# Lecture 8 (Part 1) — September 20, 2007

*Prof. Lillian Lee* Scribe: *Cristian Danescu Niculescu-Mizil*

## 1 Review

Recall that by following the analysis of [1] and combining the mixture of 2-Poisson model with the Robertson-Spärck Jones scoring function [2] we obtained the following scoring formula for a document $d$ with the term frequency vector $\vec{d}$:

$$score_q(d) = \prod_{\substack{j:\, q[j]>0 \\ d[j]>0}} \frac{tr_j + (1-tr_j)\left(\frac{\mu_j}{\tau_j}\right)^{d[j]} e^{\tau_j-\mu_j}}{tg_j + (1-tg_j)\left(\frac{\mu_j}{\tau_j}\right)^{d[j]} e^{\tau_j-\mu_j}} \times \frac{tg_j\ e^{\mu_j-\tau_j} + (1-tg_j)}{tr_j\ e^{\mu_j-\tau_j} + (1-tr_j)}$$

$$\stackrel{rank}{=} \sum_{\substack{j:\, q[j]>0 \\ d[j]>0}} \log\left[\frac{tr_j + (1-tr_j)\left(\frac{\mu_j}{\tau_j}\right)^{d[j]} e^{\tau_j-\mu_j}}{tg_j + (1-tg_j)\left(\frac{\mu_j}{\tau_j}\right)^{d[j]} e^{\tau_j-\mu_j}} \times \frac{tg_j\ e^{\mu_j-\tau_j} + (1-tg_j)}{tr_j\ e^{\mu_j-\tau_j} + (1-tr_j)}\right] \tag{1}$$

where:

- $\tau_j$ and $\mu_j$ are the means of the Poisson distributions for the term $v_j$ in the on-topic and off-topic case, respectively:

$$P(A_j = d[j]|T_j = y) = Poisson(\tau_j) = \frac{\tau_j^{d[j]}}{d[j]!}e^{\tau_j} \tag{2}$$

$$P(A_j = d[j]|T_j = n) = Poisson(\mu_j) = \frac{\mu_j^{d[j]}}{d[j]!}e^{\mu_j} \tag{3}$$

- $tr_j$ is the probability of being on the topic of the term $v_j$, given relevance:

$$tr_j = P(T_j = y|R_q = y) \tag{4}$$

- $tg_j$ is the probability of being on the topic of the term $v_j$ in general:

$$tg_j = P(T_j = y) \tag{5}$$

- we maintain the assumption that all documents have equal length.

For reference, we also repeat here the RSJ scoring function for the binary attributes case discussed in the previous lectures:

$$
\begin{aligned}
RSJ_q(d) &= \prod_{\substack{j:\, q[j]>0 \\ d[j]>0}} \frac{P(A_j = 1|R_q = y)}{P(A_j = 1)} \times \frac{1 - P(A_j = 1)}{1 - P(A_j = 1|R_q = y)} \\
&\stackrel{rank}{=} \sum_{\substack{j:\, q[j]>0 \\ d[j]>0}} \log\left[\frac{P(A_j = 1|R_q = y)}{P(A_j = 1)} \times \frac{1 - P(A_j = 1)}{1 - P(A_j = 1|R_q = y)}\right] \qquad (6)
\end{aligned}
$$

Note that — as an advantage over the RSJ scoring function for the binary attributes case — (1) is complex enough to account for the non-binary attributes case. The tradeoff is the presence of four unknown parameters for each term ($\mu_j$, $\tau_j$, $tr_j$ and $tg_j$) which we can not directly estimate. This makes this scoring function difficult to estimate and to use in practice.

