INFO-533 Homework 3 (100 points)

Deadline: TBD, 2024 11:59 pm in Brightspace

1. Use the below table data and formula to compute the **tf-idf weight** for each term in each document.

	Doc - 1	Doc - 2
Term 1	29	3
Term 2	5	30
Term 3	0	30

Document frequencies: Term1 = 15231, Term2 = 7043, Term3 = 18213
Formula: $w_{t,d} = (1 + \log_{10} tf_{t,d}) \ * \ \log_b (N / df_t)$
N = 780981
Compute for different bases b =2 and b=10
Comment on results from different bases, whether the base value affects the RELATIVE ranking of any two documents with respect to any given query if the Cosine similarity function is used. You may do this without computing the actual cosine similarity.

2. Given below are the formulas for computing term frequency weight of some term t in a document d. Consider only the case when $tf_{t,d}>=1$

Formula 1: $ltf_{t,d} = 1+\log tf_{t.d}$
Formula 2: $ntf_{t,d} = 0.4 + 0.6 * (tf_{t,d} / max\_tf_d)$, max_tf_d is maximum tf among all terms in d

Compare and analyze the properties of these two formulas and their relative advantages and disadvantages.

3. Give an example with a diagram to show that cluster pruning methods may sometimes fail to retrieve some actual top k results. Draw leaders and followers represented by dots and explain the relationship among them.

4. Given below is a list of ranked 15 documents out of which 5 documents are relevant for a given query Q. Calculate the Precision and Recall for each subset of j documents for j=1 to N (N=15). Assume these precisions as raw precisions and not interpolated precisions.

Rank	Document Id	Is Relevant?	Precision	Recall
1	doc1	No		
2	doc2	Yes		
3	doc3	No		
4	doc4	No		
5	doc5	No		
6	doc6	Yes		
7	doc7	No		
8	doc8	No		
9	doc9	Yes		
10	doc10	No		
11	doc11	No		
12	doc12	Yes		
13	doc13	No		
14	doc14	No		
15	doc15	Yes		

5. Using the values computed in the above table for recall and precision, calculate the interpolated precision and F1-value at each of the recall points mentioned in the below table. Additionally, draw the corresponding 11-point recall-interpolated precision curve.

Recall	Interpolated Precision	F1-value
0		
0.1		
0.2		
0.3		
0.4		

0.5		
0.6		
0.7		
0.8		
0.9		
1		