Poisson Mixtures

Kenneth W. Church
William A. Gale

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

Abstract

Shannon (1948) showed that a wide range of practical problems can be reduced to the problem of estimating probability distributions of words and ngrams in text. It has become standard practice in text compression, speech recognition, information retrieval and many other applications of Shannon's theory to introduce a ''bag-of-words'' assumption. But obviously, word rates vary from genre to genre, author to author, topic to topic, document to document, section to section, and paragraph to paragraph. The proposed Poisson mixture captures much of this heterogeneous structure by allowing the Poisson parameter θ to vary over documents subject to a density function ϕ. ϕ is intended to capture dependencies on hidden variables such genre, author, topic, etc. (The Negative Binomial is a well-known special case where ϕ is a Γ distribution.) Poisson mixtures fit the data better than standard Poissons, producing more accurate estimates of the variance over documents (σ^2), entropy (H), inverse document frequency (IDF), and adaptation ($Pr(x \geq 2 | x \geq 1)$).

1. Problem: Word Rates Are Highly Variable

Many applications of statistical natural language processing make use of a so-called ''bag-of-words'' assumption. Of course, is well known that word rates depend on many factors: genre, author, topic, etc. Table 1, for example, shows that ''said'' is more frequent in some types of texts and less frequent in others.

Table 1: Frequency of ''said'' Depends on Source

Source	Freq per million words
Department of Energy Abstracts	41
Groliers Encyclopedia	64
Federalist Papers	287
Hansard	1072
Harper & Row Books	1632
Brown Corpus	1645
Wall Street Journal	5600
Associated Press 1994	8514
Associated Press 1987	9525
Associated Press 1991	9861
Associated Press 1990	10,040
Associated Press 1989	10,195
Associated Press 1988	10,313

The million-word Brown Corpus (Francis and Kucera, 1982) was constructed in the 1960s to help researchers better understand how word rates vary from document to document and genre to genre. The corpus consists of 500 excerpts of approximately 2000 words each, selected from a wide variety of genres: Press (documents 1-88), Religion (89-105), Hobbies (106-141), Popular Lore (142-189), Belle-Lettres (190-264), Government and House Organs (265-294), Learned (295-374), Fiction (375-491), and Humor (492-500). Figures 1 and 2 demonstrate that this structure has a dramatic impact on the frequency of ''said.''

Figure 1 shows the frequency of ''said'' in each of the 500 documents. Figure 2 is similar except that the Brown Corpus was replaced by a corpus of 500 documents randomly generated by a binomial distribution:

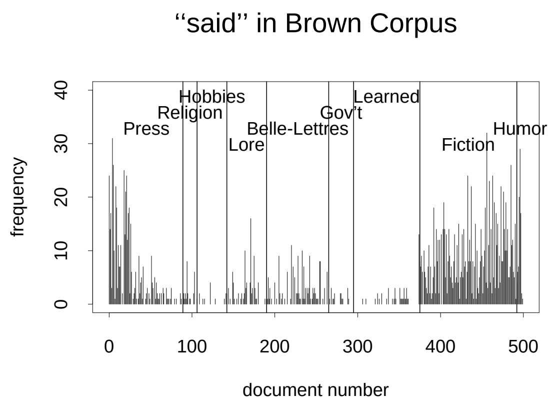

Figure 1: ''said'' is more frequent in some genres and less frequent in others.

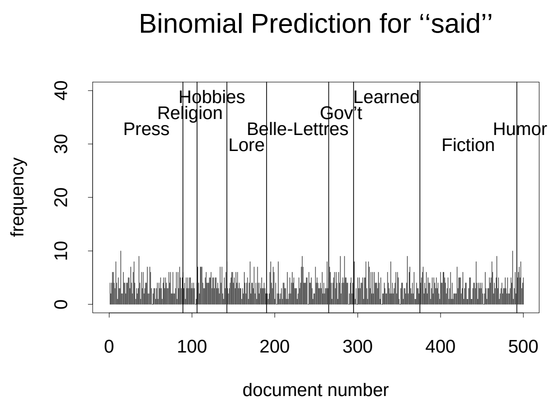

Figure 2: Word rates are relatively constant under a binomial, with no particular preference for Press, Fiction and Humor.

$$Pr_B(k) \;=\; \begin{bmatrix} N \\ k \end{bmatrix} \; P^k \; Q^{N-k} \quad \text{for } k \;=\; 0,1, \cdots ,N \qquad \text{Binomial}$$

The two parameters of the binomial, N, document length, and P, word rate, were chosen to match the Brown Corpus: $N = 2000$ words per document and $P = 3.9/2000$. P is usually interpreted as the probability of a success. In this case, P is the probability that the next word in a document is ''said.'' $Q = 1 - P$ is the probability of a failure (the next word is not ''said''). N is the number of trials, the size of the document.

The frequency varies considerably from one document to the next in Figure 1, much more so than in Figure 2. This observation can be made more precise in terms of the variance, σ^2. The variance is 36 in Figure 1, considerably larger than the variance of 3.9 in Figure 2. It has been our experience that the observed variance of the frequency of a word (or ngram) across documents is almost always larger than the mean, and therefore, larger than what would be expected under either the Binomial or the Poisson. The errors between the observed variance and the Poisson prediction tend to be particularly noticeable for content words in large and diverse collections. We suspect that the binomial and Poisson systematically underestimate the variance because they assume that there are no dependencies on hidden variables such as genre, author, topic, etc., and these factors almost always conspire to inflate the variance over what it would have been if they had not been present.

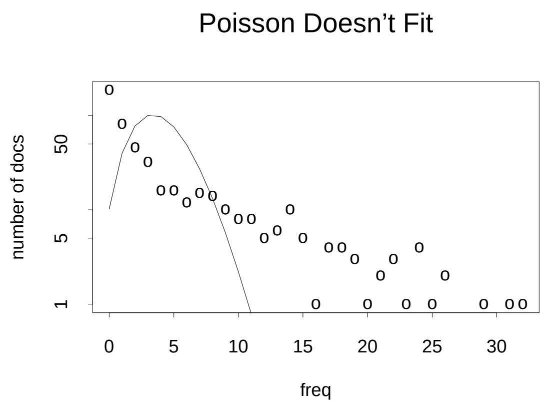

Figure 3: The Poisson (line) does not fit the observed data (circles). The data are the same as those in Figure 1. The line shows a Poisson, $Pr_P(k) = \pi(\theta, k) = \frac{e^{-\theta}\ \theta^k}{k!}$, with $\theta \approx 3.9$ instances of ‘‘said’’ per document. The number of documents for a frequency k is predicted to be $D\ Pr_P(k)$, where D is the number of documents in the collection (500).

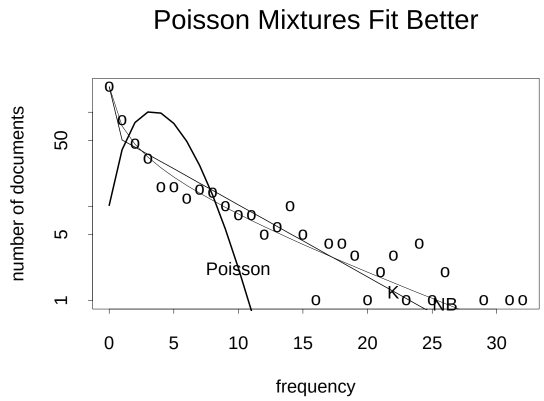

Figure 4: Poisson mixtures fit the data better than simple Poissons. Both the negative binomial (thin line) and the K-mixture (medium thickness line) fall much closer to the observed data points (circles) than the simple Poisson (very thick line). The data are the same as in Figure 1. The negative binomial is: $Pr_{NB}(k) = \begin{bmatrix} N+k-1 \\ k \end{bmatrix} P^k\ Q^{-N-k}$, with $N \approx 0.42$ and $P \approx 9.24$. The K-mixture is: $Pr_{NB}(k) = (1-\alpha)\ \delta_{k,0} + \frac{\alpha}{\beta+1}\ (\frac{\beta}{\beta+1})^k$, with $\alpha \approx 0.75$ and $\beta \approx 5.2$. $\delta_{k,0}$ is 1 when $k=0$, and 0 otherwise.

Figures 3 and 4 show that the Poisson does not fit the data in Figure 1, and that alternatives such as the negative binomial (Mosteller and Wallace, 1964, section 4.1; Johnson and Kotz, 1969, chapter 5) and Katz’ K-mixture (Katz, personal communication) are better. The negative binomial is like a Poisson, but the word rate parameter, θ, is allowed to vary over documents, subject to a density function that models the dependence of θ on all possible combinations of hidden variables such as genre, author, topic, etc. Johnson and Kotz (1969, pp. 135-136) survey a number of applications of the negative binomial in a variety of fields (medicine, psychology, economics and marketing) and conclude that ‘‘the negative binomial is frequently used as a substitute for the Poisson when it is doubtful whether the strict requirements of the Poisson,

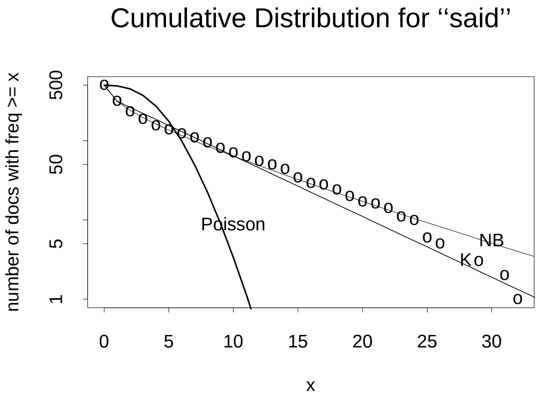

Figure 5: The cumulative probability distribution shows that the negative binomial (thin line) and the K-mixture (medium thickness line) fit the data from Figure 1 (circles) remarkably well, especially when compared with a simple Poisson (very thick line). It might appear from Figure 4 that the the negative binomial and K-mixture undershoot the data at high frequencies, but actually they overshoot the last point by a small amount.

particularly independence, will be satisfied.’’

