

# RSJ-PM Tutorial: A Tutorial on the Robertson-Sparck Jones Probabilistic Model for Information Retrieval

Dr. E. Garcia  
[admin@miislita.com](mailto:admin@miislita.com)

*Abstract* – This is a tutorial on the original Robertson-Sparck Jones Probabilistic Model. The model is based on Independence Assumptions and Ordering Principles for probable relevance and does not incorporate term frequency.

**Keywords:** probabilistic model, independence assumptions, ordering principles, idf, inverse document frequency

All Rights Reserved © 2009 E. Garcia; First Published: March 30, 2009; Last Modified: April 4, 2009.

## Introduction

In 1976, Stephen Robertson and Karen Sparck Jones proposed a probabilistic model for information retrieval based on two assumptions and principles:

### 1. Independence Assumptions:

- **I1** – The distribution of terms in relevant documents is independent and their distribution in all documents is independent.
- **I2** – The distribution of terms in relevant documents is independent and their distribution in non-relevant documents is independent.

### 2. Ordering Principles:

- **O1** – Probable relevance is based *only* on the presence of query terms in the documents.
- **O2** – Probable relevance is based on *both* the presence and absence of query terms in the documents.

Table 1 summarizes these.

		Independence Assumptions	
		<b>I1</b>	<b>I2</b>
Ordering Principles	<b>O1</b>	F1	F2
	<b>O2</b>	F3	F4

**Table 1. Assumptions-Principles Contingency Table.**

In Table 1, F1 through F4 are weighting functions. According to Robertson and Sparck Jones (1976), **I2** is more realistic than **I1** while **O2** is correct and **O1** is incorrect. The model predicts that F4 is most likely to yield the best results and is therefore the best of these functions. The purpose of this tutorial is to show you how these functions can be used to rank documents in the presence and absence of relevance information.

## Discussion

**I1** states that the presence of a term in a relevant document does not impact the presence of other terms in the same document or its presence in other relevant documents. **I1** says nothing about the distribution of terms in non-relevant documents.

**I2** extends **I1** to non-relevant documents by stating that the presence of a term in a non-relevant document does not impact the presence of other terms in the same document or its presence in other non-relevant documents. Since documents are either relevant or non-relevant to a query, this is why **I2** is more realistic than **I1**.

**O1** indicates that documents should be ranked only if they contain all of the terms specified in a query. It is an AND approach. **O1** says nothing about the absence of query terms in the documents and is therefore incorrect.

**O2** takes **O1** a little further and says that we should consider both the presence and absence of query terms. It is an OR approach. Accordingly, for a query consisting of two terms  $t_1$  and  $t_2$ , documents mentioning both terms should rank higher than those mentioning one or none of these terms.

To implement **O2**, a system using an inverted index would have to identify all terms present and not present in a document. To avoid exhaustively tracking the inverted index, we can assign zero probability of relevance to documents lacking of all the query terms. Adopting this strategy implies that we have some evidence of non-relevance. It also has the effect of artificially converting **O2**-based weights to presence-only **O1** weights. This makes **O2** more practical than **O1**.

Based on Table 1, Robertson and Sparck Jones introduced explicit expressions for  $F_1$ ,  $F_2$ ,  $F_3$ , and  $F_4$ .

## Derivation

Given a query consisting of a term and a collection of documents, the Robertson-Sparck Jones Probabilistic Model (RSJ-PM) addresses two cardinal questions:

**Is the term present in the documents?**      Answer: 1 = Yes (Present), 0 = No (Absent).

**Are the documents relevant to the term?**      Answer: 1 = Yes (Relevant), 0 = No (Non-Relevant).

This binary treatment can be extended to queries consisting of several terms. To simplify, in this tutorial we limit the discussion to one-term queries. We begin by stating the following definitions:

$r$	=	number of relevant documents that contain the term.
$n - r$	=	number of non-relevant documents that contain the term.
$n$	=	number of documents that contain the term.
$R - r$	=	number of relevant documents that do not contain the term.
$N - n - R + r$	=	number of non-relevant documents that do not contain the term.
$N - n$	=	number of documents that do not contain the term.
$R$	=	number of relevant documents.
$N - R$	=	number of non-relevant documents.
$N$	=	number of documents in the collection.

Next, a contingency table is constructed.

