Inverse Document Frequency (IDF): A Measure of Deviations from Poisson

Kenneth W. Church
William A. Gale

AT&T Bell Laboratories
Murray Hill, NJ, USA 07974
kwc@research.att.com

Abstract

Low frequency words tend to be rich in content, and vice versa. But not all equally frequent words are equally meaningful. We will use *inverse document frequency* (IDF), a quantity borrowed from Information Retrieval, to distinguish words like *somewhat* and *boycott*. Both *somewhat* and *boycott* appeared approximately 1000 times in a corpus of 1989 Associated Press articles, but *boycott* is a better keyword because its IDF is farther from what would be expected by chance (Poisson).

1. Document frequency is similar to word frequency, but different

Word frequency is commonly used in all sorts of natural language applications. The practice implicitly assumes that words (and ngrams) are distributed by a single parameter distribution such as a Poisson or a Binomial. But we find that these distributions do not fit the data very well. Both the Poisson and Binomial assume that the variance over documents is no larger than the mean, and yet, we find that it can be quite a bit larger, especially for interesting words such as *boycott* where there are hidden variables such as topic that conspire to undermine the independence assumption behind the Poisson and the Binomial. Much better fits are obtained by introducing a second parameter such as *inverse document frequency* (IDF).

Inverse document frequency (IDF) is commonly used in Information Retrieval (Sparck Jones, 1972). IDF is defined as $-log_2 df_w / D$, where D is the number of documents in the collection and df_w is the document frequency, the number of documents that contain w. Obviously, there is a strong relationship between document frequency, df_w, and word frequency, f_w. The relationship is shown in Figure 1, a plot of $\log_{10} f_w$ and IDF for 193 words selected from a 50 million word corpus of 1989 Associated Press (AP) Newswire stories ($D = 85,432$ stories).

Although $\log_{10} f_w$ is highly correlated with IDF ($\rho = -0.994$), it would be a mistake to assume that the two variables are completely predictable from one another. Indeed, the experience of the Information Retrieval community has indicated that IDF is a very useful quantity. Attempts to replace IDF with f_w (or some simple transform of f_w) have not been very successful.

Figure 2 shows one such attempt. It compares the observed IDF with $\hat{IDF}$, an estimate based on f. Assume that a document is merely a ‘‘bag of words’’ with no interesting structure (content). Words are randomly generated by a Poisson process, π. The probability of k instances of a word w is $\pi(\theta, k)$ where $\theta \approx \frac{f_w}{D}$:

$$\pi(\theta,k) = \frac{e^{-\theta}\ \theta^k}{k!} \quad \text{for } k = 0,1,\cdots \qquad \text{Poisson}$$

In particular, the probability that w will not be found in a document is $\pi(\theta,0)$. Conversely, the probability of at least one w is $1-\pi(\theta,0)$. And therefore, *IDF* ought to be:

$$\hat{IDF} = -log_2(1-\pi(\theta,0)) = -\log_2(1-e^{-\theta}) \qquad \text{Predicted IDF}$$

Figure 2 compares *IDF* with $\hat{IDF}$. Note that $\hat{IDF}$ is systematically too low, indicating that the predictions are missing crucial generalizations. Documents are more than just a bag of words.

The prediction errors are shown in more detail in Figure 3, which plots the residual IDF (difference between predicted and observed) as a function of $\log_{10} f_w$ for the same 193 words shown in Figure 2. The prediction errors are relatively large in the middle of the frequency range, and smaller at both ends. Unfortunately, we believe the words in the middle are often the most important words for Information Retrieval purposes.

