Smoothing, Normalization, Correction and Reduction

David M.W. Powers
School of Informatics and Engineering
Flinders University of South Australia

Introduction

In this paper we analyze a variety of standard techniques used in corpus linguistics and examine a number of issues relating to incorrect usage, the computational infeasibility of various approaches, and the inevitable sampling errors. We motivate much of the paper with applications from information retrieval but the techniques discussed and introduced are far more widely applicable, and some are indeed motivated from other applications of corpus linguistics.

The Singular-Valued Decomposition

We summarize first some of the well known properties and applications of singular-valued decomposition (SVD). This has been popularized recently in the context of information retrieval (IR) as latent semantic indexing or analysis (LSI/LSA) and we will use this application area as the primary example for making the algorithms and theory concrete. We will however refer to other example domains including automatic speech recognition (ASR) and electroencephalographic analysis (EEG).

We will assume that data is provided in matrix $A = a_{ij}$ and make varying assumptions about its normalization or distortion. In general both A and A^T have interpretations as representations of one domain in

terms of another, where the superscript T represent taking the transpose, viz. $A^T = A_{ji}$. Here A represents the relationship between two different aspects of our data, e.g. documents versus words, or sensors (electrodes, microphones, pixels) versus time. We will try to avoid the word ‘dimension’ as much as possible because it can be used to describe both the horizontal and vertical size and indexing of a matrix as well as the different attributes or features of a vector (e.g. a column or row of a matrix) as well as totally different modalities or sources or domains (e.g. electrode, epoch, space, time).

There are different conventions as to whether the rows or columns are interpreted as vectors, and indeed due to the duality of the analysis we can flip between the conventions at will by matrix transposition. However, strictly speaking the data should be demeaned, that is subtract off the means of the vectors, and normalized, that is divide by the standard deviations of the vectors, before applying SVD. The demeaning leads to different representations depending on whether it is done for the rows or the columns, and should be done for the vectors that are being analysed.

For example, in IR we represent the corpus as a set of document signatures based on word frequency, and it is convenient to regard the set of features, in this case words, as fixed whilst allowing the possibility of meeting new documents and queries which we want to locate in terms of the signature space. It is common to represent these in a file with one document signature per line, so the corresponding matrix matrix has one document signature per row. It is obvious that this file representation is more easily extended with additional data points in respect of new documents or queries (pseudo-documents). It is thus the rows that are demeaned and normalized, however transformations on the columns may also be carried out and TFIDF (see

below) is a kind of column normalization that compensates for the information capacity of the corresponding word (more frequent words carry less information).

Note that row normalization removes any direct sensitivity to the size of a document (though some indirect sensitivity should remain as longer documents tend to be more encyclopaedic in nature whilst small documents tend of necessity to be on a single specific topic). Note too that vector normalization removes a degree of freedom as the last value is predictable from the others given the constraint induced by normalization (the squares add to 1).

The standard notation for the SVD of A is

$$A = U S V^T$$

where U and V are orthonormal matrices representing rotations and S (sometimes L) is a matrix representing a scaling which has the eigenvalues of A, s_{ii}, represented in descending order along the diagonal, and zeros elsewhere.

The standard LSI notation varies from this including using L (sometimes S) to represent a diagonal matrix with the squareroots of the eigenvalues, called singular values, and using D (for document) and T (for term) as the names of the rotations:

$$A = D L^2 T^T$$

The original SVD formulation derives from the application of the Eigenvalue decomposition known as Principal Component Analysis (PCA) to the covariance matrix AA^T or A^TA (which ever is the smaller, this defining the rank of A):

$$AA^T = U S^2 V^T ; \; A^TA = V S^2 U^T$$

This follows by observing that as U and V are orthonormal their transposes represent the inverse rotations. More efficient calculation methods will be discussed later.

Reduction and Compression

If A is a matrix with u rows and v columns (u x v), then in the unreduced formulation, U has dimensions u x u, V has dimensions v x v, and S has dimensions u x v (viz. it is not square and thus not diagonal in the sense that implies being square). However the rank of A is

$s = \min(u,v)$

and S may be reduced to S_s with dimensions s x s, with a corresponding reduction in U or V by removing rows or columns that are multiplied by columns or rows of zeros in the unreduced S. This compression is completely lossless (apart from rounding errors already inherent in the unreduced version) and thus the subscript is usually dropped so that the same equations used for both the reduced and unreduced versions of SVD, though in reality there is no practical use for the unreduced versions.

Since the eigenvalues of S are represented in descending order, as is the usual convention, a further potentially lossy compression may be achieved by reducing S to a k x k matrix S_k by dropping the rows and columns with greater indices and the smallest eigenvalues, along with the corresponding columns resp. in U and V. If there are any 0 eigenvalues, there is no loss – and at least one 0 eigenvalue will be present (apart from rounding errors) given the vectors have been normalized as described above. If non-zero eigenvalues are removed,

then the compression is lossy and the error introduced may be characterized in terms of their values.

The interpretation of the SVD will be discussed in more detail later, but to understand the errors introduced in compression it should be noted that the principal eigenvector represents the line of best fit in the sense of minimizing the sum of squares error (SSE) and that this is rotated to become an axis and then scaled to become a unit vector. Once this source of variance has been removed, the remaining eigenvectors can be understood recursively as discovering the other tendencies of the data. An important corollary of this is that the error (SSE) introduced by compression is the sum of squares of the eigenvalues removed, and represents the variance that is not accounted for by the model (assuming A has been correctly demeaned as above). If this SSE is divided by the sum of squares of all the eigenvalues, it may be interpreted as an error rate or probability (depending on the application).

There are some papers in the LSI/IR literature that incorrectly refer to SVD as discovering independent features, but in fact it discovers only uncorrelated features and then only if the demeaning has been done as discussed above (and in the LSI literature it is normally not done). SVD minimizes the sum of squares error and is thus second order. This makes sense when dealing with data from a distribution that approaches the normal distribution, but in general independence requires dealing with higher order errors as well, and this class of algorithm is known as Independent Component Analysis (ICA). ICA algorithms usually perform SVD first and then minimize some or all higher order errors (often only fourth order).

Interpretation

We will focus on interpreting in the IR domain so that we contrast with the approach taken in LSI, but before we look at the interpretation of the SVD matrices it is useful to note the following identities and derived matrices. We use a superscript T to indicate the transpose, P to indicate the pseudoinverse, -1 to indicate the inverse, PT to indicate the transpose of the pseudoinverse or equivalently the pseudoinverse of the transpose, and -T to indicate the transpose of the inverse or equivalently the inverse of the transpose, noting that only square matrices of full rank have a true inverse. The original identities (1 to 4) are in unreduced form but the derived identities (5 to 24) are in reduced form, with SSE equal to the sum of squared eigenvalues omitted.

