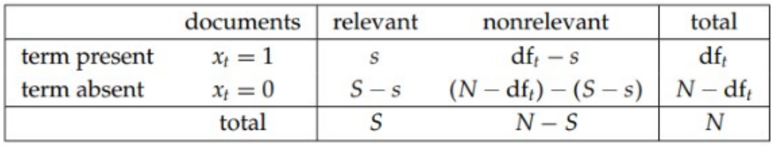
Quiz 2 IR Spring 2019 (26/03/19) Roll No--------------------------

Q1: Which of these feature weighting techniques is best to give lower weights to words that occurs very frequency in all documents? (Give one line reason) [2 points]

I. Binary occurrence
II. Term Frequency
III. Term count
IV. TF-IDF

Q2: What is the underlying assumption in Vector space model? (Answer in 1 to 2 lines) [2 points]

Q3: Consider the following contingency table and retrieval status value (RSV) formula. * [2 points]



	documents	relevant	nonrelevant	total
term present	$x_t = 1$	s	$\mathrm{df}_t - s$	df_t
term absent	$x_t = 0$	$S - s$	$(N - \mathrm{df}_t) - (S - s)$	$N - \mathrm{df}_t$
	total	S	$N - S$	N

$$RSV_d = \sum_{t \in q} \left[\; c_t \; \times \frac{(k_1 + 1)\mathrm{tf}_{td}}{k_1((1-b) + b(L_d/L_{\mathrm{ave}})) + \mathrm{tf}_{td}} \times \frac{(k_3 + 1)\mathrm{tf}_{tq}}{k_3 + \mathrm{tf}_{tq}} \right]$$

Where c_t is

$$c_t = K(N, \mathrm{df}_t, S, s) = \log \frac{s/(S-s)}{(\mathrm{df}_t - s)/((N - \mathrm{df}_t) - (S - s))}.$$

Change the above formula if we assume that all the document containing term t are relevant.

*df$_t$= document frequency of term t i.e. the number of document in which t occurstf$_{td}$: Term Frequency of t in document d
L_d : length of document d L$_{ave}$= Average length of doumentstf$_{tq}$: Term Frequency of term t in query q