		Are the documents relevant to the term?		
		1 = Yes (Relevant)	0 = No (Non-Relevant)	Collection-wide Incidence
Is the term present in the documents?	1 = Yes (Present)	r	n - r	n
	0 = No (Absent)	R - r	N - n - R + r	N - n
Total number of documents		R	N - R	N

**Table 2. Contingency Table.**

Normalizing Table 2 elements over the number of documents per columns, a table of probabilities is obtained.

		Probabilities		
		1 = Yes (Relevant)	0 = No (Non-Relevant)	Collection-wide Incidence
Probabilities	1 = Yes (Present)	r/R	(n - r)/(N-R)	n/N
	0 = No (Absent)	(R - r)/R	(N - n - R + r)/(N - R)	(N - n)/N

**Table 3. Table of Probabilities.**

Taking probability ratios, a Table of Odds is computed.

		Odds		
		1 = Yes (Relevant)	0 = No (Non-Relevant)	Collection-wide Incidence
Odds		r/(R - r)	(n - r)/(N - n - R + r)	n/(N - n)

**Table 4. Table of Odds.**

Reading from left to right, the ratios in Tables 3 and 4 are defined as follows:

$r/R$	=	probability that a relevant document contains the term.
$(n - r)/(N - R)$	=	probability that a non-relevant document contains the term.
$n/N$	=	probability that a document contains the term.
$(R - r)/R$	=	probability that a relevant document does not contain the term.
$(N - n - R + r)/(N - R)$	=	probability that a non-relevant document does not contain the term.
$(N - n)/N$	=	probability that a document does not contain the term.
$r/(R - r)$	=	odds that a relevant document contains the term.
$(n - r)/(N - n - R + r)$	=	odds that a non-relevant document contains the term.
$n/(N - n)$	=	odds that a document contains the term.

We now do some collection-wide and distribution-specific comparisons. The fraction of relevant documents containing the term ( $r/R$ ) is compared in two different ways:

- Against the fraction of documents in the collection containing the term; i.e., ( $n/N$ ).
- Against the fraction of non-relevant documents containing the term; i.e.,  $(n - r)/(N - R)$ .

Likewise, the odds that relevant documents contain the term ( $r/(R - r)$ ) is compared in two different ways:

- Against the odds that documents from the collection contain the term; i.e.,  $n/(N - n)$ .
- Against the odds that non-relevant documents contain the term; i.e.,  $(n - r)/(N - n - R + r)$ .

To account for the fact that term weights are additive we take logarithms. This yields explicit expressions for the four weighting functions given in Table 1. These are summarized in Table 5.

Weighting Function	Remarks
$F1 = \log \left[ \frac{(r/R)}{(n/N)} \right]$	F1 evaluates the ratio of the proportion of relevant documents in which the term occurs to the proportion of the entire collection in which it occurs.
$F2 = \log \left[ \frac{(r/R)}{((n - r)/(N - R))} \right]$	F2 evaluates the ratio of the proportion of relevant documents to that of non-relevant documents.
$F3 = \log \left[ \frac{(r/(R - r))}{(n/(N - n))} \right]$	F3 evaluates the ratio between the “relevance odds” for the term (i.e., the ratio between the number of relevant documents in which it does occur and the number in which it does not occur) and the “collection odds” for the term.
$F4 = \log \left[ \frac{(r/(R - r))}{((n - r)/(N - n - R + r))} \right]$	F4 evaluates the ratio between the term relevance odds and its “non-relevance odds”.

**Table 5. Table of Scoring Functions.**

In Table 5, F1 through F4 are scoring functions that evaluate the weight of term  $i$ ,  $w(t_i)$ , as log transformations. These comparisons and transformations are not arbitrary. Let see why.

In the absence of relevance information the only information available is collection-wide incidence: the  $n/N$  and  $n/(N - n)$  ratios. It seems intuitively correct to propose scoring functions that use the  $n/N$  and  $n/(N - n)$  ratios as reference points. By doing so, we are effectively comparing against collection-wide proportions.

