CS 620–Introduction to Data Science and Analytics, HW4

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

term	df_t
car	18,165
auto	6723
insurance	19,241
best	25,235

For questions 1, 2 and 3, show detailed calculations (not just the final answer).

1) (30 pts) Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, and Doc3 and the document frequency (df) of same terms in a document collection of 806,791 documents.
    a. Convert the raw term frequencies of car, auto, insurance and best using max frequency normalization (divide by tf of most common term in the document).
    b. Compute the idf weights for the terms car, auto, insurance, and best using given df in the second table (number of documents, N=806,791). Note: Use base 2 for log scale ($idf_t = \log_2(N/df_t)$).
    c. Calculate the tf-idf weights for the terms car, auto, insurance, best and create document vectors for each of the document where each vector has four components, one for each of the four terms.

2) (30 pts) Consider the queries Q1 = "best auto insurance" and Q2 = "top car insurance".
    a. Transform the queries into vector space using the same df values in the above table and calculate the tf-idf weights for the queries. (Note: DO NOT normalize the terms of this query when considering the tf values)
    b. Based on the document vectors calculated in question 1, rank the 3 documents for each query using the cosine similarity. Show detailed calculations for the cosine similarity.

3) (40 pts) Consider the transaction database in the table below:

tid	items
1	*a, b, d, f*
2	*a, d, e, f*
3	*b, c, d, e, f*
4	*b, d, f*
5	*b, e, f*
6	*c, d, f*
7	*c, e, f*
8	*b, d, e, f*

a) Show the candidate itemsets and the frequent itemsets in each level-wise pass of the Apriori algorithm at minimum support of 3.
b) Generate strong association rules from the frequent itemsets with minimum confidence of 80%

What to turn in:

Submit the answer to question 1, 2 and 3 in **Lastname-hw4.pdf** to Blackboard. Your pdf should contain the following information at the top:

CS620
HW4
@author:

Due: Sunday, Nov. 21, 2021, 11.59pm