$A = USV^T$

$A^T = VSU^T$

where U is orthonormal, $U^T = U^{-1}$

and V is orthonormal, $V^T = V^{-1}$

and $S = L^2 = U^TAV$ is diagonal

$A^P = VS^{-1}U^T$

$A^{PT} = US^{-1}V^T$

$I = L^{-1}\ U^TAV\ L^{-1} = (UL)^{-1}\ A\ (VL)^{-T}$

$AV = US$

$V^TA^T = SU^T$

$V^TA^P = S^{-1}U^T$

$A^{PT}V = US^{-1}$

$U^{T}A = SV^{T}$

$A^{T}U = VS$

$A^{P}\ U = VS^{-1}$

$U^{T}A^{PT} = S^{-1}V^{T}$

$AVS^{-1} = A^{PT}VS = U$

$S^{-1}V^{T}A^{T} = SV^{T}A^{P} = U^{T}$

$S^{-1}U^{T}A = SU^{T}A^{PT} = V^{T}$

$A^{T}U = SU^{T}A^{PT} = V$

$AV\ L^{P} = UL$

$L^{P}V^{T}A^{T} = LU^{T}$

$L^{P}U^{T}A = LV^{T}$

$A^{T}U\ L^{P} = VL$

Equations 1 and 2 are repeated here as the definition of SVD and its application to the dual problem, transposing rows and columns. The unreduced U and V matrices are orthonormal matrices which means that they are square, that the sum of squares of both rows and columns is 1, that both rows and columns represent an orthogonal basis and are uncorrelated, that the matrix may be interpreted as a rotation around the origin, and that their transpose is their inverse and represents the opposite rotation (3 and 4). S is a diagonal matrix which when it multiplies on the left scales the rows, and on the right scales the columns, by the corresponding eigenvalues. The product, sum and

inverse of square diagonal matrices are simply the diagonal matrices composed of the product, sum and reciprocal of the respective elements. The pseudoinverse S^P of the unreduced form of S formed by replacing the eigenvalues on the diagonal by their reciprocals and taking the transpose becomes the true inverse S^{-1} in the reduced form. (When a matrix X is not square it does not have a unique inverse but will in general have multiple pseudoinverses that satisfy $X^PX = XX^P = I$.)

Thus we have (pseudo)inverses for all three matrices in the decomposition and this allows us to determine pseudoinverses of A & A^T (6 & 7) as well as the identities that are the basis of LSI. In particular, identity 8 shows that UL maps the same latent values to documents that VL maps to rows. Identities 6 to 24 are all derived by the process of multiplying both sides of a previous identity by the (pseudo)inverse of a term that is to be cancelled from one side.

Representations 17 to 20 reveal various additional interpretations of U and V and their transposes. In particular U and V are not only rotations, but are also normalized representations of the vectors in their own right: U is a representation of the rows of A rotated by V and normalized to the unit hypersphere by S^{-1}, and V is a similarly a rotated and normalized representation of the rows of A^T. In their unreduced form U and V are actually quite uninteresting as they simply represent unit vectors along orthogonal axes. In particular, the unreduced U and V matrices are not useful for classification or clustering as each is equidistant from all the others. However in reduced form they are very useful for visualizing two dimensions at a time as the deviation from the unit circle represents the variance in the hidden dimensions – they will be on the circle if they can be described fully in the depicted latent dimensions alone, whilst if there is little relevance to those dimensions

they will be located near the origin. Note that unless A is square and unnormalized, U and V will always have been reduced (to s-1 columns, the rank minus the degree of freedom normalized out).

Representations 9 to 16 are those that are most appropriate for purposes of compression, and where the reduction is lossless the results using L2 and Cosine metrics, and the clusters formed by clustering using these metrics, will be identical to (or in the case of visualizations, rotations of) those obtained with the original matrix, but they should be obtained more cheaply due to the reduction in the matrix sizes.

Representations 21 to 24 are preferred by LSI in the belief that only they will allow words and documents to be depicted or accounted as near each other when they relate to the same topic. In fact, all of representations 8 to 24 have good properties in this respect with representations 9 to 16 giving the original results, 17 to 20 giving results that completely normalize out the strengths of the different latent variables (LSI topics or senses) and 21 to 24 representing a compromise that does neither. This can be seen by observing that 9 to 16 are simple rotations of the original data and all the others are derived by application of a simple scaling by either L^{-1} for the LSI representations 21 to 24 or S^{-1} for the normalized representations.

Identity 8 merely shows that the same scaling L applied to U and V produces a pair of dual transformations and representations UL and VL, each of which can transform A (or A^T) into the other and when both are applied the distortions cancel. On the other hand identity 7 shows that the bilateral application exposes the variances that apply in both contexts.

Noise, Artefact and Error

Data compression and efficiency is not the only reason for reducing the matrix A to a more compact representation. In fact, often the reason is that rather than increasing error the reduction can actually reduce error of various kinds.

The first kind of error is noise. In LSI, noise means the apparently arbitrary choice amongst different synonyms or other alternate expressions, idioms or metaphors. The variance that results from this is often modelled using some variant of the Poisson distribution (and we will discuss normalizations appropriate to these models below). The theory is that the low order eigenvalues correspond to this irrelevant variation.

The second kind of error is artefact. In IR, artefact may refer to some systematic bias in the data, for example the register and content bias introduced by the use of the Wall Street Journal as a source of information about English. In EEG processing, artefact refers to signal that originates from sources other than the brain, in particular ocular and other muscular contamination. Artefact can result in very strong effects that correspond to large eigenvalues, so these reductions must be made by identification of the corresponding components. In the case of EEG this is achieved more effectively by using blind signal separation (BSS) techniques such as ICA.

Other kinds of errors that indeed produce very small artefacts include those that result from the finite precision used in the matrix calculations, and these should always be removed as they really represent zero variance.

Where eigenvectors are removed that correspond to noise, artefact or error, whether the eigenvalues are big or small, the sum of squared

eigenvalues actually represents the net reduction in error, and this is routinely reflected in improved results in subsequent processing. Factor analysis can be understood in terms of subtracting an estimate of the noise distribution from the covariance at an intermediate point in the SVD process and will be discussed below.

Unfortunately, normalization is often neglected or incorrectly performed, which subverts the correct application of the theory, and an unnecessary error may be retained or introduced by dimension reduction. This is particular true in the IR application and the commonly espoused application of SVD in LSI, so we treat this example in detail.

Normalization in our usage includes demeaning as well as dividing by column or row sums or other normalization factors, but should sometimes also include distortion of the original data by the inverse of some non-linear distorting process.