Even though the mixtures fit the data in Figure 4 better than standard Poisson over the entire frequency range, some readers may be concerned about the tail of the mixtures. The two mixtures in Figure 4 appear to undershoot the data on the right side of the graph, though in fact, they are actually slightly too high, as shown in Figure 5. In any case, the mixtures fit the data much better than the Poisson.

Table 2: More content → Less df

freq	df	Word
140	38	Kennedy
141	62	East
140	68	letter
140	71	production
140	75	son
140	82	Well
141	83	statement
141	90	increased
141	90	results
140	97	thinking
140	99	start
141	99	addition
141	101	showed
141	107	decided
	122	*Binomial* or *Poisson*

2. Some Words Deviate From Binomial More Than Others

‘‘said’’ is not particularly unusual. Table 2 shows 14 words that appeared 140 or 141 times in the Brown Corpus. ‘‘Kennedy,’’ the president of the United States when the Brown Corpus was collected, bunches up into just 38 documents. The less ‘‘loaded’’ words farther down the list tend to be less bunched up, but they are all more bunched up than what would be expected under either a Binomial or a Poisson model. Under the Binomial, $df = D(1-(1-P)^N) \approx 122$, where $D = 500$ documents, $N = 2000$ words per document, and $P = 140$ per million. Under the Poisson, $df = D(1-e^{-\theta}) \approx 122$, where $\theta = 140/500$.

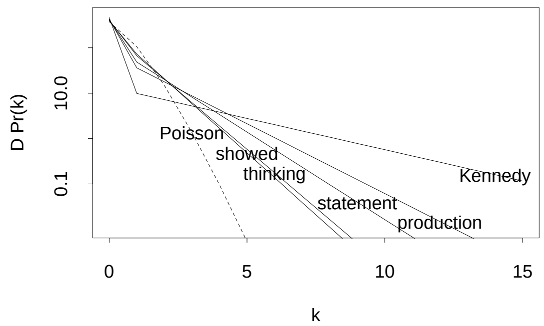

Figure 6: The size of the tail, $Pr(k \geq T)$, varies inversely with *df*. ''Kennedy'' has a large tail because it has a small *df*, and conversely, ''showed'' has a small tail because it has a large *df*. The solid lines show five words fit with the data in Table 2 under a K-mixture model. All five words have smaller *df*s and larger tails than Poisson (dotted line).

The proposed mixture models make use of these differences in *df* to produce improved estimates of $Pr(k)$. Figure 6 shows the K-mixture fits for five words selected from Table 2. The tail, $Pr(k \geq T)$, varies inversely with *df*. ''Kennedy'' has a large tail because it has a small *df*, and conversely, ''showed'' has a small tail because it has a large *df*. In all five cases, though, the tail is larger than Poisson (dotted line). It has been our experience that almost all words (and ngrams) have a larger tail and a smaller *df* than would be expected under a Poisson model.

Figures 7 and 8 examine ''Kennedy'' and ''showed'' in more detail. Figure 7 shows that if ''Kennedy'' is mentioned once in a document (or genre), then it is relatively likely that he will be mentioned again in the same document (or genre). As a result, he is mentioned in fewer documents (and genres) than other words with the same frequency. The bunching-up effect is less pronounced for ''showed,'' but even this word displays strong deviations from Binomial/Poisson behavior, especially in the Learned genre.

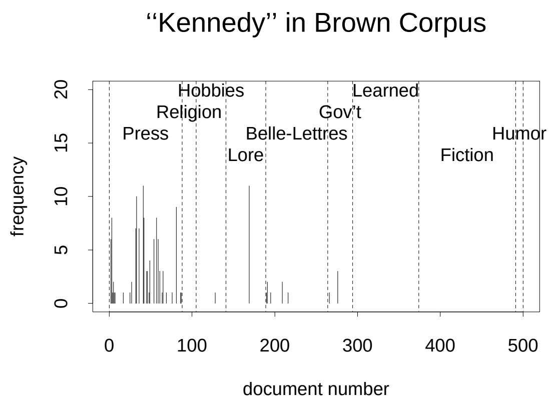

Figure 7: Content words like ''Kennedy'' tend to be very contagious. They do not appear in very many documents/genres, but when they do appear, they are often found in abundance.

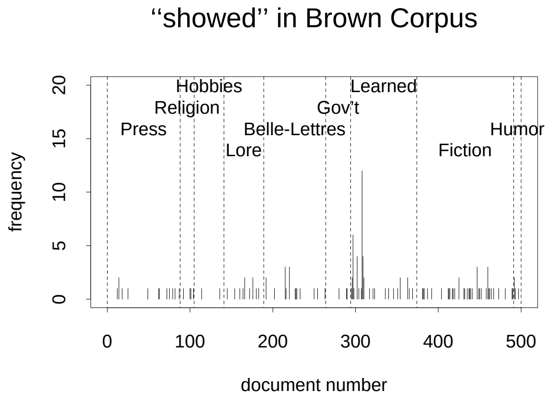

Figure 8: ‘‘showed,’’ a word with relatively little content, is almost as diffuse as a binomial, though there are a few spikes in the Learned genre.

Table 3 is similar to Table 2. Table 3 contrasts words that always appear in the same document ($df = 1$) with words that always appear in different documents ($df = freq$). The first class are much more bursty than the binomial would predict. Note that they tend to be semantically loaded words, e.g., proper nouns, acronyms, technical terms. The second class are as diffuse as would be expected by the binomial. They tend to be relatively devoid of strong dependencies on hidden variables such as the content of the document, topic, genre, etc.

Table 3: Bursty words tend to have more content

freq	df = 1 (bursty)	df = freq (diffuse)
15	Blackman, Dandy, Drug’s, Eugenia, Fromm’s, Hardy’s, Juanita, Selden, Ulyate, collage, tappet	Naturally, Norman, Otherwise, Somehow, Thank, cease, claiming, clue, confident, indispensable, landed, originated, plunged, restricted, sweep, termed
16	Gilborn, Handley, Hanford, Nicolas, Styka, Willis, clover, leveling, secants, thyroglobulin	Already, Back, None, Right, absurd, appearing, collect, delighted, deserves, devised, discussing, faster, inherited, legitimate, lined, link, men’s, persuade, piled, praise, refuse, severely, shops, sole, spreading, thereafter, unnecessary, waved
17	Angie, BOD, Giffen, Krim, Lalaurie, Lizzie, Moreland, Nadine, TSH, Trevelyan, accelerometer	35, Go, K., artificial, capture, consistently, designated, expecting, formally, grasp, lit, obscure, pushing, respective, spontaneous, surprisingly, vitality
18	Andrei, Barco, Helion, Keys, Kitti, Langford, Madden, Saxon, Stevie, Upton, effluent, nonspecific	Beyond, avoided, birthday, emphasized, escaped, gather, instantly, packed, proceed, repeatedly, sixty, submit, surrounded
19	Haney, Killpath, Letch, tetrachloride, tsunami	Which, alike, amazing, bold, happily, notable, overwhelming, remainder, rid, rush, savage, whereby

Thus far, we've mentioned three ways to think about ''variability'' of word rates over documents.

1. Variance (Statistics)
2. Document Frequency (Information Retrieval)
3. ±Content (Linguistics)

The variance is perhaps the most natural way of thinking about variability, especially in the statistics literature. Document Frequency (df) is closely related to Inverse Document Frequency (IDF), a quantity often used in information retrieval (Sparck Jones, 1972). IDF is usually not thought of as an indicator of variability, though it may have certain robustness advantages over variance, since the sample variance is notoriously sensitive to outliers. Three other measures of variability will be introduced shortly: entropy (H), burstiness (B) and adaptation (P21).

3. The Poisson

Let $Pr(k)$ be the probability that a document will have exactly k instances of the term. Let $Pr_P(k)$ be this probability under the Poisson model:

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

3.1 Parameter Estimation

This paper will present a number of theoretical models, the Poisson and various mixtures of Poissons. After presenting each of these models, we will mention a few ways to estimate the n parameters of the model. The simplest method is often the so-called *method of moments* which equates the the theoretical values for the first n moments with their sample estimates, and solves the system of n equations for the n unknowns to obtain estimates for the n parameters. In the case of the Poisson, there is only one parameter, and consequently, there is only one equation. The method of moments simply equates the theoretical value for the mean, θ, with the sample mean, $\overline{t}$, the mean of the term frequencies, $t = t_1, \ldots, t_D$. Each element, t_i, denotes the frequency of the term in the i^{th} document. (For now, we will assume that the documents all have the same length.)