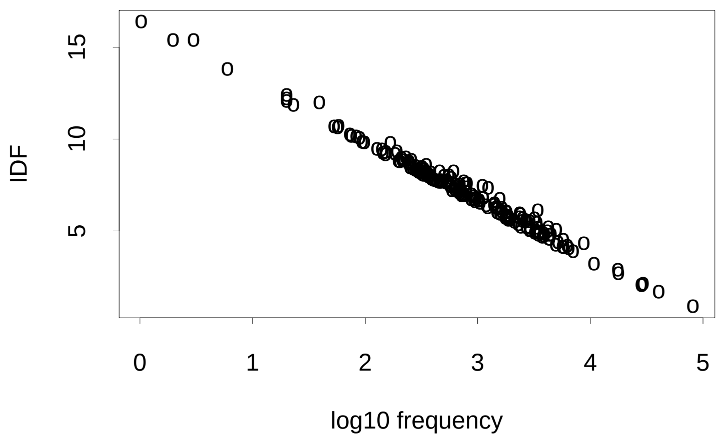

Figure 1: IDF is highly correlated with log frequency ($\rho = -0.994$). The circles show $\log_{10} f$ and IDF for 193 words selected from a corpus of 1989 Associated Press Newswire stories ($D = 85,432$).

2. A Good Keyword is far from Poisson

To get a better look at the crucial differences between IDF and f in the middle frequency range ($f \approx 10^3$), we selected a set of 53 words for further investigation with $1000 < f < 1020$ in the 1989 AP corpus. The 53 words are shown in Table 1, sorted by *df*. Note that the words near the top of the list tend to be more appropriate for use in an information retrieval system than the words toward the bottom of the list. Stories that mention the word *boycott*, for example, are likely to be about boycotts. In contrast, stories that mention the word *somewhat* could be about practically anything.[1]

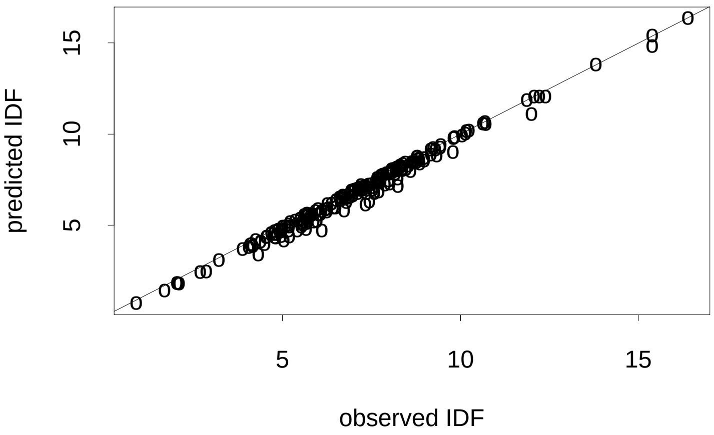

Figure 2: The observed IDF is systematically lower than what would be expected under a Poisson, $-log_2(1 - e^{-f/D})$. All but 6 of the circles fall below the $x = y$ line. The data are the same as in Figure 1.

Why is IDF such a useful quantity? One might try to answer the question in terms of information theory (Shannon, 1948). IDF can be thought of as the usefulness in bits of a keyword to a keyword retrieval system. If we tell you that the document that we are looking for has the keyword *boycott*, then we have narrowed the search space down to just 676/D documents.

But, this answer doesn't explain the fundamental difference between *boycott* and *somewhat*. *boycott* has an IDF of $-log_2 676/D = 7.0$ bits, only a little more than *somewhat*, which has an IDF of $-log_2 979/D = 6.4$. And yet, *boycott* is a reasonable keyword and *somewhat* is not.

A good keyword, like *boycott*, picks out a very specific set of documents. The problem with *somewhat* is that it behaves almost like chance (Poisson). Under a Poisson, the 1013 instances of *somewhat* should be found in approximately $D(1 - \pi(\theta, 0)) \approx D(1 - \pi(1013/85432, 0)) \approx 1007$ documents. In fact, *somewhat* was found in 979 documents, only a little less than what would have been expected by chance. Good keywords tend to bunch up into many fewer documents. *boycott*, for example, bunches up into only 676 documents, much less than chance ($D(1 - \pi(1009/85432, 0)) \approx 1003$). Almost all words are more ''interesting'' in this sense than Poisson, but good keywords like *boycott* are a lot more interesting than Poisson, and crummy ones like *somewhat* are only a little more interesting than Poisson.