Where different transformations are introduced prior to application of SVD, the interpretation of the noise that is eliminated will change correspondingly. We will discuss several different transformations below from other perspectives, but at this point we will characterize the effect of subtractive filtering for several methods.

Simple demeaning has the effect of removing an additive bias and is optimal in the second order sense of minimizing the sum of squared error given the distribution is sufficiently normal. The effect and advantages of subtracting other measures of central tendency are discussed later, but when applying SVD the arithmetic mean must be zero in order for the squared eigenvalues to represent sum of squares error. We will also consider below the effect of subtracting the

covariance matrix of a noise distribution from the covariance matrix of our original data during Factor Analysis.

In the frequency domain, for example after applying the Fast Fourier Transform (FFT) to a speech signal, subtracting a measure of central tendency or a noise distribution corresponds to eliminating background noise or an uninformative carrier by eliminating its characteristic noise signature. For example, if samples are taken of background air-conditioning and computer-noise, this has a characteristic frequency distribution that can be directly subtracted. For speech, centring on the fundamental and normalizing the variance cancels out basic frequency characteristics due to sex. If in the frequency domain we take the log and similarly subtract a measure of central tendency or a noise distribution, the effect is quite different. In particular for an acoustic signal, this approach can cancel out basic convolutive effects due to multiple paths and echo or reverberation. These are standard steps in signal/speech processing.

These last examples illustrate model-based elimination of noise prior to or during the SVD process, whereas the previously discussed reduction of the decomposition is hypothesized to eliminate certain kinds of noise subsequent to the decomposition and may be model-based (*a priori*), significance-based (eliminating the least significant eigenvectors), accuracy-based (eliminating eigenvectors whose contribution is less than the accuracy of calculation), or empirically-based (*a posteori* elimination of eigenvectors that are associated with identifiable sources of error or artefact).

IDF, TFIDF and Entropy

We now consider the most common preprocessing technique used in IR, TFIDF (Term Frequency * Inverse Document Frequency). There

are a number of common variations of TFIDF that relate to smoothing (e.g. adding one to all values or to zero frequencies) and normalization (e.g. normalizing TF by the frequency of the most common stem in a document rather than the length of the document. We will ignore these considerations, along with stemming, stopping and other processes unique to IR.

The two basic rates that underlie TFIDF are Term Frequency (TF) and Document Frequency (DF) and are expressed relative to a corpus of N documents of average size D, individual size D_d and lexicon size L. If TC_w represents the number of occurrences of a specific term w in the corpus, then $TF_w = TC_w/N$ is its expected relative term frequency and TC_{wd} and $TF_{wd} = TC_{wd} \cdot D/D_d$ represent its count resp. normalized relative term frequency in the specific document d with the expected value of TF_{wd} under the assumption $D_d=D$ being $TF_w = E_D(TF_{wd})$. Similarly if DC_w represents the number of documents w occurs in, $DF_w = DC_w/N$ is its Document Frequency.

Note that TF_w and TF_{wd} are both assumed to be rates relative to a typical document of size D and we have made explicit an assumption that is both ubiquitous and iniquitous, but typically only implicit, in the IR literature – namely that all documents are roughly the same size D. This is clearly not true in most domains, but it can be made truer by dividing documents up into segments or contexts of approximately the same size, choosing D to represent the typical size for topicality of a term. This may be regarded as an assumption that all terms tend to have bursts according to the same probability distribution (again not true), or used to help define what we mean by topicality and thus assist in parameterizing a more accurate model of the distribution.

The Information conveyed by DF_w is $IDF_w = -\log_2(DF_w)$, the misnamed Inverse Document Frequency, being the number of bits required to

efficiently represent the fact that term w occurs in a document. This is a useful measure in its own right (Sparck-Jones, 1972) has a Bayesian justification as well as the information-theoretic motivation given here (Robertson, 2004), and captures the far greater importance of the fact of occurrence in a document versus the precise count or (relative) frequency. Note that much of the work using these measures assumes $D_d \approx D$ is (approximately) constant, and several corpora (e.g. Brown, BNC) use extracts of (approximately) constant size rather than full documents. It is also misleading to think of documents as being about various topics, as different parts of a document may touch on, or even spend considerable time on a topic and then move on to another, thus it can be useful to talk about contexts rather than documents, and these cannot in general be assumed to be of constant size either. We can however select extracts of constant size that relate to a specific topic.

We define $TFIDF_{wd}$ as $TF_{wd}*IDF_w$. As an entropy estimate this represents the contribution of word w to the size of the document when each occurrence is represented in IDF_w bits, but this is not an efficient representation given words actually tend to occur much more than once in the documents they occur in. Thus it is somewhat surprising that TFIDF should for so long have been regarded (and empirically demonstrated) as superior to all other tested weighting schemes (Aizawa, 2003; Robertson, 2004; R&Y text, XXX).

The K-mixture and an improved entropy estimate

Church & Gale (1995), pre-empting Katz (1996), proposed a model dealing with both the probability a document was a relevant context for a word (α) and the rate of occurrence of a term in a relevant context (β).

This formulation of the Katz model, the K-mixture, represents the probability of a term occurring k times in a document as

$P_w(k) = (1-\alpha)\ 0^k + (\alpha/[\beta+1])\ (\beta/[\beta+1])^k$ where 0^k is 1 if k=0 and otherwise 0.

Setting the parameters α and β for a term w may be done empirically, but it is usual to relate them to TF and DF. The rate of occurrence of a term w in a relevant context (β) may be estimated from the rate of occurrence in documents in which it is known to occur ($\gamma = TF_w/DF_w$), but γ would be a significant overestimate for β since these contexts already are known to contain one occurrence. Thus it is usual to set $\beta = \gamma-1$, however this now probably underestimates β as the context is reduced by one word and there is usually a minimum distance before a term recurs. According to this definition, β is set from the rate of additional occurrence in documents the term is known to occur in ($\beta = [TF_w- DF_w]/DF_w$). We may now note that $P_w(0) = DF$ is the probability of a null context and the sum of the independent cases of irrelevant contexts and null relevant contexts, from which we can show that $\alpha = TF_w/\beta$.

Note that Equation 25 uses γ as a denominator discounting both α and β, and the corresponding terms $\eta = (\alpha/[\beta+1])$ and $\delta = (\beta/[\beta+1]) = [TF_w - DF_w]/TF_w$ both have intuitive interpretations with η being the rate of occurrence of relevant contexts with no occurrences of the term (seen by setting k=0) and δ being the probability of an (additional) occurrence in a relevant document (seen by setting k=k+1). The function of γ here can also be understood in terms of a change from average rates in terms of documents to average rates in terms of individual words.