If we recall,  $\log(N/n)$  is the so-called Inverse Document Frequency (IDF) and  $\log((N - n)/n)$  is its “odds version” also known as IDF Probabilistic (IDFP). Considering these as weighting functions we can write

$$F0 = \log \left[ \left( \frac{N}{n} \right) \right] = \text{IDF} \quad (\text{Eq. 1})$$

$$F00 = \log \left[ \left( \frac{N - n}{n} \right) \right] = \text{IDFP} \quad (\text{Eq. 2})$$

As weighting functions, F0 and F00 evaluate the weight of term  $i$ ,  $w(t_i)$ , but with one caveat: without incorporating relevance information. These are just collection-wide estimators of the discriminatory power of a term (*term specificity*). Indeed according to Robertson (2004), an IDF value is an RSJ weight in the absence of relevance information. This is also true for IDFP. From F1 and F3, it is evident that

$$F1 = \log \left[ \left( \frac{r}{R} \right) \right] + \text{IDF} = \log \left[ \left( \frac{r}{R} \right) \right] + F0 \quad (\text{Eq. 3})$$

$$F3 = \log \left[ \frac{r}{(R-r)} \right] + \text{IDFP} = \log \left[ \frac{r}{(R-r)} \right] + F00 \quad (\text{Eq. 4})$$

That is, F1 and F3 compensate for the lack of relevance information in IDF and IDFP weights by adding to these a relevance component.

Note that F1 and F3 are related by comparing the relevant document distribution of a term to its entire collection distribution. In the case of F2 and F4, these functions are related by comparing relevant and non-relevant distributions. It is possible to derive IDF and IDFP from F2 and F4 by making specific assumptions about the degree of relevance information available. For instance, IDFP can be obtained by setting  $R = r = 0$  in F4 (Robertson, 2004).

Regarding the use of logarithms, we must remember that these are additive: the log of a product is a sum of logs. This additive property is frequently assumed in IR with term matching coefficients (Robertson, 2004, Robertson & Sparck Jones, 1976).

## Using the Model Predictively

The RSJ model can be used in two different ways: retrospectively and predictively. According to Robertson and Sparck Jones, if the model is used retrospectively, the use of proportions as estimates is recommended.

However if the model is used predictively, it will breakdown when  $n, r, N, R, n-r, N-n$ , or  $N-R = 0$ . This can be avoided by adding a correction factor  $k$  to the entries of Table 2. See Table 6.

		Are the documents relevant to the term?		
		1 = Yes (Relevant)	0 = No (Non-Relevant)	Collection-wide Incidence
<b>Is the term present</b>	1 = Yes (Present)	$r + k$	$n - r + k$	$n + 2k$
<b>in the documents?</b>	0 = No (Absent)	$R - r + k$	$N - n - R + r + k$	$N - n + 2k$
Total number of documents		$R + 2k$	$N - R + 2k$	$N + 4k$

**Table 6. Contingency Table with Correction Factor  $k$ .**

Using the model predictively means making inferences about the probabilities on the basis of sample information available. This is problematic for small samples. In their original paper, the authors used  $k = 0.5$  (the so-called *point-5 correction*), obtaining the scoring functions depicted in Table 7.

RSJ Predictive Functions		$k = 0.5$	
$F1 = \log \left[ \frac{(r + k)/(R + 2k)}{((n + 2k)/(N + 4k))} \right]$		$F1 = \log \left[ \frac{(r + 0.5)/(R + 1)}{((n + 1)/(N + 2))} \right]$	
$F2 = \log \left[ \frac{(r + k)/(R + 2k)}{((n - r + k)/(N - R + 2k))} \right]$		$F2 = \log \left[ \frac{(r + 0.5)/(R + 1)}{((n - r + 0.5)/(N - R + 1))} \right]$	
$F3 = \log \left[ \frac{((r + k)/(R - r + k))}{((n + 2k)/(N - n + 2k))} \right]$		$F3 = \log \left[ \frac{((r + 0.5)/(R - r + 0.5))}{((n + 1)/(N - n + 1))} \right]$	
$F4 = \log \left[ \frac{((r + k)/(R - r + k))}{((n - r + k)/(N - n - R + r + k))} \right]$		$F4 = \log \left[ \frac{((r + 0.5)/(R - r + 0.5))}{((n - r + 0.5)/(N - n - R + r + 0.5))} \right]$	

Table 7. RSJ Model with correction factor  $k$ .

### A Working Example

Robertson and Sparck Jones (1976) applied their model to a collection of 200 documents of which 5 were relevant to terms  $a$ ,  $b$ ,  $c$ ,  $d$ , and  $e$ . In Table 8 we have reproduced their results. For comparison purposes we computed results for  $F0$  and  $F00$ . Results were computed using both versions of the model.