---

1. There is a weak tendency for nouns to appear higher on the list than non-nouns, though tendency is too weak to explain the pattern of the systematic deviations from Poisson. In addition, there are plenty of exceptions in both directions: *rape, pool, grants, code* and *premier* are not necessarily nouns, and *sweeping, leads, bound* and *worry* are not necessarily non-nouns.

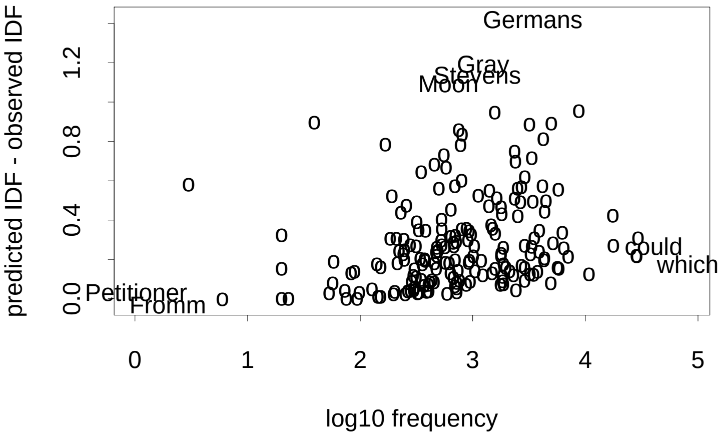

Figure 3: The Prediction errors are systematically positive. The errors tend to be larger in the middle of the frequency range (*Germans*), and smaller at both ends (*Fromm*, *which*). The data are the same as in Figures 1-2.

On this account, a good keyword is one that behaves very differently from the null hypothesis (Poisson). We conjecture that the best keywords tend to be found toward the middle of the frequency range, where there are relatively large deviations from Poisson, as illustrated in Figure 3. This hypothesis runs counter to the standard practice in Information Retrieval of weighting words by IDF, favoring extremely rare words, no matter how they are distributed.

Of course, IDF is but one of many ways to show deviations from chance. Figure 4 shows the distributions for *boycott* and *somewhat*. Note that *somewhat* is much ‘‘closer’’ to Poisson in almost any sense of closeness that one might consider. Three measures of ‘‘closeness’’ are presented in Table 2: IDF, variance (σ^2), and entropy (H). Table 2 compares the top 10 words in Table 1 (labeled ‘‘better keywords’’) with the bottom 10 words in Table 1 (labeled ‘‘worse keywords’’). The better keywords have more IDF, more variance and less entropy than what would be expected under a Poisson with $\theta \approx f / D \approx 1000/85{,}432 \approx 0.012$.

3. How robust are these deviations from chance?

We were concerned that the crucial deviations from Poisson behavior might not hold up if we looked at another corpus of similar material. Figure 5 shows the word *boycott* in five different years of the AP news. The ‘‘fat tails’’ show up in each of the five years. Clearly, the non-Poisson phenomenon is robust.

Figures 6 and 7 compare IDF and $\log_{10} \sigma^2$ for the 53 words in Table 1, and find that IDF and $\log_{10} \sigma^2$ are reasonably stable across years. The correlations of IDF and $\log_{10} \sigma^2$ across years are presented in Tables 3-4. All of the correlations are quite large. The correlations for IDF are perhaps somewhat larger than those for $\log_{10} \sigma^2$, suggesting that IDF may be somewhat more robust, which is not