## 2 Analysis of the scoring function

Acknowledging this problem we will now try to find a more tractable scoring function that has approximately the same behavior as (1). For this purpose in the following we analyze the behavior of the terms of (1)[1] , seen as a functions of $d[j]$:

$$
f(d[j]) = \log\left[\frac{tr_j + (1 - tr_j)\left(\frac{\mu_j}{\tau_j}\right)^{d[j]} e^{\tau_j - \mu_j}}{tg_j + (1 - tg_j)\left(\frac{\mu_j}{\tau_j}\right)^{d[j]} e^{\tau_j - \mu_j}} \times \frac{tg_j\ e^{\mu_j - \tau_j} + (1 - tg_j)}{tr_j\ e^{\mu_j - \tau_j} + (1 - tr_j)}\right] \qquad (7)
$$

(a) For $d[j] = 0$ we have:

$$
\begin{aligned}
f(0) &= \log\left[\frac{tr_j + (1 - tr_j)e^{\tau_j - \mu_j}}{tg_j + (1 - tg_j)e^{\tau_j - \mu_j}} \times \frac{tg_j\ e^{\mu_j - \tau_j} + (1 - tg_j)}{tr_j\ e^{\mu_j - \tau_j} + (1 - tr_j)}\right] \\
&= \log\left[\frac{tr_j + (1 - tr_j)e^{\tau_j - \mu_j}}{tg_j + (1 - tg_j)e^{\tau_j - \mu_j}} \times \frac{tg_j\ e^{\mu_j - \tau_j} + (1 - tg_j)}{tr_j\ e^{\mu_j - \tau_j} + (1 - tr_j)} \times \frac{e^{\tau_j - \mu_j}}{e^{\tau_j - \mu_j}}\right] \\
&= \log\left[\frac{tr_j + (1 - tr_j)e^{\tau_j - \mu_j}}{tg_j + (1 - tg_j)e^{\tau_j - \mu_j}} \times \frac{tg_j + (1 - tg_j)e^{\tau_j - \mu_j}}{tr_j + (1 - tr_j)e^{\tau_j - \mu_j}}\right] \\
&= \log(1) = 0 \qquad (8)
\end{aligned}
$$

[1]Note that in these lecture notes we fix a problem with the presentation given in class: we analyze here the terms of the *log* version of the scoring function (1) and, by doing so, we provide a better justification for the proposed approximation.

(b) For $d[j] \to \infty$ the behavior of (7) is determined by the asymptotic value of $\left(\frac{\mu_j}{\tau_j}\right)^{d[j]}$. Note that it is natural to assume that $\mu_j < \tau_j$: we expect to encounter more query terms in on-topic documents than in off-topic documents. Therefore, as $d[j] \to \infty$ we have $\left(\frac{\mu_j}{\tau_j}\right)^{d[j]} \to 0$ and:

$$f(d[j]) \to \log \left[ \frac{tr}{tg} \times \frac{tg_j \; e^{\mu_j - \tau_j} + (1 - tg_j)}{tr_j \; e^{\mu_j - \tau_j} + (1 - tr_j)} \right]. \tag{9}$$

Making the additional assumption that $\mu_j - \tau_j << 0$, and thus $e^{\mu_j - \tau_j} \simeq 0$, we have:

$$f(d[j]) \to \log \left[ \frac{tr}{tg} \times \frac{tg_j \; e^{\mu_j - \tau_j} + (1 - tg_j)}{tr_j \; e^{\mu_j - \tau_j} + (1 - tr_j)} \right] \simeq \log \left[ \frac{tr}{tg} \times \frac{(1 - tg_j)}{(1 - tr_j)} \right]. \tag{10}$$

Now, if we are also willing to accept that $tr_j = P(T_j = y|R_q = y) \approx P(A_j = 1|R_q = y)$ and that $tg_j = P(T_j = y) \approx P(A_j = 1)$ (i.e. that the probability of being on topic is approximated by the probability of containing the term indexing that topic) then (10) tells us that, for $d[j] \to \infty$, $f(d[j])$ approximates the terms of the RSJ weight (6). Knowing from our previous analysis of the RSJ terms that they can be interpreted (under certain assumptions) as inverse document frequency[2] we can conclude that in this limit case $f(d[j]) \to IDF_j$.