Note that the rate of occurrence of a relevant context (α) may be estimated from the rate of occurrence of documents containing the term (DF_w), although this again would be an underestimate as there are in general relevant documents in which the term could be expected to occur but it does not (perhaps a synonym is used). The correction factor is the ratio of all relevant contexts to non-null contexts which is equivalent to the ratio of all occurrences to extra occurrences, $1/\delta$ – as the rate of extra occurrences is assumed to be the same in a relevant context irrespective of whether it is non-null or null (viz. whether or not w actually occurs in that context).

We now re-express the K-mixture (25) as the more compact P-mixture (26) which is parameterized directly by two probabilities, showing the interrelationships compactly in terms of the derived rates and their interpretations as probabilities and expectations. Note that α is expressed as the sum of the geometric progression formed by the second term of the mixture summed over all k. Similarly β and $\beta+1=\gamma$, which are expectations rather than probabilities, are eschewed in this formulation in favour of (conditional and joint) probabilities δ, ε and η, allowing the P-mixture formula and other identities to be verified directly by multiplication of probabilities and conditional probabilities or expectations.

Identities 27 to 35 relate to the usual parameterization of the K-mixture known as the Katz model. We emphasize the pivotal role in the mixtures of the ratio γ (the expected number of occurrences of w in a D-term context that is known to *contain* w). The ratio β (the expected number of occurrences of w in a D-term context that is known to be *relevant* to w) and the normalized ratio δ (the estimated probability of an occurrence of w in a D-term context that is known to be *relevant* to w) depend solely on γ. The reciprocal of γ, the ratio ε, is directly

interpretable as the probability of non-occurrence in a relevant context and is thus preferred. ζ is introduced (32) as the solution of the quadratic that shows how we can derive γ from TF_w and η, and takes on the pivotal role in an improved parameterization we introduce subsequently. λ and μ (33 & 34) are conventionally used in Poisson models based on term frequency and document frequency, but we discount these by β to derive probabilities α and η resp.

Assumptions. We make explicit the *assumption* of the K-mixture that documents may be approximated as being of fixed size D that represents the expected scope of relevance of a term. D is typically regarded as being of the order 200 to 500 tokens or 1000 to 2000 words. According to Zipf's law the top 150 words accounts for 50% of the tokens in a corpus, and these are largely grammatical function words (closed class words). In addition at least 50% of the so-called content words (open class words) are also very general (generic) and this leaves 20 to 25% of tokens as topical *terms*. Thus we prefer to refer to D-term *contexts*, rather than documents, to emphasize that *relevance* (topicality) is local, that the probability P_D depends on D, and that the large D assumption behind a Poisson distribution does not apply in a D-term context.

We also make explicit the *assumption* that the fact that a document d *contains* a term w implies that w is *relevant* to d (but not vice-versa), express this as $d \ni w \rightarrow d\ Я\ w$, and generally omit $d\ Я\ w$ (*relevant for*) when it is implied by $d \ni w$ (*contains*).

$$P_w(k) \equiv (1-\alpha)0^k + \alpha[1-\delta]\delta^k = (1-\alpha)0^k + \eta\delta^k = (1-\eta/\varepsilon)0^k + \eta(1-\varepsilon)^k$$

$$\alpha \equiv \lambda / \beta = \mu / \delta = \eta / \varepsilon = \gamma\eta = \mu + \eta = P_D(d\ Я\ w)$$

$$\beta \equiv \gamma - 1 = \zeta - ½ = [\lambda-\mu] / \mu = \mu / \eta = \delta/[1-\delta] = E_D(TF_{wd} \mid d\ Я\ w)$$

$$\gamma \equiv \beta + 1 = \zeta + ½ = \lambda / \mu = \alpha / \eta = 1/[1-\delta] = E_D(TF_{wd} \mid d \ni w)$$

$\delta \equiv [\zeta - \frac{1}{2}] / [\zeta + \frac{1}{2}] = \mu / \alpha = \beta / \gamma = \beta\varepsilon = 1 - \varepsilon = [\lambda - \mu] / \lambda = P_D(d \ni w \mid d \text{ Я } w)$

$\varepsilon \equiv 1 / \gamma = 1 - \delta = \mu / \lambda = P_D(d \not\ni w \mid d \text{ Я } w)$

$\zeta \equiv [\lambda/\eta + \frac{1}{4}]^{\frac{1}{2}} = \gamma - \frac{1}{2} = \beta + \frac{1}{2}$

$\eta \equiv \mu / \beta = \alpha / \gamma = \alpha\varepsilon = \alpha[1-\delta] = P_D(d \not\ni w \ \& \ d \text{ Я } w)$

$\lambda \equiv TF_w = \alpha\beta = \gamma\mu = \alpha\gamma\delta = \beta\gamma\eta = E_D(TF_{wd})$

$\mu \equiv DF_w = \alpha\delta = \beta\eta = \alpha\beta/\gamma = \alpha\beta[1-\delta] = \alpha\beta\varepsilon = P_D(d \ni w) = P_D(d \ni w \ \& \ d \text{ Я } w))$

An improved *term entropy* estimate based on the above assumptions and model may thus be defined (36), noting that the model is defined by any two of the mutually independent parameters (viz. at most one of the pairwise dependent β, γ, δ, ε and ζ).

$TE_{wd} = -\log(P_w(TF_{wd})) = -\log(\delta)*TF_{wd} - \log(\eta) = -\log(DF_w/\alpha)*TF_{wd} - \log(DF_w/\beta)$

Rearranging, TE_{wd} may be viewed as a combination of the simple TFIDF and IDF models, or a smoother 'add one' TFIDF model, with the subtraction of redundant information as revealed by a simple ITF model, since IDF ignores extra occurrences of w and ITF models the overestimated entropy of the expected β extra occurrences of w amongst the $\lambda=TF_w$ occurrences treated as introducing new relevant contexts (37 & 38).

$TE_{wd} = TF_{wd}*IDF_w + IDF_w - ITF_w = (TF_{wd}+1)*IDF_w - ITF_w$

$ITF_w = -\log(\beta) - TF_w*\log(\alpha) = -\log(\nu)$ where $\nu = \beta\alpha^{\lambda}$

There is however some fine tuning that will performed later when we present our parameterization models – in particular TE_{wd} is not well

defined in the K-mixture when $TF_w = DF_w$ (e.g. at most one occurrence of word w in any document).

Note that there are many different notations used for the probabilities and expected values discussed here, that statisticians tend to use the term frequency to indicate a count irrespective of the total sample or corpus size, where as in natural language processing we are more interested in the stricter definition of frequency as a *rate* of occurrence relative to some specific sample size or time period, and in particular in the above discussion we use frequency relative to some known number of words or a specific average document size – in statistics, this usage is referred to as relative frequency. Counts are not interpretable without knowledge of the sample size or time period. The average interval between occurrences, which corresponds to wavelength in the time domain, is also a useful scale that will be discussed later.