Retrospective Mode ( $k = 0$ )						Predictive Mode ( $k = 0.5$ )					
$a$	N	R	n	r		$a$	N	R	n	r	
	200	5	5	1			200	5	5	1	
F0	F00	F1	F2	F3	F4	F0	F00	F1	F2	F3	F4
1.60	1.59	0.90	0.99	0.99	1.08	1.53	1.51	0.93	1.04	1.04	1.15
$b$	N	R	n	r		$b$	N	R	n	r	
	200	5	5	4			200	5	5	4	
F0	F00	F1	F2	F3	F4	F0	F00	F1	F2	F3	F4
1.60	1.59	1.51	2.19	2.19	2.89	1.53	1.51	1.40	1.99	1.99	2.59
$c$	N	R	n	r		$c$	N	R	n	r	
	200	5	100	1			200	5	100	1	
F0	F00	F1	F2	F3	F4	F0	F00	F1	F2	F3	F4
0.30	0.00	-0.40	-0.40	-0.60	-0.62	0.30	0.00	-0.30	-0.31	-0.48	-0.49
$d$	N	R	n	r		$d$	N	R	n	r	
	200	5	100	4			200	5	100	4	
F0	F00	F1	F2	F3	F4	F0	F00	F1	F2	F3	F4
0.30	0.00	0.20	0.21	0.60	0.62	0.30	0.00	0.18	0.18	0.48	0.49
$e$	N	R	n	r		$e$	N	R	n	r	
	200	5	20	3			200	5	20	3	
F0	F00	F1	F2	F3	F4	F0	F00	F1	F2	F3	F4
1.00	0.95	0.78	0.84	1.13	1.20	0.98	0.94	0.75	0.82	1.08	1.15

Table 8. RSJ weights for five terms ( $a$ ,  $b$ ,  $c$ ,  $d$ , and  $e$ ) with  $k = 0$  and  $k = 0.5$ .

In Table 9 we have reordered these results with respect to the  $r/n$  ratio (the probability that documents containing the term are relevant).

		R/N (%) = 2.5			Results with k = 0						Results with k = 0.5						
	r	n	r/N (%)	n/N (%)	r/n (%)	F0	F00	F1	F2	F3	F4	F0	F00	F1	F2	F3	F4
<i>c</i>	1	100	0.5	50	1	0.30	0.00	-0.40	-0.40	-0.60	-0.62	0.30	0.00	-0.30	-0.31	-0.48	-0.49
<i>d</i>	4	100	2	50	4	0.30	0.00	0.20	0.21	0.60	0.62	0.30	0.00	0.18	0.18	0.48	0.49
<i>e</i>	3	20	1.5	10	15	1.00	0.95	0.78	0.84	1.13	1.20	0.98	0.94	0.75	0.82	1.08	1.15
<i>a</i>	1	5	0.5	2.5	20	1.60	1.59	0.90	0.99	0.99	1.08	1.53	1.51	0.93	1.04	1.04	1.15
<i>b</i>	4	5	2	2.5	80	1.60	1.59	1.51	2.19	2.19	2.89	1.53	1.51	1.40	1.99	1.99	2.59

**Table 9. Reordering of Table 8 results with respect to the  $r/n$  ratio.**

To understand better these results, we have also computed the  $R/N$  and  $n/N$  ratios. It is clear that:

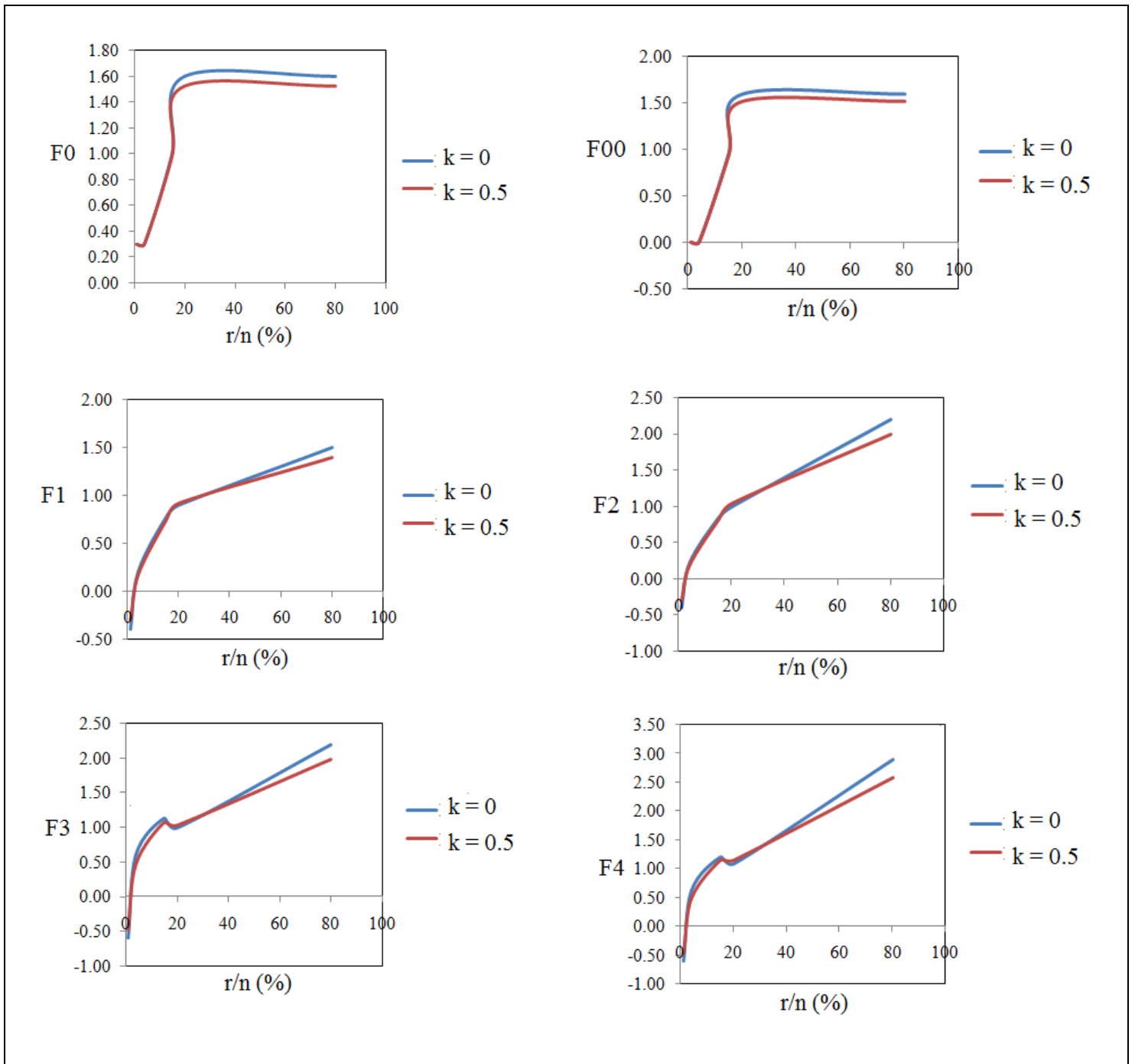
- A weight of zero is obtained when  $R/N = r/n$ . In addition, the theory predicts that  $F00$  should give a zero IDFP weight when  $n/N = 0.5$  and a negative weight when  $n/N > 0.5$ .
- For  $r/n < R/N$  all four functions ( $F1$  through  $F4$ ) give a negative weight to  $c$  since documents chosen at random from those containing the term  $c$  is less likely to be relevant than one chosen at random from the whole collection.
- As  $r/n$  increases and  $n/N$  decreases,  $F1$  through  $F4$  separate terms as expected:  $b > a$  and  $d > c$ .
- The relationship of  $a$  and  $e$  shows that the four functions do not necessarily rank terms in the same order.
- With  $k = 0$ ,  $F2$  and  $F3$  assign the same weight to  $a$  (0.99) and to  $b$  (2.19). This is also observed when  $k = 0.5$ .
- $F4$  assigns higher weights than  $F1$ ,  $F2$ , and  $F3$ .
- $F0$  and  $F00$  assign higher weights than  $F4$  when both  $r/N$  and  $r/n$  are small (for  $a$ , these respectively are 0.5 % and 2.5 %).