3.2 Variability

Although θ is usually thought of as an estimate of the mean word rate, it can also interpreted as an estimate of the variance of the word rate over documents. Under the Poisson, the mean and the variance are both equal to θ. As mentioned above, one of the problems with the Poisson is that the sample variance is usually larger than the mean, often by a considerable margin. Some of the other models that will be presented soon will allow for more flexibility in the variance.

The variance is just one of many ways to measure the variability of word rates over documents. Five measures will be considered: variance (σ^2), entropy (H), burstiness (B), inverse document frequency (IDF), and adaptation (P21). Under the Poisson, all of these summary statistics (and any others that one might wish to consider) depend on a single parameter, θ. Many have a simple closed form:

$$E_P(k) = \theta \qquad \text{mean}$$

$$\sigma_P^2 = \theta \qquad \text{variance}$$

$$IDF_P = -\log_2 Pr_P(k \ge 1) = -log_2(1 - e^{-\theta}) \qquad \text{inverse document frequency (IDF)}$$

$$H_P = -\sum_{k=0}^{\infty} Pr_P(k) \log_2 Pr_P(k) \qquad \text{entropy (H)}$$

$$B_P = \frac{E_P(k)}{Pr_P(k \ge 1)} = \frac{\theta}{1 - e^{-\theta}} \qquad \text{burstiness (B)}$$

$$Pr_P(k \geq 2 | k \geq 1) = \frac{Pr_P(k \geq 2)}{Pr_P(k \geq 1)} = \frac{1 - e^{-\theta} - \theta e^{-\theta}}{1 - e^{-\theta}} \quad \text{adaptation (P21)}$$

(The notation $Pr(k \geq T)$ is used as a short hand for $\sum_{i \geq T} Pr(i)$.)

As mentioned above, the variance is perhaps the most natural way of thinking about variability. IDF is borrowed from information retrieval (Sparck Jones, 1972). Entropy is found in information theory (Bell, 1990; Cover, 1991). Accurate predictions of entropy are important for many applications of Shannon's theory. Burstiness is Katz' innovation (Katz, personal communication). It is like the mean, except that it ignores documents with no instances of the word. Burstiness is a convenient quantity for working with the K-mixture.

We have found $Pr(k \geq 2 | k \geq 1)$ to be useful for modeling adaptation. Adaptive language models, which have become popular recently in the speech recognition literature (e.g., Lau *et al*, 1993, and references therein). Under standard independence assumptions, it is extremely unlikely that lightning would strike twice (or half a dozen times) in the same document. But text is more like a contagious disease than lightning. If we see one instance of a contagious disease such as tuberculosis in a city, then we would not be surprised to find quite of few more. Similarly, if a few instances of ''said'' have already been observed in a document, then there will probably be some more.

4. Empirical Estimates of Variability

The mean and the five measures of variability can all be estimated empirically from *t*, the vector of term frequences: $t_1, ..., t_D$. Let $Pr_E(x)$ be the fraction of the *D* documents that have a term frequency of *x*.

$$E_E(t) = \bar{t} = \frac{1}{D} \sum_{t=1}^{D} t_i \quad \text{mean}$$

$$\sigma_E^2 = E_E((t - \bar{t})^2) = \frac{1}{D-1} \sum_{i=1}^{D} (t_i - \bar{t})^2 \quad \text{variance}$$

$$IDF_E = -log_2 Pr_E(x \geq 1) \quad \text{inverse document frequency (IDF)}$$

$$H_E = -\sum_{x=0}^{\infty} Pr_E(x) \ \log_2 Pr_E(x) \quad \text{entropy (H)}$$

$$B_E = \frac{\bar{t}}{Pr_E(x \geq 1)} \quad \text{burstiness (B)}$$

$$Pr_E(x \geq 2 | x \geq 1) = \frac{Pr_E(x \geq 2)}{Pr_E(x \geq 1)} \quad \text{adaptation (P21)}$$

Table 4 shows some typical values for these five measures. The words were selected from the Brown Corpus, by extracting all upper case words in a narrow frequency range: 125-150. The list is sorted by variance.

The Poisson does not predict these values very well. The two rows labeled ''Poisson'' were computed from a Poisson with a θ of 0.25 and 0.29. The θs for the 18 words all fall in this range, and therefore, if the Poisson was an appropriate model of these words, the observed variance should be in the range 0.25 – 0.29, the observed IDF should be 2.00 – 2.18, and so on. The data in Figure 4 show, however, that the Poisson systematically overestimates the entropy, and underestimates the other values. The errors are particularly large for content words where there are relatively important dependencies on hidden variables such as genre. ''Church,'' for example, has a large variance because there is a genre on Religion. Similarly, ''Government'' has a large variance because there is a genre on Government and House Organs. ''Island'' might be somewhat more surprising, but it turns out that the Brown Corpus was collected in Rhode Island and consequently, ''Island'' is highly associated with the genre on Government and House Organs. Word rates vary from genre to genre, topic to topic, author to author, document to document, section to section, paragraph to paragraph. These factors tend to decrease the entropy and increase the other test variables.

Table 4: Variability is Highly Associated with Content

mean	var	IDF	entropy	burstiness	adaptation	
0.29	3.51	3.45	0.66	3.17	0.50	Government
0.27	2.63	3.53	0.60	3.14	0.44	Island
0.25	2.23	3.68	0.59	3.21	0.54	Church
0.27	1.99	3.65	0.60	3.35	0.55	Federal
0.29	1.80	3.53	0.64	3.35	0.60	Christian
0.28	1.75	3.72	0.59	3.68	0.58	Kennedy
0.26	1.67	3.84	0.55	3.66	0.80	Soviet
0.28	1.52	3.02	0.78	2.27	0.42	East
0.29	1.31	2.71	0.86	1.93	0.36	William
0.29	1.24	2.90	0.82	2.15	0.39	North
0.28	1.17	2.81	0.83	1.96	0.37	French
0.26	0.96	2.89	0.79	1.90	0.32	George
0.27	0.72	2.70	0.86	1.74	0.35	City
0.26	0.66	2.64	0.87	1.65	0.35	During
0.28	0.58	2.61	0.92	1.71	0.43	Well
0.25	0.54	2.71	0.85	1.64	0.39	I’ve
0.27	0.43	2.37	0.94	1.40	0.28	Yet
0.29	0.38	2.19	0.98	1.31	0.25	Here
0.25	0.25	2.18	0.89	1.13	0.12	*Poisson*
0.29	0.29	2.00	0.97	1.15	0.14	*Poisson*

The columns in Table 4 are highly correlated, as shown in Table 5. According to the Poisson, all of these values should be predictable from θ, and therefore there should be no systematic structure within a group of words that share the same θ. However, it has been our experience that large correlations such as those in Table 5 are to be expected within a group of words (or ngrams) with similar means.

Table 5: Pairwise Correlations of Values in Table 4

	variance	IDF	entropy	burstiness	adaptation
variance		0.80	–0.80	0.81	0.57
IDF	0.80		–0.99	0.98	0.88
entropy	–0.80	–0.99		–0.98	–0.85
burstiness	0.81	0.98	–0.98		0.89
adaptation	0.57	0.88	–0.85	0.89	

Table 6 shows the IDF and variance in five years of Associated Press (AP) for the same set of words (except for *I’ve*).[1] Table 7 shows that the estimates in one year are good predictors of the estimates in another year. IDF tends to have larger correlations from one year to the next than variance, indicating that IDF is more somewhat more robust than variance. Estimates tend to degrade over time, as indicated by larger correlations in adjacent years than in non-adjacent years. If you want to predict the IDF in next year’s AP, it is better to use data from last year, than for data from a decade ago. The time structure varies from one word to another, of course. The IDF for *George* was relatively low in 1988 when George Bush was running for President. Similarly, the IDF for *East* was relatively low in 1989 and 1990 when the Berlin Wall was falling down.

