Probability Ranking Principle

D={ d1, d2, d3,…d10} (Fixed Collection)

Q= q1 (Fixed Query)

Retrieval /Relevance

	Rel	NRel
Ret	3	2
NRet	1	4

P(Rel)=4/10=.4

P(NRel)=6/10= .6

P(Ret)=5/10 = 0.5

P(NRet)=5/10 = 0.5

P(dk/Rel,Ret) =3/5 =0.6

P(dk/NRel,Ret)=2/5=0.4

P(dk) =0.1

For the same query a new document dk is to be ranked.

P(Rel/dk)= { P(dk/Rel)* P(Rel)} / p(dk) = (.4*.1)*.4 / .1 = 0.16

P(NRel/dk) ={P(dk/NRel)*P(NRel)} /p(dk) = (0.6* 0.1)* 0.6 / 0.1 = 0.36

As P(NRel/dk) > P(Rel/dk) hence it is Non-Relevant.

Posterior probabilities are now:

P(dk/Rel) = P(Rel/dk) * P(Rel) / P(dk) = (0.16 * .4) /0.1= 0.64

P(dk/NRel) =P(NRel/dk)* P(NRel) / P(dk) = (0.36 * 0.6) / 0.1 = 0.116

We will be more accurate if we have P(Rel/dk,User profile, query, context ,..etc)

Ranking of Document through PRP when relevance is given.

Example:
Consider the following given document-collection and a query.

d1: virus microscopic organism
d2: virus infects cell organism
d3: virus infects computers
d4: tiny virus security
q: virus tiny organism

From the system d1 and d2 are relevant to this query and d3 and d4 are not. Using the probabilistic model for IR with given relevance, rank these documents using probability ranking principle. Show all intermediates steps of calculations. Compute whether the document **"d5: virus computer virus"** is relevant or not.

Words	**cell**	**computers**	**infects**	**microscopic**	**organism**	**security**	**tiny**	**virus**
$N_1(W)$	1	0	1	1	2	0	0	2
	0.5	0.1667	0.5	0.5	0.833	0.1667	0.1667	0.833
$N_0(W)$	0	1	1	0	0	1	1	2
	0.1667	0.5	0.5	0.1667	0.1667	0.5	0.5	0.833

Let **N** be the number of documents in the collection. Here N=4
Let **R** be the number of relevant documents in collection. Here R=2
Let $\mathbf{n_t}$ is the number of documents contains term t.
Let $\mathbf{r_t}$ is the number of relevant documents contains term t.
Let p(t) is the probability of term for relevant document. We used $N_1(w)$ for it.
Let u(t) is the probability of term for non-relevant document. We used $N_0(w)$ for it.

$N_1(w)$ is computed as, $N_1(w) = (r_t + 0.5) / (R+1)$

$N_0(w)$ is computed as, $N_0(w) = (n_t - r_t + 0.5) / (N-R+1)$

Now, we want to rank all documents with the probability of relevance with the query. P(R=1/ di,q) for i=1,2,3, and 4.

RSV for d1 = P(R=1/d1,q) = rank for all terms common in d1 and q $[(p_t/1-p_t) * (1-u_t/u_t)]$

RSV for d1= P(R=1/ d1,q) $=^{rank}$ [(0.833/0.1667)X(0.833/0.1667)]X0.5 = 12.58

RSV for d2= P(R=1/ d2,q) $=^{rank}$ (0.833/0.1667)X(0.833/0.1667)X0.5 = 12.58

RSV for d3= P(R=1/ d3,q) $=^{rank}$ (0.833/0.1667)X(0.1667/0.833)X0.5 = 0.5

RSV for d4= P(R=1/ d4,q) $=^{rank}$ (0.1667/0.833 x 0.5/0.5) X(0.833/0.1667 X 0.1667/0.833)X0.5 = 0.09

Hence ranking will be either d1, d2, d3 and d4 or d2, d1, d3 and d4.

Now, knowing the class of d5;

P(R=1/ d5,q) = ((0.833/0.1667)X(0.1667/0.833))X0.5 = 0.5

P(R=0/ d5,q) = ((0.1667/0.833)X (0.833/0.1667))X0.5 = 0.5

P(R=1/ d5,q) is not greater than P(R=0/ d5,q) hence d5 is non-relevant.

Ranking of Document through PRP when relevance is not given.

Example:
Consider the following given document-collection and a query.

d1: virus microscopic organism
d2: virus infects cell organism
d3: virus infects computers
d4: tiny virus security
q: virus tiny organism

Using the probabilistic model for IR with given relevance, rank these documents using probability ranking principle. Show all intermediates steps of calculations. Compute whether the document **"d5: virus computer virus"** is relevant or not.

Words	**cell**	**computers**	**infects**	**microscopic**	**organism**	**security**	**tiny**	**virus**
N(W)=	1	1	2	1	2	1	1	4
	3.5/5	3.5/5	2.5/5	3.5/5	2.5/5	3.5/5	3.5/5	0.5/5

N(W) = (N – Nw + 0.5) / (N + 1) where w is the term from the data.

Now, we want to rank all documents with the probability of relevance with the query. P (R=1/ di,q) for i=1,2,3, and 4.

RSV for d1= P(R=1/ d1,q) $=^{rank}$ (0.5/5) X (2.5/5) = 0.05

RSV for d2= P(R=1/ d2,q) $=^{rank}$ (0.5/5) x (2.5/5) = 0.05

RSV for d3= P(R=1/ d3,q) $=^{rank}$ (0.5/5) =0.1

RSV for d4= P(R=1/ d4,q) $=^{rank}$ (0.5/5) x (3.5/5) = 0.07

Hence ranking will be d3,d2,d1,and d4 or d3,d1,d2,and d4
Now, knowing the relevance of d5;

P(R=1/ d5,q) = (0.5/5)X(0.5/5)x (3.5/5) = 0.07

It seems marginally relevant. As P(R=1/ d5,q) is not very high.

N(w) is similar to tf*idf if we take P(t) for w, as 1/2 and Q(t) for w, as |D| / dft
pt reflects that we have no information about relevant documents
qt under the assumption that nos. of relevant documents are much much lesser than # documents