Furthermore, the presentation by Church & Gale (1995) of Katz's K-mixture pre-empts the publication in the same journal of Katz's (1996) account of his model, and the specific notation and formulation of the models are quite different in the two papers. The K-mixture is presented as an additive model in which two components contribute to the probability estimate for documents that do not contain a term. However, the corresponding model of Katz (1996) is a two parameter variant, G', of a more general three parameter model, G, that has only one term applicable to any specific term frequency. In fact, G' and $G=G''$ can be considered as first order and second order variants (resp.) of a more general model extending the straightforward zeroth order model, G^o, that depends only on the mean document frequency $\mu=DF_w$.

The following presents this model in a form that use the constructive notation proposed in Katz (1996) but deliberately eschews the adoption of the Greek letters used by him in his preferred presentation form, as

they differ from those used above based on the usage of Church & Gale (1995). The first concept introduced by Katz is Burstiness, $B_m = E_w(k \mid k \geq m)$, with B_0 corresponding to the *distribution mean*, B_1 to *average burstiness* (before topicality is established), and B_2 to *topical burstiness* (in model G, two occurrences are necessary and sufficient to establish topicality):

$$B_m \equiv \sum_{r \geq m} P_w(r)\, r \,/\, \sum_{r \geq m} P_w(r)$$

Katz defines $P_m = E_w(k+1|k, k \geq m)$ as the *average conditional probability* of additional occurrences after seeing at least m, being the ratio of repeats to opportunities:

$$P_m \equiv \sum_{r \geq m} P_w(r)\,(r - m) \,/\, \sum_{r \geq m} P_w(r)\,(r - (m-1)) = 1 - 1 \,/\, [B_m - (m-1)]$$

A model of order m is defined as having a conditional probability of repeat occurrence that is independent of the number k of previously observed occurrences once the burst size is at least m. In this model $P_w(k+1|k, k \geq m) = C_m$ is assumed to be a constant independent of k for all $k \geq m$ in a model of order m, and thus the average of all such *constant conditional probabilities* $C_m = P_m = P_k$ for all $k \geq m$:

$$C_m \equiv \sum_{r > k \geq m} P_w(r) \,/\, \sum_{r \geq k \geq m} P_w(r) = P_m$$ *(cf. Kat*

The model order m thus corresponds to the assumption $P_m \approx P_k$ for all $k \geq m$ and allows us to define $P_w(k)$ for these models as

G^o : $(1-P_0)P_0^{\,k}$

G' : $(1-P_0)0^k + P_0(1-P_1)P_1^{\,k-1}[1-0^k]$

G'' : $(1-P_0)0^k + P_0(1-P_1)0^{k-1} + P_0P_1(1-P_2)P_2^{\,k-2}[1-0^k-0^{k-1}]$

$G^{\{m\}}$: $\sum_{j<m} 0^{k-j}\,(1-P_j)\prod_{i<j} P_i \;+ [1-\sum_{j<m} 0^{k-j}]\,(1-P_m$

An generalized *term entropy* for $k = TF_{wd}$ based on the this generalized Katzian $G^{\{m\}}$ model may thus be defined as

$$-\sum_{j<m} 0^{k-j} \log(1-P_j)+\sum_{i<j} \log P_i \; - [1-\sum_{j<m} 0^{k-j}][\log(1-P_m)+(k-m)\log P_m+\sum_{i<m} \log P_i]$$

These order m models essentially treat the j<m cases individually, and define the j≥m cases as if $\delta = \delta_m = P_m \approx C_m$ was defined by an order 1 model (G′ or K-) employing a membership threshold of m. Relating these parameters to the Greek parameters used earlier, we get

$$P_0 \equiv \mu = DF_w;\; P_1 \equiv \delta_1 \equiv \delta;\; P_k \equiv \delta_k$$

where these P_j values may be empirically calculated using (39 or 40), or estimated for k≤m or k=m using (41) for a smoother model and/or a more accurate tail. Simplifying and taking advantage of the exponential Kronecker forms, we get probabilities for k

$$\sum_{j\le m} \delta_{i<j\le m} (1-\delta_j)\, \delta_{k,j\le m} \qquad \textit{where} \quad \delta_{i<j\le m} \equiv \prod_{i<j} \delta_i \;;\; \delta_{k,j<m} \equiv 0^{k-j} \;;\; \delta_{k,j=m} \equiv \delta_m^{k-j}$$

Note that these forms are less convenient than the K-mixture and P-mixture due to the Kronecker terms $\delta_{k,j<m} = 0^{k-j}$ that appear for k>0 in this formulation, but are avoided in the previously given mixtures (we are not concerned with the k=0 case documents that omit the term). Nonetheless, we find this new notation both elegant and intuitive as we see that the higher order model is a smoothing model using a mixture of elementary models of different orders where δ terms for component models of lower order j<m reflect a probability of contribution 0 to higher or lower orders, whilst the δ term matching the model order j=m reflects a probability of P_m for each successive term above the model order. This convenience is also observed in the P-mixture (26).

Katz (1996) has verified that the G-model performs considerably better than several lower order models including G′ and Church and Gale (1995) have similarly demonstrated the competitiveness of G′ in the form of their K-mixture, albeit validating on the training data in both cases rather than on independent data. A number of other researchers have used the technique or the concept of burstiness (Umemura, ZZZZ) in various applications and have found it effective, although it tends to be the K-mixture form that is used even when only Katz (1996) is cited (e.g. Gao et al., XXXX, YYYY).

The scope of demeaning and normalization

There is however a prior question relating to demeaning that we will pick up here – how should normalize your data. The first aspect is whether you demean, by row or by column, over the entire matrix or by both rows and columns successively? These are four distinct possibilities (really five but two are mathematically equivalent, but will have different errors introduced numerically). The second aspect concerns whether you should scale the data, and a similar set of choices are available.

Demeaning is recommended, it is technically incorrect to apply SVD without it and it reduces the compression error. While technically demeaning should be done for the vectors in focus, in practise we would like to consider both interpretations of the data and a similar error reduction is achieved either way – we will call this orthogonal demeaning and it may in certain cases be more convenient. Double demeaning (vector and orthogonal successively) reduces the error further and while theoretically less justifiable it may be useful if it is desirable to reduce the error due to compression, and/or if the dual interpretation is also useful. Matrix demeaning involves subtracting the

mean of the entire matrix from all entries and does not achieve any useful effect in general, achieving a fairly meaningless translation in both primary and dual spaces.