Table 1: More IDF (less df) → More Content

df	w	df	w	df	w	df	w
435	governors	724	pool	827	unity	937	worry
506	festival	740	restaurants	845	bed	940	containing
551	gang	745	grants	847	coastal	946	explained
553	bullion	752	scheme	851	educational	951	bound
563	attendants	754	code	853	lying	953	leads
623	rape	761	premier	853	neighbor	955	happens
639	palace	775	wire	863	tragedy	960	improving
676	boycott	781	customer	867	acquire	960	welcomed
687	routes	783	rooms	874	restored	961	triggered
690	incentives	786	engineering	905	legitimate	966	sweeping
695	poverty	803	color	910	deliver	968	fairly
718	donations	811	possession	914	types	969	heading
722	lawsuits	815	projected	929	reject	979	somewhat
						986	noting

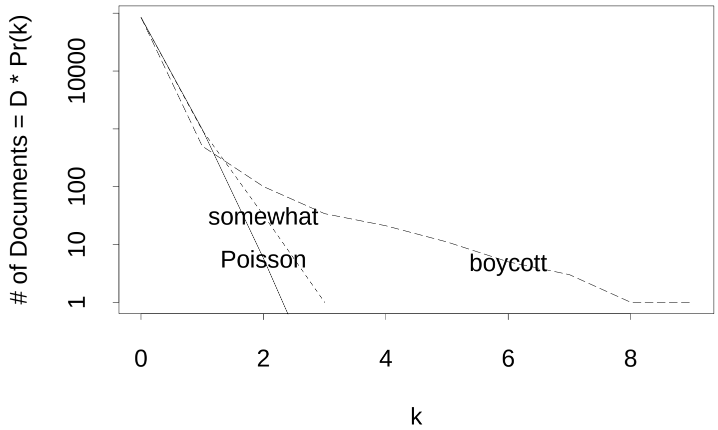

Figure 4: Most words have a fatter tail than Poisson (solid line). The deviations from Poisson are more salient for good keywords like *boycott*, than for crummy keywords like *somewhat*.

surprising given that empirical estimates of variance are notoriously subject to outliers. None of the correlations in Tables 3 and 4 can be attributed to word frequency effects since the 53 words were all chosen with almost the same 1989 frequency.

In general, the correlations in Tables 3-4 are larger near the diagonal, suggesting that estimates degrade over time. If you want to predict next year's IDF, it is better to use this year's estimate than a ten-year-old estimate.

Table 2: Good keywords have more IDF, more var and less entropy than Poisson

Better Keywords				Worse Keywords			
IDF	var	entropy		IDF	var	entropy	
7.6	0.060	0.057	governors	6.5	0.013	0.092	leads
7.4	0.044	0.064	festival	6.5	0.013	0.092	happens
7.3	0.043	0.067	gang	6.5	0.013	0.092	improving
7.3	0.028	0.068	bullion	6.5	0.013	0.092	welcomed
7.2	0.042	0.068	attendants	6.5	0.013	0.092	triggered
7.1	0.032	0.073	rape	6.5	0.013	0.093	sweeping
7.1	0.028	0.074	palace	6.5	0.013	0.093	fairly
7.0	0.027	0.077	boycott	6.5	0.013	0.093	heading
7.0	0.026	0.078	routes	6.4	0.013	0.093	somewhat
7.0	0.025	0.078	incentives	6.4	0.012	0.092	noting
6.4	0.012	0.092	*Poisson*	6.4	0.012	0.092	*Poisson*

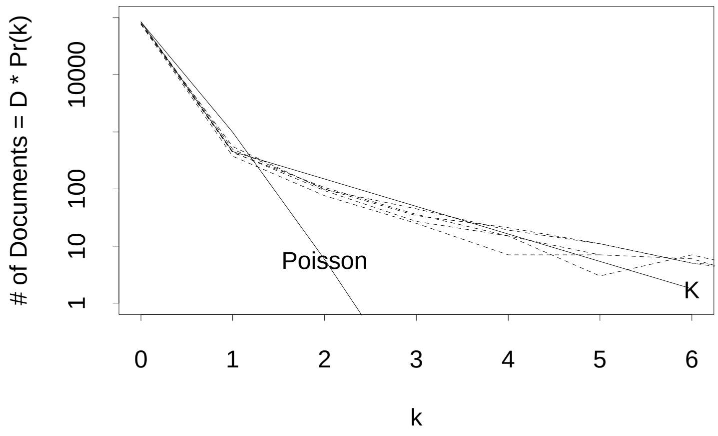

Figure 5: The strong deviations from Poisson for the word *boycott* show up very clearly in the AP in 1988, 1989, 1990, 1991 and 1992 (dotted lines). Katz' K-mixture (Katz, personal communication), the solid line labelled ''K,'' fits the data better than the Poisson.

