CS 172 Handout
Probabilistic Retrieval Model

We are interested in using the following document-term matrix and the associated relevance information as training data for a probabilistic retrieval model. A 1 entry indicates that the term occurs in a document, and 0 means it does not. Also, we have a column indicating R or NR which indicates the relevance of the document with respect to queries in the training data that was provided.

	T_1	T_2	T_3	T_4	T_5	T_6	Relevance
D_1	1	0	1	1	0	0	R
D_2	0	1	0	1	0	1	R
D_3	1	0	1	1	1	0	NR
D_4	0	1	1	0	1	1	NR
D_5	1	1	0	1	0	0	NR
D_6	1	1	0	1	1	1	NR
D_7	0	0	0	0	0	1	R
D_8	0	0	1	1	1	0	NR
D_9	1	1	1	0	1	1	R
D_{10}	1	0	0	1	1	0	NR

Using the basic probabilistic retrieval mode, compute the relevance and non-relevance probabilities associated with terms T1 through T6.

	R	**NR**
T_1	2/4	4/6
T_2	2/4	3/6
T_3	2/4	3/6
T_4	2/4	5/6
T_5	1/4	5/6
T_6	3/4	2/6

Then using these probabilities and the given query Q = (1,1,0,1,0,1), compute the discriminant Disc (Q, D_{11}) and Disc (Q, D_{12}) for each of the following new documents:

D_{11} = (0,1,1,0,0,1)
D_{12} = (1,0,1,1,0,1)

Based on the discriminants, should these documents be retrieved? Explain your answer

$$\text{Disc}(Q, D_{11}) = \frac{P(D_{11}|R)P(R)}{P(D_{11}|NR)P(NR)} =$$

$$\frac{\left(1-P(t_1|R)\right)*P(t_2|R)*(1-P(t_4|R))*P(t_6|R)*P(R)}{\left(1-P(t_1|NR)\right)*P(t_2|NR)*(1-P(t_4|NR))*P(t_6|NR)*P(NR)} = \frac{(1-\frac{2}{4})(\frac{2}{4})(1-\frac{2}{4})(\frac{3}{4})(4/10)}{(1-4/6)(3/6)(1-\frac{5}{6})(2/6)(6/10)} =$$