The correct normalization for AA^T to be a covariance matrix is demeaning by row, since we are treating A as a set of row vectors. More importantly this is appropriate if we are analysing, clustering or visualizing row vectors, e.g. each row describes a document in terms of the words contained in it, in typical IR usage, and we are interested in the relationships between documents rather than between words. For A^TA to be a covariance matrix we should demean by column, and this is appropriate if we are dealing with the column vectors, which corresponds to us being more interested in the relationship between words than documents.

The minimal normalization for AA^T or A^TA to be a covariance matrix is thus demeaning, but this reduces the degrees of freedom for the vectors by one (this reduces the rank by one if A is square or we demeaned along the shorter dimension). Following up with standardization of A by dividing by the corresponding (row or column) standard deviation means AA^T or A^TA is a correlation matrix, whilst the more usual normalization to unit vectors (sum of squares 1) maps vectors to the unit hypersphere.

Undemeaned L1 normalization (normalizing the sum of magnitudes to be 1) by row or column without demeaning gives rise to conditional probabilities in the case of frequency data. However L1 normalization with or without demeaning is arguably inappropriate for SVD processing which is based on minimizing sum of squares error. Furthermore, L1 normalization after demeaning no longer admits a probabilistic interpretation.

L1 normalization of an entire undemeaned frequency matrix (so the sum of all elements is 1) gives rise to probabilities, but normalization of the entire matrix (L1 or L2) is usually of little value as biases such as document size are preserved.

This leaves the double normalization, where we normalize the rows then the columns. Demeaning by row means all rows now sum to zero and have zero mean. Thus following up by demeaning the columns means we are subtracting a row of columns means with zero row mean from each row, so the row mean remains zero. Thus this procedure, or its column then row converse, lead to a matrix that has zero mean for all rows and columns, and formally the two approaches are equivalent.

Geometrically if the data is considered as a NxM matrix of N M-dimensional row vectors, demeaning by rows translates so the centroid is at the origin, and distances of the vectors from the origin (length) reflect variance. Demeaning by columns projects onto hyperplane $\Sigma_i x_j = 0$, and the vectors reflect dissimilarity of the attribute dimensions. L2-normalization of N vectors by scaling to unit variance or unit hypersphere means the sum of squares adds to N (naïve sample variance), N-1 (unbiased sample variance) or 1, but in any case that the vectors lie on the surface of a hypersphere. Each of these three steps, individually or together, costs a degree of freedom in the data, although L2-normalization is not linearly resolvable so does not affect rank (except to the extent it approximates L1-normalization).

Moverover, note that L1-normalization of non-negative attributes (e.g. frequency data to probabilities) projects them onto the hyperplane $\Sigma_i x_i = 1$ so that subsequent demeaning by columns does not cost an additional degree of freedom but represents a simple translation to $\Sigma_i x_j = 0$. On the other hand, L1-normalization after orthogonal demeaning does still lose the third degree of freedom. Since L1-normalization sets

the row sum to 1, the row sum of the mean vector (centroid) is also 1 and standard vector demeaning following L1-normalization obviates any need for orthogonal demeaning since the row mean sum of 1 is subtracted off each row sum of 1 giving a row mean of 0.

It is not possible in general to simply scale both columns and rows simultaneously to achieve L1 or L2 normalization, though in fact that effect is achieved by the SVD rotation and scaling process since the original square versions of both U and V are L2 normalized by both row and column. A normalization by scaling row then column versus column then row is generally quite different and the last normalization therefore should correspond to the vector normalization required for subsequent processing. However, the empirical investigation of Powers (1997) found the double normalization and the matrix normalization gave no advantage or made things worse: on average it reduced the performance achievable in subsequent clustering. On the other hand, the study showed that L1 vector normalization helped more than any other normalization (in this case frequency data was used and it factored out the size of the document), but L2 vector normalization was second best (it also helped with size).

Our conclusion is that the double demeaning is potentially useful as in general it reduces the sum of squares variance represented by the S matrix of the SVD over what is achieved by either column or row demeaning alone, it is theoretically insensitive to the order in which it is performed, and it means that the SVD analysis may be interpreted validly for both columns and rows.

Vector normalization by linear scaling may be helpful if size is not regarded as a desirable factor, and indeed the principal eigenvector will generally correspond to size (and thus not be useful) if this normalization is not performed. However, after normalization one of

the largest eigenvalues may still correspond to a size-correlated eigenvector (in IR this is largely due to the different lexicon sizes of large and small documents – compare a dictionary or an encyclopaedia to a personal home page or a newsgroup submission).

Geometrically L2 normalization after demeaning places vectors on the unit hypersphere (circle in 2D; sphere in 3D) around the origin, whilst L1 normalization places vectors on a unit hyperhedron (diamond in 2D; octahedron in 3D). L2 normalization is more appropriate for a subsequent SVD processing step with its second order nature and is thus a reasonable postdemeaning normalization step, whilst L1 is the best way to normalize for size and is a highly appropriate predemeaning step but is in the IR context is unnecessary if documents are constrained to be the same size. However, for IR with unconstrained document size, the following steps are recommended:

1. L1-normalization of vectors (map to conditional probabilities, reducing freedoms)

2a Optional non-linear transformation of vectors (e.g. map to conditional information)

2b Optional orthogonal demeaning (translate by $\Delta x_i = -1/M$ from $\Sigma_i x_i = 1$ to $\Sigma_i x_i = 0$)

3. Vector demeaning (mapping centroid vector to the origin, reducing freedoms).

Note that as discussed above demeaning step 2b is not required after L1-normalization unless a non-linear transformation at step 2a undoes the L1-normalization, as vector demeaning will also achieve orthogonal demeaning. Double demeaning (explicit or implicit) will reduce the error of SVD-reduction. In order for this sequence not to hide error we implicitly assume that the size information lost in step 1 is not directly

informative (size of document does not provide information about what the document is about) and similarly that step 2 is information preserving. This is automatically so for transformations that are reversible, as is the case for 2b. This is also the case for 2a if monotonic functions are used. Given data is strictly positive log$x$, tan$x$, tanh$x$ or x^k are all possibilities, but the >0 condition need not obtain and indeed in IR matrices are usually sparse in violation of this condition. This would thus appear to exclude direct application of log, which is unfortunate as translation to the information domain would appear to be semantically advantageous (as we feel proportional increases in frequency are equivalent) whilst preserving sparsity is computationally advantageous (in terms of both time and memory).

Postnormalization (after step 3) is not recommended as it will destroy the double demeaning. This includes standardization (so that A^TA is a correlation matrix, as commonly performed in ICA algorithms) and any other form of L2 normalization in particular (and in general any transformation that is non-linear for either rows or columns) means that the loss of degrees of freedom is no longer reflected in rank.