Another way to confirm that our measurements of IDF, variance and H have consequences across years in the AP data, is to note that measurements of IDF, variance and H in 1989 can be used to predict word frequency in some other year. The correlations are shown in Table 5. They may not not be large, but they are too large to be due to chance and they all point in the same direction. The correlations cannot be attributed to variations in frequency in 1989, since all 53 words have almost the same 1989 frequency. Clearly, there are some interesting systematic relationships between IDF/variance/H and f that hold up to replication across multiple years in the AP, measurement errors, and other sources of noise.

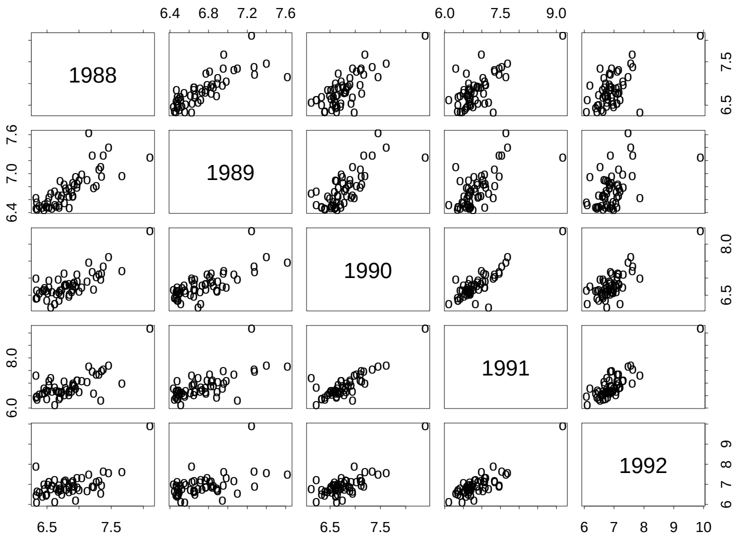

Figure 6: IDF in one year of the AP is very predictive of IDF in another (for the 53 words in Table 1). Each scatter plot compares IDF in one year with IDF in another. The fact that most of the points line up fairly well indicates that IDF values are strongly correlated across years. The correlations are shown in Table 3.

4. Katz' K-mixture

Clearly, the Poisson does not fit our data very well, especially for good keywords like *boycott*. This is, however, a negative result. Can we say something more constructive?

Katz (personal communication) proposed the following alternative to the Poisson. $Pr_K(k)$ is the probability of *k* instances of *w* in a document.

$$Pr_K(k) = (1-\alpha)\ \delta_{k,0} + \frac{\alpha}{\beta+1}\ \left(\frac{\beta}{\beta+1}\right)^k \qquad \text{K-mixture}$$

$\delta_{k,0}$ is 1 when $k=0$, and 0 otherwise. Katz' K-mixture distribution can be thought of as a mixture of Poissons. Suppose that, within documents, *boycott* is distributed by a Poisson process, but, across documents, the Poisson parameter θ is allowed to vary from one document to another depending on how much the document is about boycotts. In other words, $Pr_K(k)$ can be expressed as a convolution of Poissons with a density function ϕ:

$$Pr(k) = \int_0^\infty \phi(\theta)\ \pi(\theta,k)\, d\theta \quad \text{for } k = 0,1,\cdots \qquad \text{Poisson Mixture}$$

In this way, the θs can depend on an infinite number of unknowable hidden variables, e.g., what the documents are about, who wrote them, when they were written, what was going on in the world when they were written, etc., but we don't need to know these dependencies for any particular document. All we need to know is ϕ, the density of θs, aggregated over all possible combinations of hidden variables.