1. The tools for accessing the AP corpus made it difficult to work with *I’ve.*

Table 6: Estimates from Five Different Years of Associated Press (AP) Newswire

IDF					Variance					
1988	1989	1990	1991	1992	1988	1989	1990	1991	1992	
7.13	7.03	6.85	6.48	6.47	0.01	0.01	0.01	0.01	0.02	Yet
6.46	6.60	6.61	6.42	6.42	0.03	0.02	0.03	0.04	0.03	Well
5.74	5.58	5.70	5.92	5.93	0.03	0.03	0.03	0.02	0.02	Government
5.36	5.28	5.40	5.43	5.51	0.07	0.09	0.07	0.08	0.05	Island
5.24	5.37	5.50	5.56	5.61	0.11	0.12	0.08	0.08	0.07	Church
5.29	5.09	5.20	5.73	5.57	0.10	0.36	0.21	0.06	0.07	Christian
5.28	5.66	5.56	5.26	5.42	0.42	0.14	0.15	0.48	0.19	Kennedy
4.83	4.71	4.49	4.44	4.39	0.20	0.24	0.25	0.23	0.26	French
4.73	5.01	4.83	4.58	4.23	0.05	0.03	0.14	0.20	0.16	Here
4.56	4.61	4.61	4.58	4.76	0.05	0.05	0.05	0.05	0.04	During
4.07	3.79	3.24	3.49	4.21	0.27	1.53	1.44	0.37	0.18	East
3.99	3.75	3.99	3.86	4.33	0.17	0.22	0.16	0.17	0.11	Federal
3.94	3.95	4.00	3.98	3.92	0.68	1.33	0.56	0.59	0.41	North
3.78	3.66	3.90	3.85	4.10	0.11	0.13	0.11	0.11	0.09	William
3.47	3.51	3.54	3.61	3.69	0.31	0.38	0.39	0.41	0.32	City
3.29	3.21	3.04	3.06	3.66	2.69	2.44	3.24	3.05	0.65	Soviet
2.98	4.01	4.28	4.29	3.86	0.32	0.13	0.09	0.12	0.29	George

Table 7: Correlations of Columns in Table 6

IDF					Variance					
1988	1989	1990	1991	1992	1988	1989	1990	1991	1992	
1.00	0.96	0.93	0.92	0.95	1.00	0.85	0.92	0.99	0.86	1988
0.96	1.00	0.98	0.96	0.94	0.85	1.00	0.94	0.85	0.80	1989
0.93	0.98	1.00	0.98	0.94	0.92	0.94	1.00	0.94	0.78	1990
0.92	0.96	0.98	1.00	0.96	0.99	0.85	0.94	1.00	0.84	1991
0.95	0.94	0.94	0.96	1.00	0.86	0.80	0.78	0.84	1.00	1992

Table 8 shows the residual IDF, the difference between the observed IDF and the value predicted under a Poisson model. Table 8 also shows the residual variance.

$$\textit{Residual IDF} = IDF_E - IDF_P = IDF_E + \log_2(1 - e^{-\theta})$$
$$\textit{Residual } \sigma = \sigma_E - \sigma_P = \sigma_E - \theta$$

The residuals in Table 8 are large and systematic. They are so systematic that they can be used to distinguish better keywords from less good keywords. It is customary in Information Retrieval to weight words by IDF, but we conjecture that Residual IDF might be even better. Consider *Here* and *French*. They have roughly the same IDF, but quite different Residual IDFs. Intuitively, *French* seems to be a better keyword. If a document mentions the word *Here*, it could be about practically anything. The words in Table 6 are sorted by 1988 IDF and the words in Table 8 are sorted by the 1988 Residual IDF values. We believe that the words at the top of Table 8 tend to be better keywords than the words at the top of Table 6.

Moreover, Residual IDF does a better job of weighting words across years. Consider the words *George* and *East*. In both cases, IDF is relatively large when Residual IDF is relatively small, and vice versa. The Residual IDF values seem to make more sense. The weight for *George* for example ought to be relatively large in the election years of 1988 and 1992, and the weight for *East* ought to be relatively large in 1989 and 1990 when East Germany merged with the west.

The magnitude of the deviations from Poisson depends on many factors, some linguistic and some not. The deviations are more noticeable in larger and more diverse collections, especially for words that are associated with the particular type of diversity that is most salient in the collection. The effects are larger for content words, presumably because variations in content tend to dominate other factors such as stylistic variation. In general, more semantic content leads to more variance, more IDF, less entropy, more

Table 8: Residuals from Poisson

Residual IDF					Residual Variance					
1988	1989	1990	1991	1992	1988	1989	1990	1991	1992	
1.68	1.57	1.64	1.62	1.05	2.29	2.05	2.77	2.59	0.47	Soviet
1.20	0.88	0.81	1.39	0.88	0.36	0.10	0.11	0.41	0.14	Kennedy
0.94	1.20	0.79	0.83	0.73	0.55	1.17	0.44	0.47	0.29	North
0.80	0.84	0.77	0.72	0.75	0.14	0.17	0.17	0.15	0.18	French
0.60	0.66	0.60	0.57	0.62	0.06	0.08	0.05	0.05	0.04	Church
0.59	1.26	1.05	0.61	0.51	0.18	1.34	1.19	0.22	0.10	East
0.58	1.24	1.01	0.50	0.49	0.06	0.29	0.15	0.03	0.04	Christian
0.51	0.58	0.60	0.61	0.61	0.18	0.23	0.25	0.27	0.19	City
0.46	0.53	0.50	0.52	0.43	0.03	0.05	0.04	0.05	0.02	Island
0.37	0.38	0.34	0.37	0.30	0.09	0.11	0.08	0.08	0.04	Federal
0.35	0.24	0.20	0.28	0.46	0.15	0.05	0.03	0.05	0.19	George
0.29	0.33	0.54	0.61	0.35	0.01	0.01	0.02	0.02	0.01	Well
0.16	0.06	0.05	0.09	0.20	0.00	0.00	0.00	0.00	0.00	Yet
0.15	0.18	0.20	0.18	0.17	0.03	0.04	0.03	0.03	0.02	William
0.15	0.13	0.12	0.11	0.12	0.01	0.00	0.01	0.00	0.00	Government
0.05	0.07	0.07	0.07	0.05	0.01	0.00	0.00	0.00	0.00	During
0.05	0.02	0.23	0.30	0.22	0.01	0.00	0.10	0.14	0.10	Here

burstiness and more adaptation.

5. The Two Poisson Model

The Two Poisson Model is a simple example of a Poisson mixture.

$$Pr_{2P}(x) = \alpha\ \pi(x, \theta_1) + (1-\alpha)\ \pi(x, \theta_2)$$

The Two-Poisson model is used in the Information Retrieval literature (Bookstein and Swanson, 1974; Harter, 1975) to account for the fact that relevant documents tend to have different frequencies from irrelevant documents.

Harter (1975) showed how to use the method of moments to fit the three parameters of the Two-Poisson model, θ_1, θ_2 and α, from the first three moments. Let R_i be the i^{th} moment around zero. It can be estimated empirically by $R_i \approx \sum_k k^i\ Pr_E(k)$. $\alpha = \frac{R_1 - \theta_2}{\theta_1 - \theta_2}$. θ_1 and θ_2 are the roots of the quadratic equation: $a\theta^2 + b\theta + c = 0$, where:

$$a = R_1^2 + R_1 - R_2$$
$$b = R_1^2 - R_1 R_2 + 2R_1 - 3R_2 + R_3$$
$$c = R_2^2 - R_1^2 + R_1 R_2 - R_1 R_3$$

Tables 9-10 show that two Poissons outperform a single Poisson. The obs(erved) values in Table 9 are borrowed from Table 4. The est(imated) values were computed with the Two Poisson model. The Two Poisson model has smaller errors than the standard Poisson model. The err(ors) are computed by taking the difference between the estimated values and the observed values. The RMS (root mean square) errors in Table 10 are computed by:

$$err = \sqrt{\sum_{word} (est - obs)^2} \qquad \text{RMS error}$$

The RMS errors show that the Two Poisson model outperforms the Poisson in all three respects (IDF, Entropy and Adaptation), though the improvement in entropy is particularly impressive.

Table 9: Estimation Errors Under Two Poisson

IDF			Entropy			Adaptation			
est	obs	err	est	obs	err	est	obs	err	
2.64	3.45	–0.80	0.75	0.66	0.09	0.11	0.50	–0.39	Government
3.04	3.53	–0.50	0.66	0.60	0.06	0.14	0.44	–0.30	Island
2.87	3.68	–0.80	0.69	0.59	0.10	0.11	0.54	–0.43	Church
3.19	3.65	–0.46	0.65	0.60	0.05	0.19	0.55	–0.36	Federal
3.19	3.53	–0.34	0.69	0.64	0.05	0.27	0.60	–0.33	Christian
3.76	3.72	0.03	0.58	0.59	–0.01	0.49	0.58	–0.09	Kennedy
3.18	3.84	–0.66	0.65	0.55	0.10	0.19	0.80	–0.61	Soviet
2.67	3.02	–0.35	0.79	0.78	0.01	0.15	0.42	–0.27	East
2.41	2.71	–0.30	0.87	0.86	0.01	0.14	0.36	–0.22	William
2.68	2.90	–0.22	0.82	0.82	0.00	0.19	0.39	–0.20	North
2.52	2.81	–0.29	0.84	0.83	0.01	0.15	0.37	–0.22	French
2.75	2.89	–0.13	0.80	0.79	0.01	0.19	0.32	–0.13	George
2.65	2.70	–0.04	0.86	0.86	0.00	0.24	0.35	–0.11	City
2.49	2.64	–0.15	0.86	0.87	–0.01	0.16	0.35	–0.19	During
2.57	2.61	–0.04	0.93	0.92	0.01	0.39	0.43	–0.04	Well
2.62	2.71	–0.10	0.86	0.85	0.01	0.24	0.39	–0.15	I’ve
2.36	2.37	–0.01	0.94	0.94	0.00	0.25	0.28	–0.03	Yet
2.20	2.19	0.01	0.98	0.98	0.00	0.24	0.25	–0.01	Here

Table 10: RMS (Root Mean Square) Errors for Words in Table 9

	Poisson θ=0.25	Poisson θ=0.29	Two Poisson
IDF	1.00	1.17	0.39
Entropy	0.39	0.41	0.05
Adaptation	0.35	0.33	0.27

Unfortunately, two Poissons are probably not enough, as illustrated in Figure 9. Bookstein and Swanson (1974, p. 317) came to the same conclusion, and suggested a Three Poisson model, though they noted that it would require even more parameters than the Two Poisson model. They then suggested the negative binomial, which can be viewed as a continuous mixture of infinitely many Poissons:

> ‘‘The multiple Poisson distribution may, of course, be generalized to a continuous association strength... The negative binomial distribution, which was examined by Mosteller and Wallace, is of this form...’’ (Bookstein and Swanson, 1974, p. 317)

6. Poisson Mixtures

Poisson Mixtures can be thought of as a generalization of the Two Poisson Model where the mixing parameter, α, is replaced with an arbitrary density function, ϕ. The density function ϕ is intended to capture dependencies on hidden variables such as genre, topic, author, etc. The general form of a Poisson mixture is:

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

where π is a Poisson:

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

and ϕ is an arbitrary density function. A density function should integrate to 1. That is, $\int_0^\infty \phi(\theta)\, d\theta = 1$.