## Revisiting k

Robertson and Sparck Jones adopted the idea of using the correction factor  $k$  from Cox (Robertson & Spark Jones, 1976; Cox, 1970). From the above discussion, it is still unclear the role of  $k$  in their information retrieval model. In particular,

- What is the effect of varying  $k$  for a given weighting function across terms?
- What is the effect of varying  $k$  for a given term across weighting functions?

To address these questions, in Figure 1 we inspected the effect of varying  $r/n$  for each of the scoring functions. These were inspected at  $k = 0$  and  $k = 0.5$ .

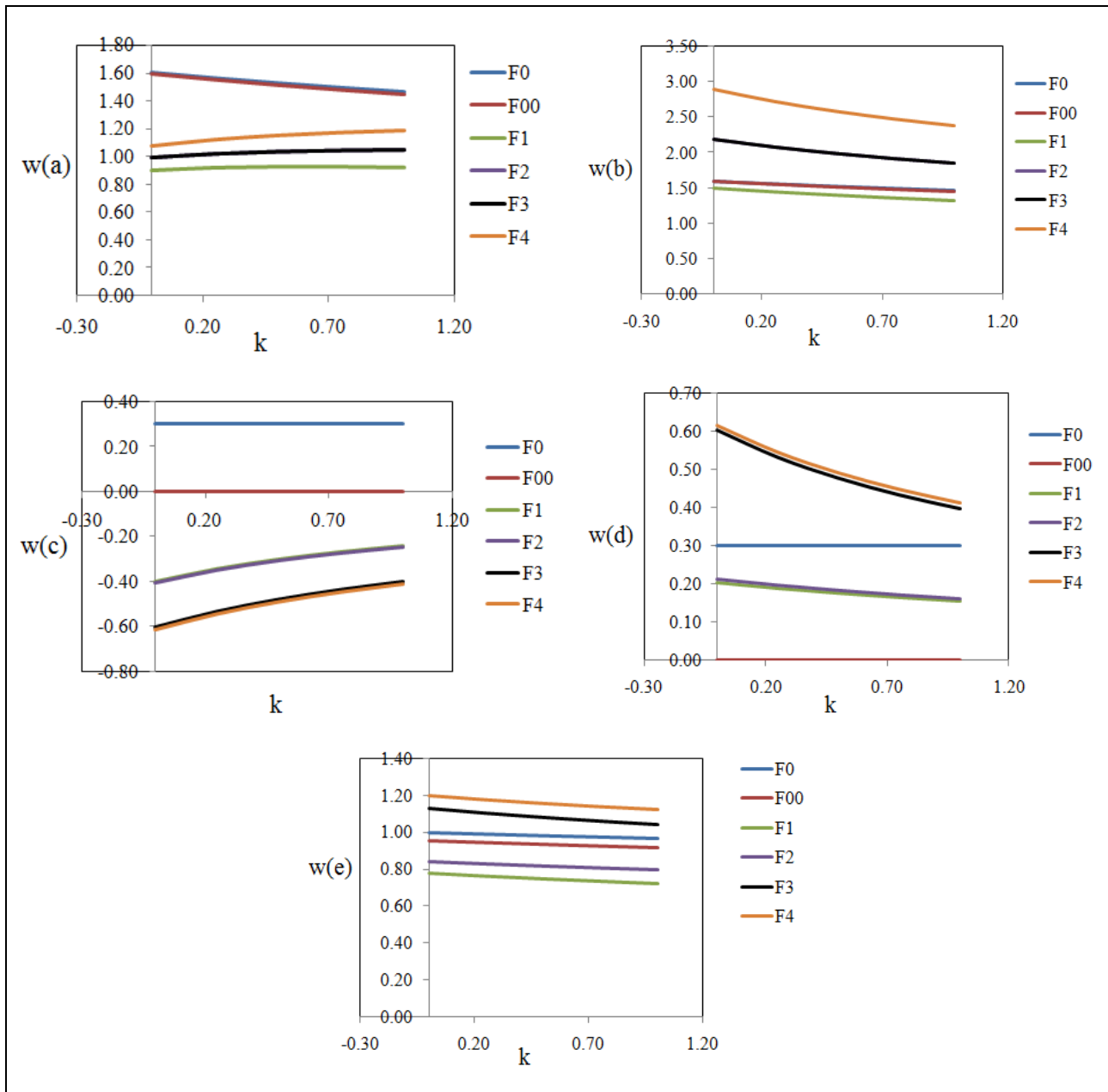


**Figure 1. Profile curves of term weights ( $w$ ) vs.  $r/n$  ratios at  $k = 0$  and  $k = 0.5$  for  $a$ ,  $b$ ,  $c$ ,  $d$ , and  $e$  terms.**

Note that varying  $k$  from 0 to 0.5 essentially dampens down the curves. These results confirm the generalized perception that  $k$  is a smoothing correction.