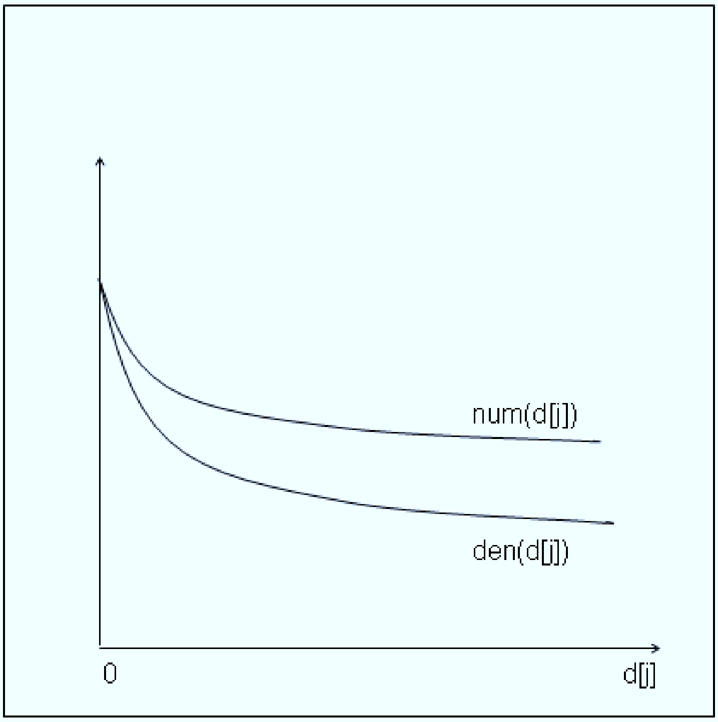


Figure 1: The monotonicity of the numerator $num(d[j])$ and of the denominator $den(d[j])$ of (7).

[2]We refer here to the logarithmic version of the inverse document frequency: $IDF_j = log(N/n_j)$, where $N$ is the number of documents in the corpus and $n_j$ is the number of documents in witch the term $v_j$ occur.

(c) For $0 < d[j] < \infty$ we will analyze the numerator $num(d[j])$ and the denominator $den(d[j])$ of the argument of the log in (7) separately. Both $num(d[j])$ and $den(d[j])$ are exponentially decreasing because $\mu_j < \tau_j$ (as we discussed in (b)). Also, from (8) we know that at $d[j] = 0$ the numerator equals the denominator. By employing the same reasoning as in (b) for $num(d[j])$ and $den(d[j])$ separately, we obtain the following linear horizontal asymptotes:

$$\lim_{d[j]\to\infty} num(d[j]) = tr_j \times (1 - tg_j) \tag{11}$$

$$\lim_{d[j]\to\infty} den(d[j]) = tg_j \times (1 - tr_j). \tag{12}$$

Assuming that for terms $v_j$ in the query the probability of being on the topic of $v_j$ is greater for relevant documents than the probability of being on the topic of $v_j$ in general documents (i.e. $tr_j > tg_j$), the asymptotic value of the numerator is greater than that of the denominator, and thus $num(d[j])$ decreases slower than $den(d[j])$ as illustrated in Fig. 1. Therefore, $num(d[j])/den(d[j]))$ is monotonically increasing for $0 < d[j] < \infty$ and, given the monotonicity of the log function, $f(d[j])$ is monotonically increasing for $0 < d[j] < \infty$.

Summing up our analysis, we conclude that the terms $f(d[j])$ of (1) are monotonically increasing from 0 to a value identifiable with $IDF_j$.

## Exercise

a) Suppose we are interested in finding out about the breeding habits of a certain species of chipmunks, namely the alpine chipmunks. We construct the query "alpine chipmunks breeding" and submit it to Google$^{TM}$. Out of the obtained ranking we extract the following results:

| **Name** | **Rank** | **Address** |
|---|---|---|
| $\text{Doc}_1$ | 1 | www.nps.gov/history/history/online_books/grinnell/mammals63.htm |
| $\text{Doc}_2$ | 2 | animaldiversity.ummz.umich.edu/site/accounts/information/Tamias_alpinus.html |
| $\text{Doc}_3$ | 8 | sfgate.com/cgi-bin/article.cgi?f=/c/a/2005/11/27/ING66FMV901.DTL |
| $\text{Doc}_4$ | 21 | ilmbwww.gov.bc.ca/risc/pubs/tebiodiv/pisc/piscml20-06.htm |

where the "Rank" column refers to the ranking given by the search engine.

