CS 299–Introduction to Data Science, HW4

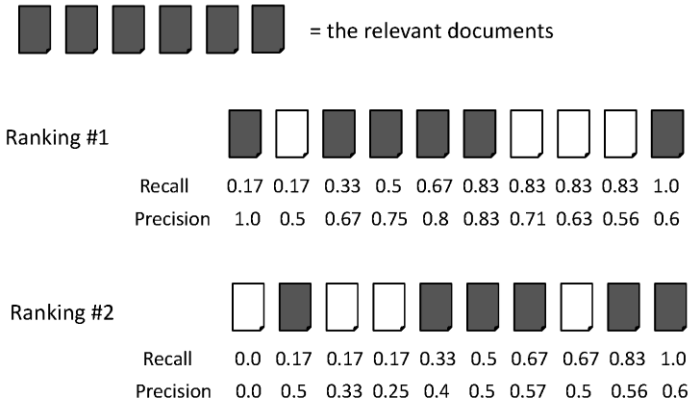
	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

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

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, best" using max frequency normalization (tf of most common term in the document).
    b. Compute the idf weights for the terms "car, auto, insurance, 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 table where each vector has four components per each document, 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 by calculating cosine similarity.

3) Consider the 2 ranking algorithms in the figure below.
    a. Calculate the confusion matrix values (tp, fp, tn, fn) for 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.



How to turn in: On paper at the beginning of the class. **Due: May 14, 2.00pm**
(Suggestion: You can use Microsoft Excel or similar application to calculate the answers and copy paste the work in each step to your assignment submission**)**