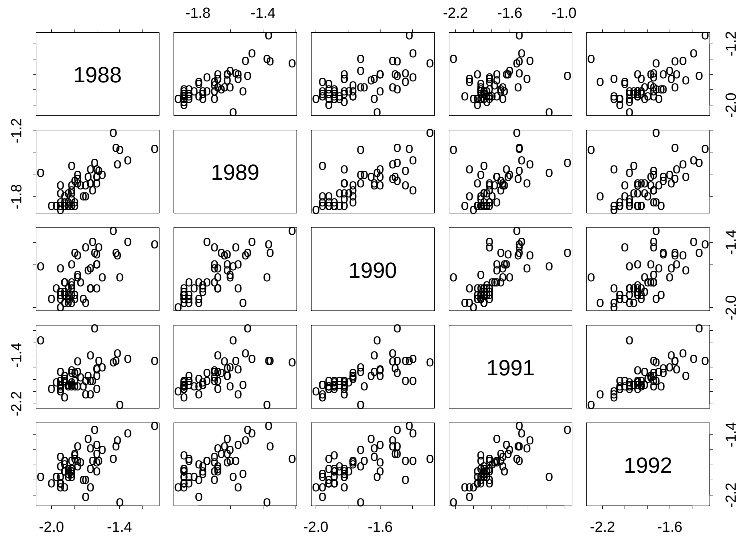

Figure 7: $\log_{10}\sigma^2$ is also predictable from one year to the next, though maybe not as predictable as IDF (for the 53 words in Table 1). The correlations are shown in Table 4.

Table 3: Correlations of IDF across years

	1988	1989	1990	1991	1992
1988		0.80	0.76	0.68	0.60
1989	0.80		0.75	0.67	0.48
1990	0.76	0.75		0.85	0.76
1991	0.68	0.67	0.85		0.84
1992	0.60	0.48	0.76	0.84	

Table 4: Correlations of log var across years

	1988	1989	1990	1991	1992
1988		0.74	0.61	0.25	0.67
1989	0.74		0.73	0.42	0.51
1990	0.61	0.73		0.50	0.61
1991	0.25	0.42	0.50		0.62
1992	0.67	0.51	0.61	0.62	

Table 5: Correlations of IDF, log var and H in 1989 with log f in other years

	1988 log f	1990 log f	1991 log f	1992 log f
1989 IDF	–0.18	–0.14	–0.20	–0.17
1989 log var	–0.13	–0.11	–0.14	–0.12
1989 H	0.17	0.15	0.20	0.16

In the case of Katz' K-mixture, $\phi(\theta)$ is assumed to be $(1-\alpha)\;\delta(\theta)+\frac{\alpha}{\beta}\;e^{-\frac{\theta}{\beta}}$. $\delta(k)$ is Dirac's delta function, ∞ when $k=0$, and otherwise, 0.

Katz' K-mixture has two parameters, α and β. The α parameter determines the fraction of relevant and irrelevant documents. $1-\alpha$ of the documents have no chance of mentioning *boycott* $(\theta = 0)$ because they are totally irrelevant to boycotts. The β parameter determines the average θ among the relevant documents.

The two parameters, α and β, can be fit from almost any pair of variables considered thus far, e.g., *f*, IDF, σ^2, H. We have found that *f* and IDF are particularly easy to work with, and are more robust than some others such as σ^2.

$$\beta \approx \frac{f}{D} 2^{IDF} - 1$$

$$\alpha \approx \frac{f}{D} \frac{1}{\beta}$$

It has been our experience that Katz' K-mixture fits the data much better than the Poisson, as can be seen in Figure 5. Unlike the Poisson, the K-mixture has two parameters, α and β, and can therefore account for the fact that IDF and *f* are not completely predictable from one another.