Take a quick look at these web-pages and judge their relative relevance to the query yourself. Then rank them according to the following approximation of the 2-Poisson model scoring function (first proposed in [1]):

$$score_q(d) = \sum_{\substack{j:\, q[j]>0 \\ d[j]>0}} \frac{d[j]}{k + d[j]} \times idf_j \tag{13}$$

and compare your results with the Google$^{TM}$ ranks and with your own expectations. Set $k = 1.5$ and make an informed choice of the inverse document frequency $idf_j$. Note that for simplicity we are employing the version of the scoring function which assumes equal document length — the documents above were selected to have roughly the same size.

b) Now let's look more in detail at the term frequency related part of (13):

$$tfpart_j^k(d) = \frac{d[j]}{k + d[j]} \tag{14}$$

In [1] Robertson and Walker motivated the choice for this expression by the fact that this leads to a scoring function that has approximately the same behavior as the 2-Poisson model score function. We claim that there is another aspect that makes this $tfpart$ preferable over other alternatives. Find and discuss this advantage and analyze the effects of modifying $k$, going beyond the most obvious answer. Relate this discussion to our example.

c) In the lecture notes, in our analysis of the behavior of the factors of the 2-Poisson model scoring function we assumed that $\mu_j - \tau_j << 0$ and therefore $e^{\mu_j - \tau_j} \simeq 0$. Discuss a case when this assumption does not hold and use our setting to exemplify this case.

**Solutions:**

a) The Google$^{TM}$ ranking matches our intuition, except for the relatively high ranking of Doc$_3$, which only mentions alpine chipmunks as an example, and contains nothing related to their breeding habits. We consider Doc$_4$ to be more relevant than Doc$_3$, given that it talks about breeding habits of chipmunks (even though not about alpine chipmunks).

Given the indexing of the sum in (13), we only need to calculate $d[j]$ and $idf_j$ for the terms that appear both in the query and documents: "chipmunk" (we do not distinguish between the singular and plural form),"alpine" and "breeding". We calculate $idf_j$ using the formula:

$$idf_j = ln \frac{|C|}{\# \; docs \; in \; C \; containing \; v_j} \tag{15}$$

where $C$ is the corpus from which the documents was retrieved: the set of English language web-pages indexed by Google$^{TM}$ (approximate size: $4,320,000,000$ documents). We get the denominators by searching for the individual terms and reading the approximate number of indexed documents containing those words. The inverse document frequencies obtained this way and the term frequencies are:

| | **chipmunk(s)** | **alpine** | **breeding** |
|---|---|---|---|
| idf | 7.10 | 4.50 | 4.62 |
| Doc$_1$ | 38 | 19 | 2 |
| Doc$_2$ | 15 | 12 | 3 |
| Doc$_3$ | 3 | 5 | 3 |
| Doc$_4$ | 76 | 4 | 3 |

The ranking we obtain using the scoring function (13) is [Doc$_1$,Doc$_2$,Doc$_4$,Doc$_3$] which matches our intuition:

| | **Doc$_1$** | **Doc$_2$** | **Doc$_3$** | **Doc$_4$** |
|---|---|---|---|---|
| Score | 13.65 | 13.54 | 11.28 | 13.32 |

b) First we notice that $k$ allows us to gauge the importance that (13) gives to term frequencies (for values of k that are not overly large). To realize this we consider two documents $d$ and $f$ in which a query term $j$ has different frequencies: $d[j] > f[j]$. To see how $tfpart$ contributes to distinguishing these documents we look at:

$$tfpart_j^k(d) - tfpart_j^k(f) = \frac{d[j]}{k + d[j]} - \frac{f[j]}{k + f[j]} \tag{16}$$

as a function of $k$. As can seen in Fig. 2, for $k$ smaller than a certain value, $tfpart_j^k(d) - tfpart_j^k(f)$ is monotonically increasing: the larger the value of $k$, the more the gap between the frequencies matters.