Recall that the SVD process produces two orthonormal matrices, before reduction. Even after rank reduction the column vectors are effectively L2-normalized and orthogonalized, but the row vectors are shorter in proportion to the squareroot of the reduction in rank – with rank reduction from N to K, length is $\sqrt{(K/N)}$. With lossy compression the shortening is given by the foreshortening due to the magnitudes of the components in the ignored dimensions, viz. $1 - \sqrt{(\Sigma_i x_i)}$.

Renormalization is thus not desirable for purposes of 2D visualization or other lossy reduction paradigms, as the length of the vectors (distance from centre) is indicative of the information that is present in unshown/unretained dimensions. For purposes of comparison (distance

or similarity measures) in the absence of lossy reduction, renormalization is unnecessary as it remains L2-normalized.

In this context, dot product corresponds to D2=cos(θ) where θ is the angle between vectors, varying from +1 for positively correlated through 0 for uncorrelated to –1 for negatively correlated (as θ goes from 0 to π/2 to π), and Euclidean distance corresponds to L2=2sin(θ/2), varying from 0 for identity to 2 for diametrically opposite (as θ/2 goes from 0 to π/2). Note that θ is itself a linear distance measure whilst dot product and Euclidean distance represent non-linear (inverted sigmoid and half-sigmoid) distortions of θ. Furthermore, $L2 = (2–2D2)^{1/2}$ and $D2 = 1 – L2^2/2$. The Manhattan distance, L1, the Canberra distance and other variants of the Minkowski metrics, Lp p≠2, are not rotation invariant and hence particularly inappropriate in conjunction with SVD. ICA as an alternative to SVD represents a further rotation and thus is pointless if rotation invariant distance measures are used, unless it is used to identify and eliminate irrelevant dimensions (as a form of lossy compression designed to lose noise rather than information).

The sparsity preserving log transformation (splog)

We noted above in several places that demeaning should be performed before SVD, but often is not, particularly in the context of LSI. The mean represents the best fit in a sum of squares sense and so the sum of squares is not minimized if the mean is not zero, and the squared eigenvalues are not interpretable as variances. Thus in order for the error resulting from lossy reduction to be minimized it is necessary to demean, and the error introduced by the dimension reduction will tend to be doubled for LSI, since it will be biased by the mean, and the IR

frequencies may be modelled in terms of a Poisson distributions for which the mean and variance are equal.

So this raises the question of why demeaning is not performed. In some cases, it is simply ignorance of the importance of this step in what has become a black box technique. However, in the case of LSI we are dealing with term frequencies that are zero for most terms given a document, and for most documents given a term. In the case of terms, there will typically be at least a 1000 times more zeros than occurrences for any document. In the case of documents, the factor is similarly high for all terms once the frequent closed class lexical items have been excluded as stop words. The overall sparsity reflects the average sparsity of both document and term vectors, and is lost if we demean either document or term vectors as the nulls now all become – mean, and they cannot meaningfully be left at zero .

However, if we take the log (or –log) of the non-zero values and move to the information domain, the arithmetic mean in this information domain represents the geometric mean in the original frequency domain, and this is also a valid measure of central tendency that we will discuss in more detail below. We will assume we are treating document signatures, as in IR/LSI, rather than the word signatures, as in other applications such as word sense disambiguation (WSD). Thus we will subtract the mean term information in relation to the non-zero terms of each document, but in this case we will argue that we can reasonably leave the zero entries at zero.

Demeaning in the information domain gives a balanced spread around an expected information of zero (frequency equals expected frequency and the log of their ratio is 0) so that the null documents (those in which the frequency was zero) can remain at zero (0 frequency equals 0 expected frequency) for no information gain or loss. Positive figures

correspond to the document's information gain for the word: higher occurrences mean less information is gained by being told a word in the document is the target, reflecting the fact that the document has more information about this word. Negative figures correspond to information loss which tends to be spread over many words as, except for the excluded closed class words and a few other instances of similar generality, words other than the topical words will occur less often than usual in the document simply because they are not the topic and are being displaced by topic-relevant words.

Note that the expected count for a term in an irrelevant document is close to zero and is much less than the expected count in a relevant document (and indeed we have assumed that the occurrence of a word defines a document as relevant). For terms that do not occur in a document, that we ignored in taking the log and the mean above, we note that since we know there a 0 occurrences of the term, 0 is arguably the expected number of occurrences in the document, and so there is no deviation from this expectation and hence no information gain or loss. The demeaning model based on this argument we will call splog0. This is simplistic and we will shortly analyze the error introduced by these assumptions about demeaning and the fact that we are essentially demeaning two subsets of the documents separately.

Scales and distortions

We noted earlier that sometimes it is appropriate to normalize using some non-linear processing step. In fact, the difference between SVD and higher order ICA can largely be understood in terms of ICA matching a sigmoid to the cumulative distribution function of the data (the probability of a value being less than or equal to k, being the sum

up to k of the probabilities for discrete possibilities i, or the integral up to k of a continuous probability distribution function p(i)).

We have now seen some applications of the non-linear logarithmic distortion, which has raised the question of whether this is a good thing from various perspectives.

In general statistical analyses are based on specific assumptions and these include assumptions about the scale of measurement. For example, it only makes sense to add values that are measured on an equal interval scale. In the case of frequency, our intuition is that adding one extra occurrence of a term to a document with one occurrence is far more significant than adding it to a document with a thousand occurrences. Thus we think that it is the percentage increase that is important. Taking the logarithm means that variations by the same percentage are shifted by equal intervals, and this is part of the value of the concept of information.

Another useful assumption that can be made about scales is that there is an absolute zero and it makes no sense to have values less than zero – only in such a scale does it make sense to talk about ratios, something being twice something else, for example. For this reason, absolute scales are also known as ratio scales. The ratio scale property holds for measurements of size or interval, but is not true for measurement of location. In the IR application, zero frequency of a term and zero interval between occurrences of a term both thus delimit absolute measurement scales. Note that frequency and wavelength have a reciprocal relation, and so do frequency and interval. Both average interval and probability can be seen as being derived from frequency, but equally well frequency could be regarded as being derived from an underlying probability distribution or an underlying repetition interval (buses are due every 15 minutes => 4 per hour => a 1/15 probability of

a bus arriving during any particular minute). The Poisson distribution actually derives from this idea of the expected interval to the next occurrence, noting that it cannot be smaller than zero as that would mean the latest bus arrived before the last bus, which is a contradiction.

Zipf's Law tells us that if we order the events (e.g. occurrence of terms) by their probability and number the most frequent term (1) to the least frequent term (L), this *rank* is inversely proportional to the frequency, or directly proportional to the interval. Taking the rank is a standard statistical technique used when the data does not obey the assumptions of the paradigm used and so parametric techniques are not appropriate. After taking the rank we end up with an absolute equal interval scale and we can now use techniques that make such assumptions, the composite approach of taking the rank and using a parametric method being a non-parametric method that does not require these assumptions about the underlying distribution.

