National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Information Retrieval**	**Course Code:**	**CS 4051**
Program:	**BS (Data/Computer Science)**	**Semester:**	**Spring 2023**
Duration:	**60 mins**	**Total Marks:**	**15**
Paper Date:	**8-April-23**	**Weight**	**%**
Section:	**BDS-6A, BDS-6B, BCS-8A, BCS-8B**	**Page(s):**	**2**
Exam:	**Midterm 2 Exam**		

Instruction/Notes: Attempt the examination on the answer sheet in the same order as given in the here. Submit question paper with the answer sheet

Q1) (a) Given a query q_1, the set of documents relevant to the users is D^* = {d1, d10, d35, d40, d45, d55, d56}. The IR system retrieves the following documents D = {d10, d11, d35, d49, d45}. Leftmost document is top ranked. Compute Average Precision. [2 Marks]

k	Retrieved	R/N	Precision@Relevance	
1	d10	R	1	1
2	d11	N	-	-
3	d35	R	2/3	0.667
4	d49	N	-	-
5	d45	R	3/5	0.6

Average Precision=0.755

Q1) (b) [2 Marks] Presented with a list of documents in response to a search query, an experiment participant is asked to judge the relevance of each document to the query. Each document is to be judged on a scale of 0-3 with:

0 not relevant
3 highly relevant, and
1 and 2 "somewhere in between".

Compute Normalized Discounted Cumulative Gain (NDCG) for the following IR systems
IR System: [2, 3, 0, 1, 2]

i	rel_i	$log_2(i+1)$	$\frac{rel_r}{log_2(r+1)}$
1	2	1	2
2	3	1.585	1.893
3	0	2	0
4	1	2.322	0.431
5	2	2.585	0.774

$$DCG_k = \sum_{r=1}^{k} \frac{rel_r}{log(r+1)} = 2 + 1.893 + 0 + 0.432 + 0.774 = 5.097$$

$$NDCG_k = \frac{5.097}{5.097} = 1$$

Q1) (c) In Normalized Discounted Cumulative Gain (NDCG), we normalize the Discounted Cumulative Gain (DCG) for each topic with a normalizer. What is this normalizer? Why do we need to do this normalization step? Justify with an example. [2 Marks]

Example:
System 1: 3,2,3,0,1,2
System 2: 3,3,3,2,2,2,1,0

- Ideal DCG at 6 is (the best value) DCG for 3,3,3,2,2,2
- Normalize DCG with Ideal DCG value.
- NDCG for System 1 = DCG/IDCG =1.
- NDCG for System 2 = 0.785

Q2) Given the three-document corpus and a stop word list below, answer the following questions AFTER removing stopwords. [1 +2 = 3 Marks]

d_1	information retrieval is process of index search retrieval
d_2	retrieval is used for evaluation of search results retrieval retrieval
d_3	evaluation in information in evaluation process search
Q	information retrieval
Stopwords	is , of, in, for, to

Calculate similarity of document **d_2** with the query using maximum likelihood (language modeling) (use three document corpus given above).

a) No smoothing

b) Dirichlet Smoothing (mu = 4)

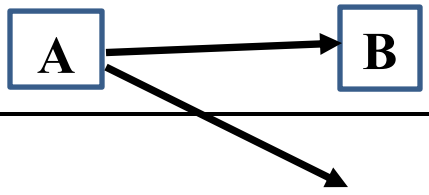
$$a)\ MLE_{d2,q} = \frac{freq(information, d_2)}{|d_2|} \times \frac{freq(retrieval, d_2)}{|d_2|} = \frac{0}{7} \times \frac{3}{7} = 0$$

$$b)\ MLE_{d2,q} = \frac{freq(information, d_2) + \frac{\mu.freq(information,C)}{|C|}}{\mu + |d_2|} \times \frac{freq(retrieval, d_2) + \frac{\mu.freq(retrieval,C)}{|C|}}{\mu + |d_2|}$$

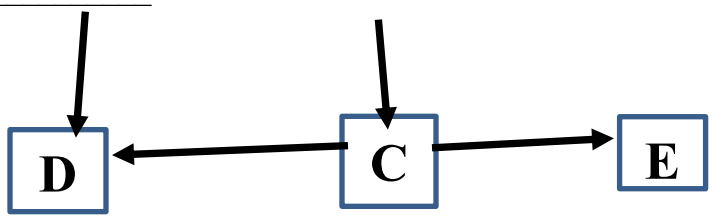
$$b)\ MLE_{d2,q} = \frac{0 + \frac{4\times2}{18}}{4+7} \times \frac{3 + \frac{4\times5}{18}}{4+7} = \frac{0.444}{11} \times \frac{4.111}{11} = 0.0404 \times 0.3737 = 0.0151$$

Q3 Only for Section BDS-6A and BDS-6B

Q 3)(a) Compute page rank of all nodes of following graph. Damping factor d = 0.9. Perform only two iterations of page rank algorithm. [4 Marks]



Name: _____ Reg #: _____ Section: _____



Q3)(b) Suppose that P, Q, and R are different web pages. Explain how it can happen that adding a link from P to Q can raise the PageRank of R. Explain how it can happen that adding a link from P to Q can lower the PageRank of R. In both cases, you should show a specific graph where this happens, though you need not work out the actual numerical values. [2 Marks]

Q3 Only for Section BCS-8A and BCS-8B

Q3) (a) [6 Marks]

First Word (w_1) \ Second Word (w_2)	baa	black	sheep	wolf
baa	20	30	15	0
black	5	0	40	0
sheep	0	0	0	20
wolf	0	0	0	0

i. Using Witten-Bell, find the bigram w_1w_2 probability of all unseen starting with "baa"
ii. Using Witten-Bell, find the bigram w_1w_2 probability of each unseen starting with "sheep"
iii. Calculate P(baa|baa) with add-k smoothing where k=0.75.

First Word (w_1) \ Second Word (w_2)	baa	black	sheep	wolf	Total N(w1)	Seen T(w1)	Unseen Z(w1)
baa	20	30	15	0	65	3	1
black	5	0	40	0	45	2	2
sheep	0	0	0	20	20	1	3
wolf	0	0	0	0	0	0	4

$$i. \frac{T(baa)}{N(baa)+T(baa)} = \frac{3}{65+3} = 0.0441$$

$$ii. \frac{T(sheep)}{N(sheep)+T(sheep)} \times \frac{1}{z(sheep)} = \frac{1}{20+1} \times \frac{1}{3} = 0.01587$$

$$iii. p(baa|baa) = \frac{C(baa\ baa)+K}{C(baa)+K|V|} = \frac{20+0.75}{65+(0.75\times 4)} = 0.3051$$