We can explain this behavior analytically by calculating the derivative of (16) with respect to $k$:

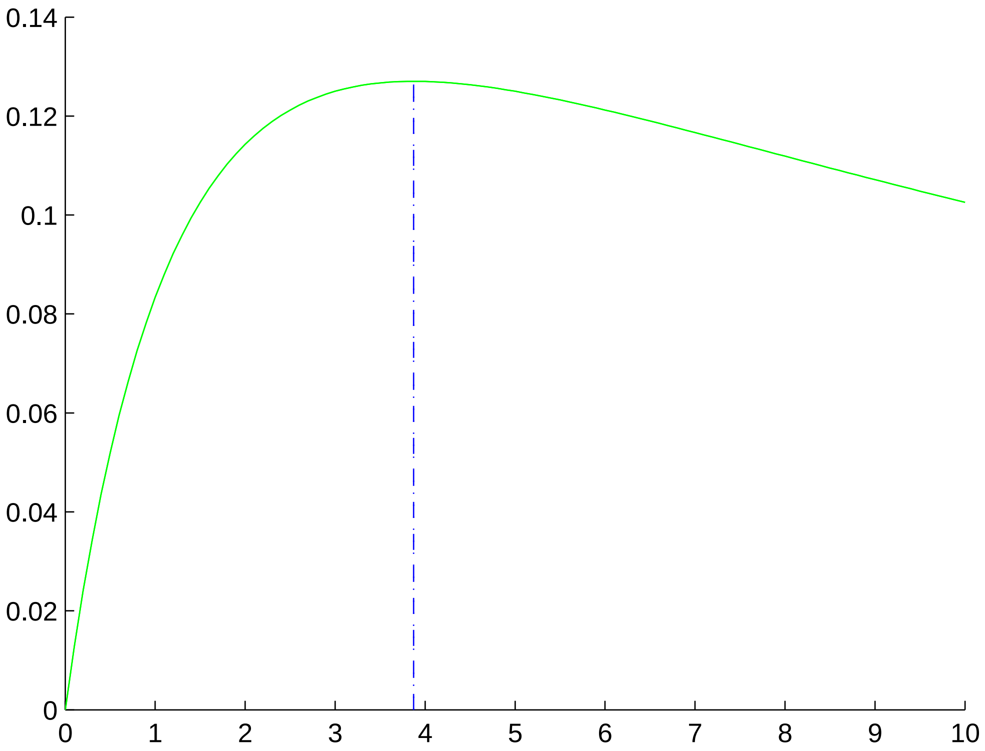


Figure 2: $tfpart_j^k(d) - tfpart_j^k(f)$ as a function of k; the dashed line represents $k^* = \sqrt{d[j]f[j]}$, the point where the function changes it monotonicity.

$$\begin{aligned} tfpart_j^k(d) - tfpart_j^k(f) &= \frac{d[j]}{k + d[j]} - \frac{f[j]}{k + f[j]} \\ &= \frac{k(d[j] - f[j])}{(k + d[j])(k + f[j])} \\ &= \frac{d[j] - f[j]}{k + (d[j] + f[j]) + d[j]f[j]/k} \end{aligned} \tag{17}$$

$$\begin{aligned} (tfpart_j^k(d) - tfpart_j^k(f))' &= \left( \frac{d[j] - f[j]}{k + (d[j] + f[j]) + \frac{d[j]f[j]}{k}} \right)' \\ &= -(d[j] - f[j]) \frac{1 + \left(\frac{d[j]f[j]}{k}\right)'}{\left(k + (d[j] + f[j]) + \frac{d[j]f[j]}{k}\right)^2} \\ &= -(d[j] - f[j]) \frac{1 - \frac{d[j]f[j]}{k^2}}{\left(k + (d[j] + f[j]) + \frac{d[j]f[j]}{k}\right)^2} \end{aligned} \tag{18}$$

Therefore, the derivative equals 0 only for $k = \sqrt{d[j]f[j]}$, is positive for $0 < k < \sqrt{d[j]f[j]}$ and is negative for $k > \sqrt{d[j]f[j]}$ and thus (16) is increasing for $0 < k < \sqrt{d[j]f[j]}$ and decreasing for $0 > k > \sqrt{d[j]f[j]}$.

However, there is a more interesting aspect of $tfpart$ that is related to the order of magnitude of the term frequencies. This can be understood by comparing $tfpart_j^k(d) - tfpart_j^k(f)$ and $tfpart_i^k(d) - tfpart_i^k(f)$ for two query terms $i$ and $j$ such that $d[i] >> d[j]$ and $f[i] >> f[j]$. In Fig. 3 we plot these as functions of $k$ for $d[i] = 50$, $f[i] = 10$, $d[j] = 5$ $f[j] = 3$ and we note that there is an interval of values of $k$ for which the small difference between small magnitude frequencies $d[j]$ and $f[j]$ matters more to the scoring function than the relatively big difference between the high magnitude frequencies $d[i]$ and $f[i]$ (given equal inverse document frequency).

