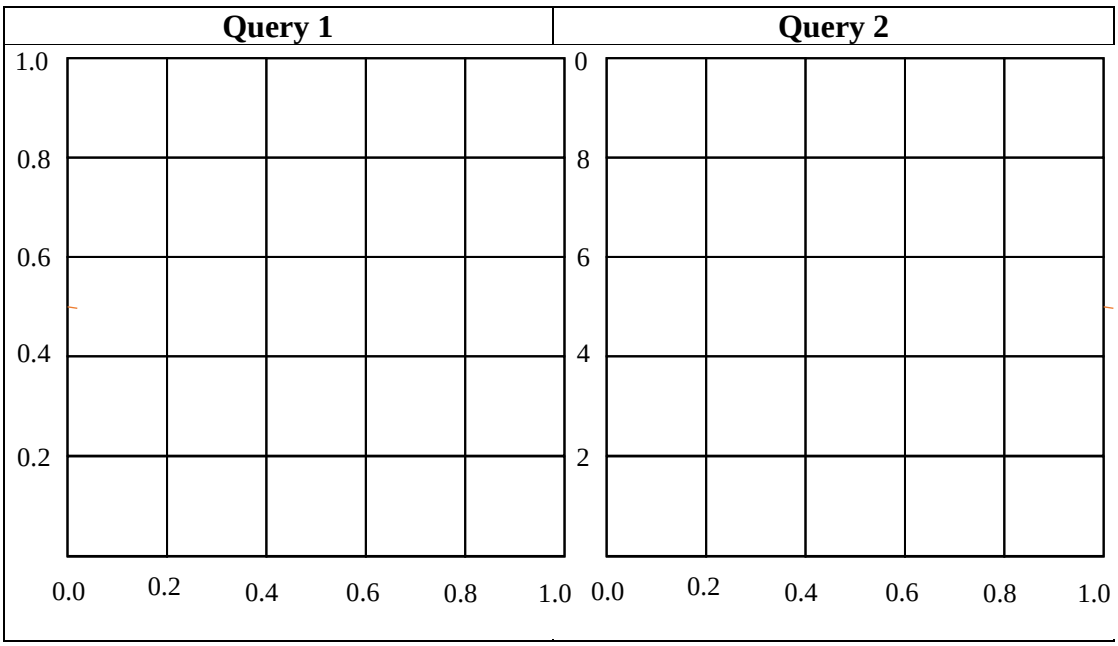
2. **[10]** The following tables show the ranked search results with relevance judgements of two queries. Assume that the two tables contain all of the documents retrieved. Plot the precision/recall of the two queries on the graphs provided below; show the graphs before interpolation with dash lines and the after-interpolation graph with solid lines. Compute MAP from the given data.

Query 1

Rank	Rel	P	R
1	No		
2	YES		
3	No		
4	YES		
5	No		

Query 2

Rank	Rel	P	R
1	YES		
2	No		
3	No		
4	YES		



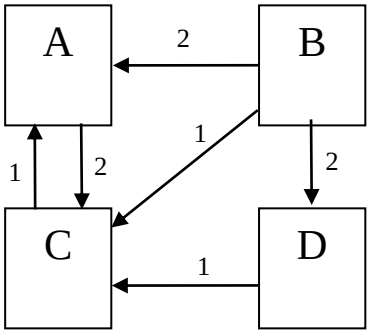
3. **[5]** Explain how you would process this query efficiently, where qt1 and qt2 are two query terms:

qt1 AND NOT qt2

Explain why the query with negation could be very expensive to process and how could you optimize the process.

4. **[5]** The Google's PageRank formula includes four factors that could affect the PageRank of a page, say, Page A. For each factor, give a rough interpretation in terms of the Random Surfer Model in no more than 3 lines.

5. **[10]** In the following web graph, each link has a weight. The PageRank of a node is passed to its children in proportion to the link weights of the node to its children. Using the weighted version of PageRank formula, compute the PageRank of the pages. The damping factor d=0.8. Use L1 normalization in each iteration.



(a) [4] Write down the PageRank formulas for the pages.

(b) [6] Compute the PageRank of the pages for the first 2 iteration.

Iteration	**0**	**1**	**Normalized PR for Iter. 1**	**2**	**Normalized PR for Iter. 2**
Page Rank (A)	1/4				
Page Rank (B)	1/4				
Page Rank (C)	1/4				

Page Rank (D)	1/4				

6. **[20]** Given the following pages and the weights of the terms they contain, the relevance of the pages indicated by the user is shown in the last column.

	Apple	**Orange**	**Relevance**
p1	0.4	0.2	Non-relevant
p2	0.2	0.9	Relevant
p3	0.7	0.4	Non-relevant
p4	0.8	0.3	Non-relevant
p5	0.9	0.7	Relevant

(a) [5] Plot the pages on a two-dimensional graph with Apple as the x-axis and Orange as the y-axis. Use ruler and pencil (or an editing tool), draw the best decision function you can observe and identify the support vectors. Briefly justify your choice is the best.

(b) [5] Write down your decision function as a linear equation. Scale the decision function so that points on the decision function return 0, points on the support vector of the relevant side returns 1 and points on the support vector of the non-relevant side returns -1.

(c) [2] Does the user prefer apple or orange? Justify briefly.

(d) [8] Explain how the decision function could be used personalizing the search results from a search engine.

7. **[5]** Describe one way to utilize collocation analysis (finding whether two words are often used together in writing) to improve search engine search quality. Since there could be many ways, your justification on how it could improve search quality is important.

8. **[5]** Give one major reason why WISE failed against PageRank in global web search engines, even though both of them utilize links in the search engine context. What would you expect WISE versus PageRank in the enterprise search environment?

9. **[5]** Give one reason for each of the following aspects for enterprise search to be more difficult than global web search engine (like google.com, bing.com)

Contents to be searched:	
Answers users want to get:	
Intranet environment	

10. **[10]**

(a) [3] Which of the following statements are true about the sentence graph of the graph-based summarization method TextRank.

T **F** a) The score of a node is the cosine similarity of the node to its adjacent nodes
T **F** b) The score of a node is the PageRank value of the node in the sentence graph
T **F** c) A link is added between two nodes if the sentences represented by the two nodes are adjacent in the article
T **F** d) The score of a node represents the vote the author casts on the importance of the sentence
e) Two sentences that are far apart in the article means the corresponding nodes are far away in the sentence graph

(i) (a) and (b) only
(ii) (c) and (e) only
(iii) (b) only
(iv) (b) and (d) only

(b) [7] In PageRank for web graph, a link from page *i* to page *j* is interpreted as a "vote" that page *i* gives to page *j* about *j*'s quality. In the graph-based summarization method TextRank:
(b.i) [2] How is the weight of a link between two sentences in a sentence graph defined?
(b.ii) [5] Why is a sentence with high PageRank in the sentence graph the best summary sentence?

11. **[5]**

(a) [2] In preference mining, we assume users exhibit the following behaviors:

T F a) The user must read the list of results from top to bottom
T F b) The user must click on pages relevant to his/her query
T F c) The user must read all of the results shown on the entire result page
T F d) The user must indicate a relevance rating for pages relevant to his/her query

(i) (b) and (d) only
(ii) (b) and (c) only
(iii) (a) and (b) only
(iv) (a) (b) and (c) only

(b) [3] Give two advantages of inferring user's preferences over two pages than inferring absolute relevance for search engine personalization.