In Figure 2 we examined the effect of varying  $k$  for each of the scoring functions. In the figure,  $w(a)$  stands for the weight assigned to  $a$ ,  $w(b)$  stand for the weight assigned to  $b$ , and so forth.

Again,  $k$  acts as a smoothing correction. However, it should be underscored that setting  $k$  does impact the scoring functions in a non-trivial way. Figure 2 shows that curve slopes differ across terms and scoring functions. This is a reflection of the several combinations of relevance/non-relevance documents used.



**Figure 2. Profile curves for  $a$ ,  $b$ ,  $c$ ,  $d$ , and  $e$  terms showing term weights ( $w$ ) for several values of  $k$ .**

The absolute values of the slopes are indicative of how sensitive the scoring functions are to  $k$ . Note that some curves overlap and are not visible while others are well discernible. When they do overlap, selecting one scoring function over the other for a particular  $k$  does not matter that much.

Last but not least when function curves are orthogonal to the  $y$  axis, using these predictively or retrospectively should return identical results. This is the case of  $F0$  and  $F00$  with terms  $c$  and  $d$ . Can you guess why? (Hint: See Table 9.)

Note that  $F4$  gives higher weights to  $b$ ,  $d$ , and  $e$ ; i.e. to terms with a high relevant document incidence,  $r$ . We can extend on this subject and argue that varying  $k$  does provide some insight as to when and why some functions assign lower or higher weights. Such a discussion is a great homework and complementary research work for this tutorial.

## Exercises

1. The following example is taken from *Information Retrieval: Algorithms and Heuristics* (Grossman & Frieder, 2004). Let Q be a query and  $d_1$ ,  $d_2$ , and  $d_3$  be documents of a collection. Thus,  $N = 3$ .

Q: gold silver truck  
 $d_1$ : Shipment of gold damaged in a fire.  
 $d_2$ : Delivery of silver arrived in a silver truck.  
 $d_3$ : Shipment of gold arrived in a truck.

Assuming term independence, rank documents in decreasing order of weights using F0 through F4 with (a)  $k = 0$  and (b)  $k = 0.5$ . Compare weighting function results for each  $k$  and between  $k$  values.

2. Show that RSJ weights scored with function F4 can be expressed as follows

$$w(t_i) = \log \left[ \left( \frac{p_i}{(1 - p_i)} \right) \left( \frac{1 - q_i}{q_i} \right) \right] \quad ; \text{ where}$$

$p_i = P(\text{document contains } t_i | \text{document is relevant})$

$q_i = P(\text{document contains } t_i | \text{document is not relevant})$

That is,  $p_i$  is the probability that the document contains  $t_i$  provided that it is relevant and  $q_i$  is the probability that the document contains  $t_i$  provided that it is not relevant.

3. Figure 1 suggests the dampening power of  $k$  is almost insignificant at lower values of  $r/n$ . Why?
4. Why the slopes of some of the curves shown in Figure 2 are either positive or negative? You need to evaluate the weighting function derivatives respect to  $k$  (i.e.,  $dF/dk$ ) using the data given in Table 9.

## References

Cox, D. R.; *Analysis of Binary Data*. Methuen, 1970, London.

Grossman, D. A. & Frieder, O. *Information Retrieval: Algorithms and Heuristics*. Springer, 2004, Netherlands.

Robertson, S. E., & Sparck Jones, K.; *Relevance weighting of search terms*, Journal of the American Society for Information Science, Volume 27, 1976 pp. 129–146.

Robertson, S. E. (2004). *Understanding Inverse Document Frequency: On theoretical arguments for IDF*. Journal of Documentation, 60, 5, 503-520.

Sparck Jones, K. (1972). *A statistical interpretation of term specificity and its application in retrieval*. Journal of Documentation, 28, 1, 11-21.