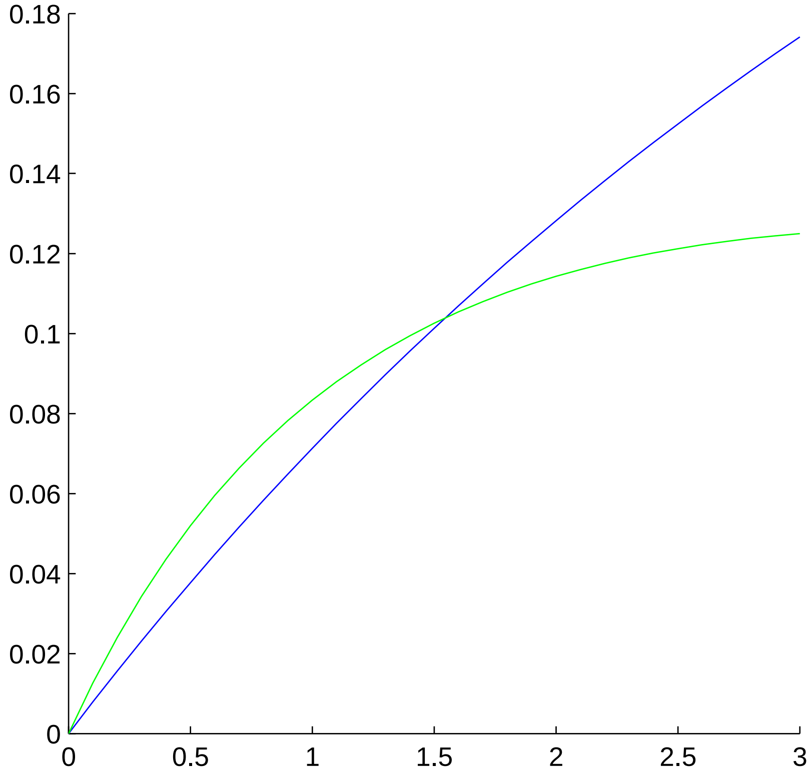


Figure 3: $tfpart_i^k(d) - tfpart_i^k(f)$ (in blue) and $tfpart_j^k(d) - tfpart_j^k(f)$ (in green) as a functions of $k$; $d[i] = 50$, $f[i] = 10$, $d[j] = 5$, $f[j] = 3$

We can briefly explain this behavior analytically by observing in (17) that the term $d[j]f[j]$ — corresponding to the magnitude of the respective frequencies — appears in the denominator. Comparing expression (17) for two query terms $i$ and $j$ such that $d[i] >> d[j]$ and $f[i] >> f[j]$ and $d[i] - f[i] \geq d[j] - f[j]$:

$$tfpart_j^k(d) - tfpart_j^k(f) = \frac{d[j] - f[j]}{k + (d[j] + f[j]) + d[j]f[j]/k} \quad (19)$$

$$tfpart_i^k(d) - tfpart_i^k(f) = \frac{d[i] - f[i]}{k + (d[i] + f[i]) + d[i]f[i]/k} \quad (20)$$

we observe that for fixed small values of $k$ the fact that $d[i]f[i]/k >> d[j]f[j]/k$ (in the denominator) undermines the effect of $d[i]-f[i] \geq d[j]-f[j]$ (in the numerator) and determines $tfpart_i^k(d) - tfpart_i^k(f)$ to be smaller than $tfpart_j^k(d) - tfpart_j^k(f)$.