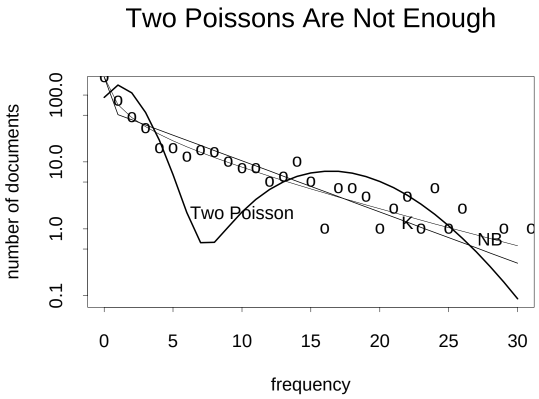

Figure 9: The Two Poisson Model (very thick line) often exhibits a serious dropout between the two θs. In this case, the two θs are 1.53 and 16.9; the mixing parameter, α, is 0.85. The negative binomial (thin line), the K-mixture (medium thickness line) and data (circles) are the same as in Figure 4.

Three special cases of ϕ will be discussed:

$$\phi_{2P}(\theta) = \alpha\ \delta(\theta-\theta_1) + (1-\alpha)\ \delta(\theta-\theta_2) \qquad \text{Two Poisson}$$

$$Pr_{2P}(x) = \alpha\ \pi(x,\theta_1) + (1-\alpha)\ \pi(x,\theta_2)$$

$$\phi_{NB}(\theta) = \frac{\theta^{N-1} e^{-\frac{\theta}{P}}}{P^N \Gamma(N)} \qquad \text{Negative Binomial}$$

$$Pr_{NB}(x) = \begin{bmatrix} N+x-1 \\ x \end{bmatrix} P^x\ Q^{-N-x}$$

$$\phi_K(\theta) = (1-\alpha)\ \delta(\theta) + \frac{\alpha}{\beta}\ e^{-\frac{\theta}{\beta}} \qquad \text{K-mixture}$$

$$Pr_K(x) = (1-\alpha)\ \delta_{x,0} + \frac{\alpha}{\beta+1}\ \left(\frac{\beta}{\beta+1}\right)^x$$

where $\delta(x)$ is Dirac's delta function; the density function is ∞ when $x=0$, and otherwise, 0. $\delta_{x,y}$ is 1 when $x=y$, and otherwise, 0. ϕ_{NB} is known as a Gamma distribution.

7. The Negative Binomial

The negative binomial is so-named by analogy with the binomial. Both $Pr_B(k)$ and $Pr_{NB}(k)$ can be formulated as the k^{th} term of a binomial expansion:

$$(P+Q)^N = \sum_{k=0}^{N} \begin{bmatrix} N \\ k \end{bmatrix} P^k\ Q^{N-k} \qquad \text{Binomial}$$

$$Pr_B(k) = \begin{bmatrix} N \\ k \end{bmatrix} P^k\ Q^{N-k} \quad \text{for } k = 0,1,\cdots,N$$

$$(Q-P)^{-N} = \sum_{k=0}^{\infty} \begin{bmatrix} N+k-1 \\ k \end{bmatrix} P^k \ Q^{-N-k} \qquad \text{Negative Binomial}$$

$$Pr_{NB}(k) = \begin{bmatrix} N+k-1 \\ k \end{bmatrix} P^k \ Q^{-N-k} \quad \text{for } k = 0, 1, \cdots$$

To assure that the probabilities sum to one, it is required that $P + Q = 1$ in the binomial case, and that $Q - P = 1$ in the negative binomial case. Both P and Q should be positive and non-zero. Consequently, $0 < Q < 1$ in the binomial case, and $Q > 1$ in the negative binomial case.

Under the binomial model, P is interpreted as the probability of success (e.g., the next word is ‘‘said’’), Q is the probability of failure (e.g., the next word is not ‘‘said’’), and N is the number of trials (e.g., the size of the document). These interpretations do not apply in the case of the negative binomial, where Q is greater than one, and N need not be an integer.

7.1 Variability

The mean word rate and all five measures of variability can be expressed in terms of the two parameters of the negative binomial: N and P. For the negative binomial,

$$E_{NB}(k) = NP \qquad \text{mean}$$

$$\sigma^2_{NB} = NPQ \qquad \text{variance}$$

$$IDF_{NB} = -\log_2 Pr_{NB}(x \geq 1) = -\log_2(1 - Q^{-N}) \qquad \text{inverse document frequency (IDF)}$$

$$H_{NB} = -\sum_{k=0}^{\infty} Pr_{NB}(k) \log_2 Pr_{NB}(k) \qquad \text{entropy (H)}$$

$$B_{NB} = \frac{NP}{1 - Q^{-N}} \qquad \text{burstiness (B)}$$

$$Pr_{NB}(k \geq 2 \mid k \geq 1) = \frac{Pr_{NB}(k \geq 2)}{Pr_{NB}(k \geq 1)} = \frac{1 - Q^{-N} - NPQ^{-N-1}}{1 - Q^{-N}} \qquad \text{adaptation}$$

Note that both the binomial and the negative binomial have the same symbolic expression for their mean (NP) and variance (NPQ). However, Q is less than one in the binomial, and greater than one in the negative binomial. Consequently, the variance is less than the mean in the binomial, and greater than the mean in the negative binomial. As mentioned previously, the sample variance is almost always larger than the sample mean, and consequently, the negative binomial is almost always more appropriate than the binomial.

7.2 Parameter Estimation

Johnson and Kotz (1969, chapter 5) present a series of methods for estimating the parameters of the negative binomial distribution. It should be possible to estimate N and P from more or less any pair of the six parameters that have been discussed here: mean, variance, IDF, burstiness, adaptation and entropy. Johnson and Kotz’ method 1 uses the mean and variance; their method 2 uses the mean and IDF. Method 1 is simply the method of moments; it equates the sample mean ($\bar{t}$) and variance (σ^2_E) with their theoretical values:

$$\bar{t} = \hat{N}\,\hat{P} \qquad \sigma^2_E = \hat{N}\,\hat{P}\,(1 + \hat{P})$$

and solves for $\hat{N}$ and $\hat{P}$:

$$\hat{P} = \frac{\sigma^2}{\bar{t}} - 1 \qquad \hat{N} = \frac{\bar{t}}{\hat{P}}$$

These equations cannot be solved if $\bar{t} \geq \sigma^2_E$, because the negative binomial is inappropriate in this case. We fall back to Poisson when the sample mean exceeds the sample variance. This usually happens when we have insufficient data, e.g., just one observation of the word in just one document.

Tables 11 and 12 show that the Negative Binomial outperforms both the Poisson and the Two Poisson for

Table 11: Estimation Errors Under Negative Binomial (Method 1)

IDF est	IDF obs	IDF err	Entropy est	Entropy obs	Entropy err	Adaptation est	Adaptation obs	Adaptation err	
3.97	3.45	0.53	0.55	0.66	–0.11	0.64	0.50	0.14	Government
3.88	3.53	0.35	0.57	0.60	–0.03	0.62	0.44	0.18	Island
3.91	3.68	0.23	0.55	0.59	–0.04	0.61	0.54	0.07	Church
3.64	3.65	–0.01	0.63	0.60	0.03	0.59	0.55	0.04	Federal
3.38	3.53	–0.15	0.72	0.64	0.08	0.57	0.60	–0.03	Christian
3.43	3.72	–0.30	0.70	0.59	0.11	0.56	0.58	–0.02	Kennedy
3.58	3.84	–0.26	0.64	0.55	0.09	0.57	0.80	–0.23	Soviet
3.29	3.02	0.27	0.74	0.78	–0.04	0.54	0.42	0.12	East
3.07	2.71	0.36	0.80	0.86	–0.06	0.51	0.36	0.15	William
3.06	2.90	0.16	0.81	0.82	–0.01	0.51	0.39	0.12	North
3.09	2.81	0.28	0.79	0.83	–0.04	0.50	0.37	0.13	French
3.09	2.89	0.20	0.78	0.79	–0.01	0.48	0.32	0.16	George
2.77	2.70	0.08	0.87	0.86	0.01	0.41	0.35	0.06	City
2.76	2.64	0.12	0.86	0.87	–0.01	0.39	0.35	0.04	During
2.53	2.61	–0.08	0.93	0.92	0.01	0.35	0.43	–0.08	Well
2.72	2.71	0.00	0.86	0.85	0.01	0.36	0.39	–0.03	I've
2.38	2.37	0.02	0.94	0.94	0.00	0.28	0.28	0.00	Yet
2.17	2.19	–0.02	0.99	0.98	0.01	0.23	0.25	–0.02	Here

