




## CS 620–Introduction to Data Science, HW4

	Doc1	Doc2	Doc3	term	df <sub>t</sub>
car	27	4	24	car	18,165
auto	3	33	0	auto	6723
insurance	0	33	29	insurance	19,241
best	14	0	17	best	25,235

For the Questions (1) and (2), clearly show the intermediate calculations. Tip: Use an Excel sheet for the calculations.

- 1) Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, and Doc3 and the document frequency 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 (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) Consider the query “best car insurance”.
  - a. Transform the query into vector space using the same df values in the above table and calculate the tf-idf weights for the query without any normalization.
  - b. Based on the document vectors calculated in question 1, rank the 3 documents for the given query using cosine similarity.
- 3) Consider the 2 ranking algorithms in the figure below.
  - a. Calculate the confusion matrix values (tp, fp, tn, fn) for the position 7 in each ranking method.
  - b. Using the confusion matrix calculated above, compute the Accuracy and Harmonic Mean at position 7 for both ranking methods.
  - c. Calculate the Average Precision for each ranking algorithms and the Mean Average Precision (MAP) for both ranking methods.
  - d. Draw Precision-Recall Curves (using interpolation) for the Ranking #1 and #2 in a same graph and explain which ranking algorithm is better in terms of the Precision-Recall curves.

 = the relevant documents

											
Ranking #1											
Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6	
Ranking #2											
Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6	