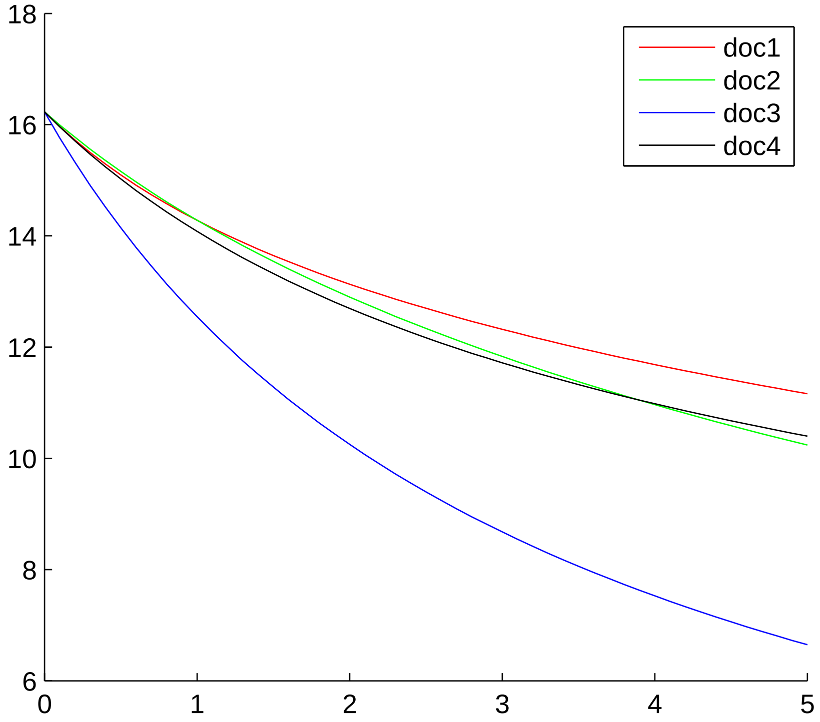


Figure 4: Relative behavior of the scoring function with respect to $k$.

Using our example, we explain why such behavior might be considered intuitive and desired: we know that $Doc_1$ and $Doc_3$ contain relatively many "chipmunk" terms, so we know that they are on the topic of "chipmunks" and we do not care so much which one contains more "chipmunk" terms; however, at this point we would like to know which of the documents talks about "alpine chipmunks", and therefore we put more emphasis on the small difference in the frequency of "alpine". And indeed, as seen in a), (13) ranks $Doc_1$ higher than $Doc_4$ even though $Doc_4$ contains double the number of "chipmunk" terms and 24 more query terms than $Doc_1$: the $tfpart$ behaves such that the relatively small magnitude difference between the count of "alpine" terms matters more. A simple analysis of our inverse document frequencies shows that this is not the effect of the $idf_j$ part of the scoring function. Also, if instead of $tfpart$ we use the simple term frequency count $tf_j(d) = d[j]$, $\text{Doc}_4$ ranks above $\text{Doc}_1$.

In Fig. 4 the behavior of the complete scoring functions for different values of $k$ is illustrated. Confirming our first observation about the role of $k$, the difference between the score of $\text{Doc}_3$ and the scores of all the other documents increases with $k$ — the difference in term frequency is taken more into account. Also, as a consequence of the impact that $k$ has in the importance that the order of magnitude of the term frequencies has, we note that for $k$ greater than a certain value $\text{Doc}_4$ ranks above $\text{Doc}_2$ and that for small $k$ $\text{Doc}_2$ ranks higher than $\text{Doc}_1$ — for those values of $k$ the fact that $\text{Doc}_2$ has an extra "breeding" term is considered more important than the 23 "chipmunk" and 7 "alpine" terms that $\text{Doc}_1$ has in excess of $\text{Doc}_3$.

c) As Robertson and Walker point out in [1], the assumption that $e^{\mu_j - \tau_j} \simeq 0$ does not hold for infrequent terms which we do not expect to have a high frequency in the results of our query. In our case "breeding" is such a term: relevant documents contain just $2 - 3$ occurrences of this term, so even if the expected rate of terms "breeding" is almost zero in other documents, the difference between $\mu_j$ and $\tau_j$ is not big enough to justify the assumption that $e^{\mu_j - \tau_j} \simeq 0$.

## References


[1] S. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. SIGIR, pp. 232-241 (1994).

[2] S. Robertson and K. Spärck Jones. Relevance weighting of search terms. Journal of the American Society for Information Science 27, 129-46 (1976).

[3] S. Robertson, S. Walker, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In Proceedings of the Third Text REtrieval Conference (TREC-3), NIST Special Publication 500-225 (1995).