Table 12: RMS (Root Mean Square) Errors

	Poisson θ=0.25	Poisson θ=0.25	Two Poisson	Negative Binomial Method 1	Negative Binomial Method 2
IDF	1.00	1.17	0.39	0.24	
Entropy	0.39	0.41	0.05	0.05	0.03
Adaptation	0.35	0.33	0.27	0.11	0.08

the 18 words in Table 4. Although the method of moments is perhaps the simplest method to implement, it has been our experience that method 2 produces more robust estimates because the sample variance is notoriously sensitive to outliers. This may explain the modest improvement found in Table 12, where RMS error was reduced from 0.05 to 0.03 for entropy, and from 0.11 to 0.08 for adaptation. Method 2 simply equates the sample mean and IDF with their theoretical values, and solves for $\hat{N}$ and $\hat{P}$. That is,

$$\bar{t} = \hat{N}\,\hat{P} \qquad -log_2 Pr_E(x \geq 1) = -log_2(1-\hat{Q}^{-\hat{N}})$$

From these equations, it follows that $\hat{P}$ should satisfy the constraint: $\dfrac{\hat{P}}{\log(1+\hat{P})} = \dfrac{\bar{t}}{-\log Pr_E(0)}$. Unfortunately, there is no closed form solution for $\hat{P}$. The following fragment of C code (Katz, personal communication) uses an iterative approximation. The input parameters `tbar` and `f0` correspond to $\bar{t}$ and $Pr_E(0)$, respectively.

```
#define PRECISION 1e-6

double
P_by_method2(tbar, f0)
     double tbar, f0;
{
  double c = tbar/-log(f0);
  double a = c-1;
  double b = c*c;
  do {
    double r = (a+b)/2;
    if (r-c*log(1+r)<0) a=r;
    else b=r;
  } while (2 * (b-a)/(b+a) > PRECISION );
  return((a+b)/2);
}
```

8. The K-mixture

Recall that the K-mixture is:

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

with $0 < \alpha < 1$ and $\beta > 0$.

The K-mixture is very close to the negative binomial, $Pr_K(x) \approx Pr_{NB}(x)$; the two are often closer to one another than either is to the data. This is convenient because the K-mixture can be easier to work with.

8.1 Variability

The mean word rate and all five measures of variability can be expressed in terms of the two parameters of the K-mixture: α and β.

$$E_K(k) = \alpha\ \beta \qquad \text{mean}$$

$$\sigma_K^2 = \alpha\ \beta\ [(2-\alpha)\beta+1] \qquad \text{variance}$$

$$IDF_K = -\log_2 Pr(x \geq 1) = \log_2 \frac{\beta+1}{\alpha\beta} \qquad \text{inverse document frequency (IDF)}$$

$$H_K = \alpha\beta \log_2 \frac{\beta+1}{\beta} + \alpha \frac{\beta}{\beta+1} \log_2 \frac{\beta+1}{\alpha} - Pr_K(0) \log_2 Pr_K(0) \qquad \text{entropy (H)}$$

$$B_K = \beta+1 \qquad \text{burstiness (B)}$$

$$Pr_K(x \geq 2 | x \geq 1) = \frac{\beta}{\beta+1} \qquad \text{adaptation}$$

In fact, the adaptation property is even stronger for the K-mixture. Unlike other models considered in this paper, the K-mixture predicts that adaptation is constant for $k \geq 2$:

$$Pr_K(x \geq k | x \geq k-1) = \frac{\beta}{\beta+1} \quad \text{for } k = 2,3, \cdots \qquad \text{constant adaptation}$$

The K-mixture also has a simple closed form expression for the tail probability:

$$Pr(x \geq k) = \alpha\left(\frac{\beta}{\beta+1}\right)^k \quad \text{for } k = 1,2, \cdots \qquad \text{tail probability}$$

8.2 Parameter Estimation

α and β can be estimated by the method of moments. That is, equate the sample mean and variance with the theoretical values, and solve for $\hat{\alpha}$ and $\hat{\beta}$. As with the negative binomial, the method of moments produces out-of-range parameter estimates when the distribution is inappropriate. We use the Poisson when this happens.

However, it is much easier to estimate α and β from the sample mean ($\bar{t}$) and the sample estimate of burstiness (B_E), since burstiness has a particularly convenient form: $B_K = \beta + 1$. We simply assume that the sample mean and the sample estimate of burstiness are the same as the theoretical values. We equate the two and solve for $\hat{\alpha}$ and $\hat{\beta}$.

$$\hat{\beta} = B_E - 1 \quad \hat{\alpha} = \frac{\bar{t}}{\hat{\beta}} \qquad \text{method of burstiness}$$

In this respect, the K-mixture is much more convenient than the negative binomial. The parameters of the K-mixture, α and β, are very closely related to the observable quantities, $\bar{t}$ and B_E. In contrast, the relationship between the parameters of the negative binomial, N and P, and $\bar{t}$ and B is much less obvious. The equations relating N and P to $\bar{t}$ and B_E do not have a closed form solution, and must be solved by iterative approximation.

The Negative Binomial (method 2) and K-mixture (method of bustiness) outperform the Poisson (method of moments) and the Two Poisson (method of moments), as can be seen by comparing the RMS errors in Table 13 with the RMS errors in Table 12. (The test on IDF in Tables 11-12 was replaced with a test on variance, because the variance was used to fit the parameters in Tables 11-12 and IDF was used to fit the parameters in Table 13.)

Table 13 suggests that the Negative Binomial is slightly superior to the K-mixture, though it has been our experience that the two methods are extremely similar to one another. The two methods are often closer to one another than either is to the data. The errors between the two methods are highly correlated. That is, the first pair of columns in Table 13 has a correlation of 0.88, the second pair has a correlation of 0.97, and the last pair has a correlation of 0.87. Although the mixtures fit the data better than the alternatives that we have considered, these large correlations indicate that there is still plenty of room for improvement.

9. Variable Length Documents

Thus far, we have been ignoring the fact that some documents are longer than others. Document lengths are modeled by adding a second parameter to the general form of the Poisson mixture. Let w be the length of the document. Then, the general form of the mixture becomes:

$$Pr(x,w) = \int_0^\infty \phi(\theta') \, \pi(w\theta', x) \, d\theta' \quad \text{for } x = 0, 1, \cdots$$

The three special cases become:

$$\phi_{2P}(\theta') = \alpha \, \delta(\theta' - \theta'_1) + (1-\alpha) \, \delta(\theta' - \theta'_2) \qquad \text{Two Poisson}$$

$$Pr_{2P}(x,w) = \alpha \, \pi(x, w\theta'_1) + (1-\alpha) \, \pi(x, w\theta'_2)$$

$$\phi_{NB}(\theta') = \frac{\theta'^{N'-1} e^{-\frac{\theta'}{P'}}}{P'^{N'} \Gamma(N')} \qquad \text{Negative Binomial}$$

$$Pr_{NB}(x,w) = \binom{N'+x-1}{x} (wP')^x \, Q^{-N'-x}$$

Table 13: Errors in K-Mixture and Negative Binomial

Variance		Entropy		Adaptation		
K-mix	NB	K-mix	NB	K-mix	NB	
–2.026	–1.516	0.054	0.044	0.18	0.08	Government
–1.284	–0.824	0.066	0.056	0.238	0.128	Island
–0.942	–0.482	0.035	0.025	0.152	0.042	Church
–0.532	0.008	0.036	0.026	0.15	0.04	Federal
–0.237	0.343	0.037	0.027	0.095	–0.015	Christian
–0.053	0.667	0.027	0.027	0.151	0.031	Kennedy
–0.12	0.53	0.033	0.033	–0.07	–0.2	Soviet
–0.598	–0.418	0.044	0.044	0.141	0.071	East
–0.546	–0.456	0.047	0.047	0.125	0.075	William
–0.374	–0.234	0.043	0.033	0.142	0.082	North
–0.441	–0.341	0.04	0.03	0.124	0.074	French
–0.307	–0.227	0.041	0.041	0.146	0.096	George
–0.128	–0.068	0.017	0.017	0.079	0.039	City
–0.124	–0.084	0.017	0.017	0.04	0.01	During
0.021	0.071	0.004	0.004	–0.017	–0.047	Well
–0.035	0.005	0.014	0.014	–0.005	–0.035	I've
–0.011	–0.001	0.001	0.001	0.012	0.002	Yet
0.003	0.003	0	0	–0.008	–0.008	Here
0.671	0.510	0.036	0.031	0.123	0.076	*RMS Error*

$$\phi_K(\theta') = (1-\alpha')\ \delta(\theta') + \frac{\alpha'}{\beta'}\ e^{-\frac{\theta'}{\beta'}} \qquad \text{K-mixture}$$

$$Pr_K(x,w) = (1-\alpha')\ \delta_{x,0} + \frac{\alpha'}{w\beta'+1}\ \left(\frac{w\beta'}{w\beta'+1}\right)^x$$

The formulas stated previously for the mean and the five measures of variability can be generalized to account for variable length documents by introducing the constraints:

$$\theta_1 = w\theta'_1 \quad \theta_2 = w\theta'_2 \qquad \text{Two Poisson}$$

$$N = N' \quad P = wP' \qquad \text{Negative Binomial}$$

$$\alpha = \alpha' \quad \beta = w\beta' \qquad \text{K-mixture}$$

10. Possible Applications

Following Shannon (1948), it has become standard practice to assume constant word rates in many important practical applications such as compression, speech recognition and information retrieval. A model of variable word rates such as Poisson mixtures could have important ramifications for many of these applications.