In related work (Church and Gale, submitted), we looked at a number of different Poisson mixtures, and found that our data can also be fit by a negative binomial, which can be viewed as a Poisson mixture where $\phi_{NB}(\theta)$ is a Gamma distribution (Johnson and Kotz, 1969). See Mosteller and Wallace (1964) for an example of how to use the negative binomial in a Bayesian discrimination task. It is straightforward to generalize the Mosteller and Wallace approach to use Katz' K-mixture or any other mixture of Poissons.

5. Conclusions

Documents are much more than just a bag of words. The Poisson distribution predicts that lightning is unlike to strike twice in a single document. We shouldn't expect to see two or more instances of *boycott* in the same document (unless there is some sort of hidden dependency that goes beyond the Poisson). But when it rains, it pours. If a document is about boycotts, we shouldn't be surprised to find two *boycott*s or even a half dozen in a single document. The standard use of the Poisson in modeling the distribution of words and ngrams fails to fit the data except where there are almost no interesting hidden dependencies as in the case of *somewhat*.

Why are the deviations from Poisson more salient for ''interesting'' words like *boycott* than for ''boring'' words like *somewhat*? Many applications such as information retrieval, text categorization, author identification and word-sense disambiguation attempt to discriminate documents on the basis of certain hidden variables such as topic, author, genre, style, etc. The more that a keyword (or ngram) deviates from Poisson, the stronger the dependence on hidden variables, and the more useful the keyword (or ngram) is for discriminating documents on the basis of these hidden dependences. Similar arguments apply in a host of other important applications such as text compression and language modeling for speech recognition where it is desirable for word and ngram probabilities to *adapt* appropriately to frequency changes due to various hidden dependencies.

We have used document frequency, *df*, a concept borrowed from Information Retrieval, to find deviations from Poisson behavior. Document frequency is similar to word frequency, but different in a subtle but crucial way. Although inverse document frequency (IDF) and $\log_{10} f$ are extremely highly correlated ($\rho = -0.994$), it would be a mistake to try to model one with a simple transform of the other. Figure 5 showed one such attempt, where f was transformed into a predicted IDF by introducing a Poisson assumption: $\hat{IDF} = -\log_2 (1 - e^{-\theta})$, with $\theta = \frac{f_w}{D}$. Unfortunately, the prediction errors were relatively large for the most important keywords, words with moderate frequencies such as *Germans*.

To get a better look at the subtle differences between document frequency and word frequency, we focused our attention on a set of 53 words that all had approximately the same word frequency in a corpus of 1989 AP stories. Table 1 showed that words with larger IDF tend to have more content. *boycott*, for example, is a better keyword than *somewhat* because it bunches up into a relatively small set of documents. Table 2 showed that variance and entropy can also be used as a measure of content (at least among a set of words with more or less the same word frequency). A good keyword like *boycott* is farther from Poisson (chance) than a crummy keyword like *somewhat* by almost any sense of closeness that one might consider, e.g., IDF, variance, entropy. These crucial deviations from Poisson are robust. We showed in section 4 that deviations from Poisson in one year of the AP can be used to predict deviations in another year of the AP.

Acknowledgments

This work benefited considerably from extensive discussions with Slava Katz.

References

Church, K., and Gale, W. (submitted) *Poisson Mixtures*.

Johnson, N., and Kotz, S. (1969) *Discrete Distributions*, Houghton Mifflin, Boston.

Katz, S. (in preparation).

Mosteller, Fredrick, and David Wallace (1964) *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, Massachusetts.

Salton, G. (1989) *Automatic Text Processing*, Addison-Wesley.

Shannon, C. (1948) ‘‘The Mathematical Theory of Communication,’’ *Bell System Technical Journal*.

Sparck Jones, K. (1972) ‘‘A Statistical Interpretation of Term Specificity and its Application in Retrieval,’’ *Journal of Documentation*, 28:1, pp. 11-21.

van Rijsbergen, C. (1979) *Information Retrieval*, Second Edition, Butterworths, London.