The probability of a string is often estimated by breaking the string up into pieces (e.g., letters, words, ngrams). The probability of the entire string, *p*, is computed by multiplying the probabilities of the pieces, p_i. The practice is often justified by introducing a multinomial assumption which assumes that the p_i's are constant. Of course, as we have seen, word rates are almost never constant. The second instance of a word ought to be less surprising than the first. There ought to be a quantity discount.

Adaptive language models have recently become quite popular in speech recognition. Lau *et al* (1993), for example, introduce a cache to remember the most recent *n* words. The estimates of p_i based on long-term evidence are combined with another set of estimates based on short-term evidence found in the cache.

In principle, it should not be necessary to introduce a cache in order to capture variable word rates. All of the required probabilities are completely determined by Pr(k). Imagine that we have a different distribution $Pr_i(k)$ for each word *i*. Initially, the probability that a word *i* will be seen in a document of length *w* is:

$\frac{Pr_i(k \geq 1)}{w}$. Suppose the first word is j. Then the probability of the next instance of j jumps to $\frac{p_j(k \geq 2 | k \geq 1)}{w-1}$ and the probabilities of the other words are adjusted downward so that all the probabilities sum to one. In this way, it ought to be possible to estimate the probability of a string without introducing a constant word rate assumption or an ad hoc device such as a cache. Of course, much work remains to be done to show that this possibility is practical for a realistic application.

Poisson mixtures are somewhat simpler to apply when it is not necessary to know the detailed order of words within a document. Information retrieval and author identification are two such cases.

10.1 Information Retrieval (IR), Author Identification and Word-Sense Disambiguation

We like to think of the probabilistic retrieval model (van Rijsbergen, 1979, chapter 6; Salton, 1989, section 10.3) as an application of well-understood Bayesian discrimination methods that have been studied extensively by Mosteller and Wallace (1964, section 3.1) in their investigation of the authorship of the Federalist Papers. The discrimination process consists of two phases: a training phase followed by a testing phase. During the training phase, we are given two (or more) classes of documents and are asked to construct a discriminator which can distinguish between the two (or more) classes of documents. These discriminators are then applied to new documents during the testing phase. In the author identification task, for example, the training set consists of several documents written by each of the two (or more) authors, e.g., Hamilton and Madison. The resulting discriminator is then tested on documents whose authorship is disputed. In the information retrieval application, the training set consists of a query (a set of one or more relevant documents and a set of zero or more irrelevant documents). The resulting discriminator is then applied to all documents in the library in order to separate the more relevant ones from the less relevant ones.

There is a wealth of information in the collection of documents that could be used as the basis for discrimination. Yet, it is common practice to treat documents as a homogeneous bag of words, and ignore much of the heterogeneous structure, including linguistic factors such as dependencies on word order and correlations between pairs of words. In other words, we are assuming that there are two (or more) sources of word probabilities, *rel* and $\overline{rel}$, in the IR application, and $author_1$ and $author_2$ in the author identification application. During the training phase, we attempt to estimate $Pr(x_w | s)$, the probability of observing x instances of w in a document generated by source s, for all words w in the vocabulary and all sources s. Then during the testing phase, we score all documents as follows and select high scoring documents as being relatively likely to have been generated by the source of interest.

$$score(doc) = \prod_{w \text{ in doc}} \frac{Pr(x_w | rel)}{Pr(x_w | \overline{rel})} \qquad \text{Information Retreival (IR)}$$

$$score(doc) = \prod_{w \text{ in doc}} \frac{Pr(x_w | author_1)}{Pr(x_w | author_2)} \qquad \text{Author Identification}$$

Mosteller and Wallace's framework can be also used to address a variety of other problems in natural language processing. In Gale *et al.* (1993), we looked at the word-sense disambiguation problem and showed that ambiguous words such as *sentence* could be disambiguated by treating the 100-word context surrounding the word as a ''document.''

$$score(doc) = \prod_{w \text{ in context}} \frac{Pr(x_w | sense_1)}{Pr(x_w | sense_2)} \qquad \text{Word-Sense Disambiguation}$$

The scoring procedure asks whether the context in question is more like the contexts of one sense or more like the contexts of the other sense. Large amounts of testing and training material were obtained by making use of parallel texts, texts such as the Canadian parliamentary debates that are available in multiple languages. We assumed that that if *sentence* was translated into French as *peine* then it was the ''judicial'' sense, and if *sentence* was translated as *phrase* then it was the ''syntactic'' sense. In subsequent work, Yarowsky (1992) has shown how to free the system from the dependence on parallel text as a source of testing and training material.

10.2 An Example of the use of Poisson Mixtures in a Discrimination Task

Mosteller and Wallace (1964) studied the Federalist Papers (1787-1788), a small collection of 85 articles (200,000 words), written by Alexander Hamilton, John Jay and James Madison, on a single question: why the citizens of the State of New York should ratify the proposed constitution. Despite the limited number of authors and subject matter, the articles cover nearly every phase of the proposed constitution. Even the Federalist Papers are far from homogeneous, as demonstrated in Figure 11. The variance of the word ''courts'' in the Federalist papers is 18 times larger than the Poisson would predict.

Suppose, for illustrative purposes, that we wanted to discriminate between documents 43-85 and documents 1-42. Call the two classes: *A* and *B*. As can be seen in Table 13 and Figure 11, the word ''courts'' might be a useful discriminator.[2] Intuitively, ten or more instances of ''courts'' in a document is very strong evidence for class *A*; one or two instances is weak evidence for class *B*. Zero instances of ''courts'' is no evidence either way, since both classes have 33 documents with no instances of ''courts.''

0 ''courts'' $\rightarrow$ no evidence either way
1-2 ''courts'' $\rightarrow$ weak evidence for *B*
10+ ''courts'' $\rightarrow$ strong evidence for *A*

This argument can be made more rigorous in terms of a log likelihood function, λ. $\lambda(x)$ indicates how much we should adjust our belief from one category to the other when presented with *x* instances of a word in a single document. Under the Poisson model, Mosteller and Wallace (1964, p. 122) show that[3]

$$\lambda_P(x) = \log \frac{Pr_P(x|\theta_A)}{Pr_P(x|\theta_B)} = x \log \frac{\theta_A}{\theta_B} + \theta_B - \theta_A \qquad \text{Poisson}$$

where θ_A and θ_B are the parameters of the Poissons for the two categories. This is similar to the document scoring procedure discussed above. It is also similar to what Bookstein (1982, pp. 122-124) referred to as the Two-Poisson Model, though it is somewhat different from what other authors mean when they refer to the Two-Poisson Model.[4] Mosteller and Wallace were concerned about the strict requirements of the Poisson, and so they repeated their study using negative binomials in $\lambda(x)$ instead of Poissons. Following their reasoning, let $\lambda_{2P}(x)$, $\lambda_{NB}(x)$ and $\lambda_K(x)$ be defined analogously to $\lambda_P(x)$, by replacing the Poisson in the numerator and denominator with either Pr_{2P}, Pr_{NB} or Pr_K. (This use of the Two-Poisson model in a likelihood ratio departs from most treatments in the information retrieval literature.) All four $\lambda(x)$ functions are shown in Figure 12, with the parameter estimates:

1. Poisson: $\theta_A \approx 2.6$, $\theta_B \approx 0.37$

---

2. In Mosteller and Wallace's study, content words like ''courts'' were placed on a stop list so that variations in content would not interfere with the author-dependent factors of interest. Information retrieval programs do just the reverse; function words such as ''the'' are put on a stop list so that stylistic factors do not interfere with the content-dependent factors of interest.

3. Mosteller and Wallace actually use the refinement mentioned in the previous section that allows for documents of different lengths, *w*. We have dropped the *w* variable throughout in order to simplify the discussion of the more complicated Poisson mixtures.

4. Harter (1975) was interested in the question of estimating the parameters of the Two-Poisson model from unlabeled training material. In the author identification case, the training material is labeled; we are told which training documents were written by Madison and which ones were written by Hamilton. Mosteller and Wallace could then estimate the two θs by simply counting the word rates in the two classes. In the information retrieval case, though, we may not be told which documents are relevant and which are not. Harter (1975) fit the two θs by assuming the observed word rates were generated by a mixture of relevant and irrelevant documents: $\alpha\pi(\theta_A, k) + (1-\alpha)\pi(\theta_B, k)$, but then he departed from Bookstein and Swanson (1974) and Bookstein (1982), and abandoned the log likelihood framework. Robertson and Walker (1994) also discuss a Two-Poisson model, though their model makes use of the notion of ''eliteness,'' and is therefore quite different from the other versions of the Two-Poission model.

2. Two Poisson: $\theta_{A,1} \approx 0.17$, $\theta_{A,2} \approx 22.2$, $\alpha_A \approx 0.89$, $\theta_{B,1} \approx 0.087$, $\theta_{B,2} \approx 1.52$, $\alpha_B \approx 0.80$
3. Negative Binomial: $N_A \approx 15.3$, $P_A \approx 0.17$, $N_B \approx 2.50$, $P_B \approx 0.15$
4. K-mixture: $\alpha_A \approx 0.23$ and $\beta_A \approx 11.11$, $\alpha_B \approx 0.62$, $\beta_B \approx 0.60$

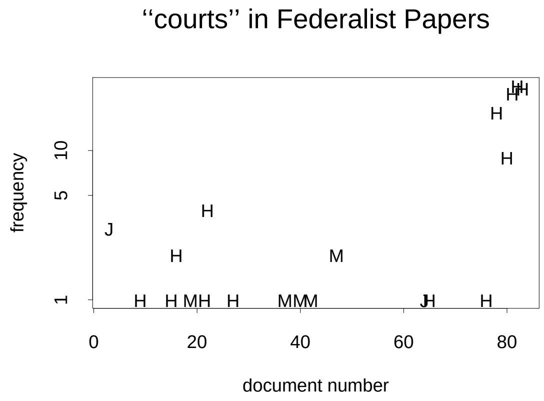

Figure 11: Although the Federalist Papers are smaller and less diverse than the Brown Corpus, the Federalist Papers are far from homogeneous. The word ''courts,'' for example, has a very striking dependence on document number, with several large bursts in the last few documents. There is also a strong dependence on authorship; ''courts'' is more associated with Hamilton (H) than Madison (M) or Jay (J).

Table 13: Data in Figure 11

Class A	doc	47	64	65	76	78	80	81	82	83	
	freq	2	1	1	1	18	9	24	27	26	
Class B	doc	3	9	15	16	20	22	27	37	40	42
	freq	3	1	1	2	1	4	1	1	1	1

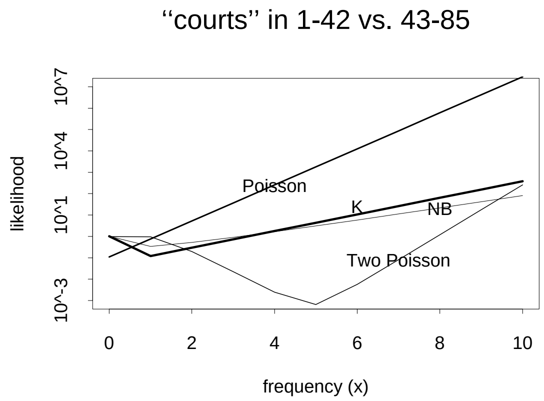

Figure 12: Four log likelihood functions: $\lambda_P(x)$, $\lambda_{2P}(x)$, $\lambda_{NB}(x)$ and $\lambda_K(x)$.

Figure 12 shows that $\lambda_P(x)$ is a straight line, both ends of which are suspect. At high frequencies (e.g., $x \geq 2$), $\lambda_P(x)$ is much too high. $\lambda_P(10)$ favors class A by the incredible odds of 30 million to one! This is especially hard to believe given that we have much less than a million data points. At the other end of the scale, $\lambda_P(0)$ predicts that no instances of ''courts'' is weak evidence (10 to 1 odds) against class A. Again,

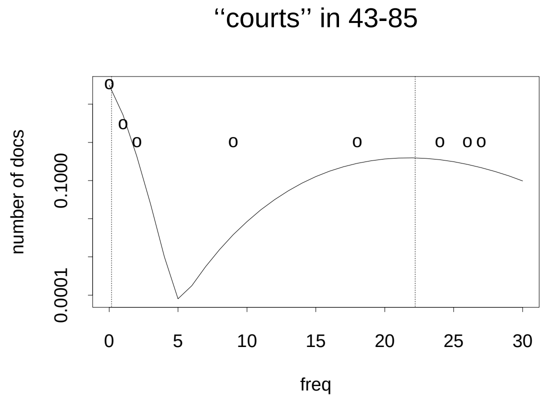

Figure 13: $\lambda_{2P}(x)$ has a dip in the middle ($x=5$) because of the dropout problem illustrated in Figure 9. The Two-Poisson model often exhibits a dropout between the two θs, which are indicated by the vertical lines at 0.17 and 22.2 in this case. A third θ would help to fill in the dropout, but a continuous mixture of infinitely many θs would be even better.

this runs counter to our intuition as stated above. Zero instances should be no evidence either way because both classes have the same number of documents (33) with zero instances of ‘‘courts.’’ Thus we conclude that the Poisson model does not describe $\lambda(x)$ very well; a straight line is too rigid to account for the description presented above.

The mixtures all have more parameters than the Poisson, and consequently, their λs need not be straight. The bent λs are more credible at both ends. Unfortunately, the Two Poisson introduces a spurious dip at $\lambda_{2P}(5)$, because of a dropout between the two θs, as shown in Figure 13. The dropouts in the Two Poisson produce strange undesirable behavior when multiplied and divided in the calculation of log likelihood functions.[5]

The negative binomial and K-mixture fit our intuition better than the Poisson or Two-Poisson. The ends are more credible than λ_P, and there is no embarrassing dip in the middle as in λ_{2P}. The size of the improvement is substantial; λ_P and λ_{2P} can be off by several orders of magnitude over much of the range of interest.

11. Conclusions

Poisson mixtures fit the data better than the standard Poisson. Under the standard Poisson, text is modeled as a homogeneous bag of words with a constant θ across documents, whereas under the proposed mixtures, the heterogeneity of text is modeled by allowing θ to vary over documents, subject to a density function ϕ, designed to capture dependencies on hidden variables such as genre, topic, author, etc. These factors effect the variance (σ^2), entropy (H), burstiness (B), inverse document frequency (IDF), and adaptation ($Pr(x \geq 2 \mid x \geq 1)$) in systematic ways. Hidden variables tend to conspire to decrease the entropy and increase the four other measures of variability over what they would have been otherwise (Poisson). The deviations from Poisson are more noticeable in larger and more diverse collections, especially for words that are associated with the particular type of diversity that is most salient in the collection.

---

5. This may not be a problem for many of the versions of the Two-Poisson Model in Information Retrieval because they generally do not multiply and divide the Two-Poisson Model in this way.

Why are the deviations from Poisson more salient for ‘‘interesting’’ words like *Government* and *courts* than for ‘‘boring’’ functions words like *Here*? Many applications such as information retrieval 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.

In the collections that we have looked at, the effects have been larger for content words (e.g., *Government*) and smaller for function words (e.g., *Here*). The size of the deviations from Poisson can be dramatic. In the previous section, the standard Poisson model predicted the incredible odds of 30 million to one in a case where the truth is probably closer to the negative binomial’s prediction of 82 to 1.

Acknowledgments

This work benefited considerably from extensive discussions with Slava Katz.

References

Bell, T., Cleary, J. and Witten, I. (1990) *Text Compression*, Prentice Hall, New Jersey.

Bookstein, A. (1982), ‘‘Explanation and Generalization of Vector Models in Information,’’ *Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 118-132.

Bookstein, A., and Swanson, D. (1974) ‘‘Probabilistic Models for Automatic Indexing, *Journal of the American Society for Information Science*, 25:5, pp. 312-318.

Cover, T., and Thomas, J. (1991) *Elements of Information Theory*, John Wiley & Sons, Inc., New York.

Francis, W., and Kucera, H. (1982) *Frequency Analysis of English Usage*, Houghton Mifflin Company, Boston.

Gale, W., Church, K. and Yarowsky, D. (1993) ‘‘A Method for Disambiguating Word Senses in a Large Corpus,’’ *Computers and Humanities*, pp. 415-439.

Harter, S. (1975), ‘‘A Probabilistic Approach to Automatic Keyword Indexing: Part I. On the Distribution of Specialty Words in a Technical Literature,’’ *Journal of the American Society for Information Science*, 26(4), 197-206.

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

Katz, S. (personal communication).

Lau, R., Rosenfeld, R. and Roukos, S. (1993) *Adaptive Language Modeling using the Maximum Entropy Principle*, ARPA sponsored workshop on Human Language Technology, Morgan Kaufmann Publishers, San Francisco, CA, ISBN 1-55860-324-7, pp. 108-113.

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

Robertson, S. and Walker, S. (1994) ‘‘Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval,’’ SIGIR, 232-241.

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.

Yarowsky, D. (1992), ‘‘Word-Sense Disambiguation Using Statistical Models of Roget’s Categories Trained on Large Corpora,’’ Coling.