Frequency and probability do not obey the appropriate assumptions for the usual statistical methods, but taking rank, reciprocal or logarithm achieves an absolute equal interval scale, otherwise known as a ratio scale. Empirically, in many language-based applications better results tend to be achieved with one of these scales than with the raw frequency or probability scales (see e.g. [Powe97]).

Note that the normalization of frequencies to probabilities or conditional properties is identical with L1–normalization on a matrix or vector (divide by the sum of magnitudes, before or after demeaning) but is undone by the L2-normalization appropriate to and performed by SVD (divide by the sum of squared magnitudes, recalling that the U and V matrices are L2-normalized). These are linear normalizations which we distinguish from the non-linear distortions represented by reciprocal and logarithm. Visually a good representation, with an

intuitively plausible equal interval scale, has a very even spread of points. A poor representation will tend to be more or less dense around the mean and will lead to the development of inappropriate clusters.

Taking the logarithm fits our intuition about ratios being important at the level of frequency as well as our intuition that information should be represented in bits (or related units such as nits or digits where natural or decimal logarithm are used rather than the usual binary logarithm).

Taking the reciprocal or the rank fits the Zipfian model and evenly spreads out the words according to their characteristic frequency (either in the corpus or in a document).

The choice of distortion will in general affect the results. For example in [Powe97], the rank distortion tended to be the most robust, finding the expected vowel class, with y as a possible inclusion, whilst the logarithmic distortion tended to class space as a possible vowel (with or without identifying the vowel class in its own right). On the other hand, the reciprocal distortion (despite its Zipfian similarity to rank) was not very successful at finding the vowel class.

In the IR application, one of the prime goals is correct ranking of documents relevant to a query, and pre-ranking of the terms has also proven to be a useful step in preliminary explorations of its utility in clustering. The TFIDF processing step discussed above, and the subsequent Katzian modification, both make use of a logarithmic distortion into the information domain. These *dis*tortions should really be understood in the etymologically accurate sense of *un*twisting the data in a way that removes *un*helpful biases.

The *un*expected value

We now return to consider in more detail our observation that the proposed sparsity preserving method of demeaning in log space (splog) effectively treats the original geometric mean as the measure of central tendency. We also note that the existence of the various alternatives is evidence that other measures are at times more useful, and in particular we note that the geometric mean is appropriate where a logarithmic distortion gives the required scale, and similarly the harmonic means is appropriate where a reciprocal distortion gives the required scale, and the respective distortions map these means to the arithmetic means required in the distorted space.

In clustering, to give a specific example, we note the existence of variant algorithms, e.g. K-medians versus K-means, K-harmonic-means, K-geometric-means and the various density-weighted K-means algorithms, all of which can be interpreted as targeting the underlying probability distributions/densities and represent the cancelling of a bias by explicit (e.g. harmonic/geometric variants) or implicit (weighted variants) application of a distortion function. However, if we applied an appropriate distortion up front, a simple K-means would give equivalent performance to the harmonic and geometric variants. The case for the weighted K-means algorithms is a little different as some are designed to still find the underlying clusters based on arithmetic means but are concerned about the distribution from the perspective of sampling (e.g. increasing weight of low cardinality clusters in conformance to Zipf's law).

Thus seeing a problem with the change in the referent of an expected value as a problem makes the assumption that frequency or probability is the correct thing to measure, whilst in fact rank, reciprocal and log

are arguably more appropriate on various statistical and intuitional grounds. Furthermore, taking the mean of the logs centres around the arithmetic mean as technically required for our various matrices to be interpreted as covariance matrices, and in fact the reason the arithmetic mean or *expected value* is preferred in statistics is largely the second order error-minimization property that applies after normalization.

A particular consequence of the methodology proposed is that the null documents are shifted relative to the non-null documents as sparsity is retained by keeping them at zero. An alternative approach would be to subtract the log of the mean rather than the mean of the log, but this no longer represent the second order optimum in terms of minimizing SSE. These two models correspond to normalizing respectively the arithmetic mean or the geometric mean to one in the original frequency space. We now analyze and compare the two models and argue that the standard model where we apply the distortion and demean is indeed preferable, but that there are also modifications that may be useful.

The discrepancy is the same either way, but affects different subsets of the data and in this case corresponds to Information Bias whilst more generally it could be called Functional Discrepancy on the basis that after applying the function the expected value (statistically) is not the expected value (semantically).

So let us consider the properties of various measures of central tendency.

The arithmetic mean is only the expected value in the semantic sense under specific conditions, e.g.

1. An aggregate of random variables and the central limit theorem applies;

2. A word's frequency or probability or occurrence intervals for an ergodic source.

In both these cases the term 'expected value' means something and it is actually the value you expect by dint of the stated assumption (e.g. the larger the interval the more likely the frequency will correspond). However, the ergodicity requirement says that any large enough sample should be representative of the whole, whilst we have actually shown a better model is that there are relevant and irrelevant contexts that demonstrate quite different distributions, and the log-demeaning process deals with these separately. Although in a document or sample of size D words, both p = f/D, the mean frequency per token, and I = D/f, the mean interval between occurrences are arithmetic means, arithmetic averages of these over different words does not give compatible results = the arithmetic mean interval for a set of words is actually the reciprocal of the harmonic mean frequency for that set of words – the arithmetic mean is not appropriate for combining frequencies or probabilities of different words if it is actually the interval that is important, and it would seem that psychologically and linguistically our cognitive limitations determine that we should get reminded periodically what the topic is by reusing the word explicitly rather than anaphorically, and that refreshing will be required after a certain interval has passed.

There are many cases where the expected value is not semantically interpretable as such, e.g.

3. The random variable represents labels rather than measurements (e.g. a dice)

4. A variable with a semantic expected value is nonlinearly distorted (e.g. log, recip.)

There are also other measures of central tendency that are used in statistics, and these come into play here as we consider the significance of the functional discrepancy between E(f(X)) and f(E(X)). We have also discovered another measure of central tendency, the entropy norm κ that is implied by the term entropy we derived from the K-mixture and P-mixture and justified in terms of balancing information about the document as measured by the fact of occurrence and the number of occurrences.

If we have a normal distribution, then the median, mode and arithmetic mean coincide. As the weight of the distribution is close around these values there is little opportunity for functional discrepancy – and this is increasingly the case for larger N and tighter distributions.

Returning to the specific functional discrepancy in moving from probability space to information space using the log function, and considering the relationship between the arithmetic and geometric means of the underlying frequency data, we note that the geometric mean is more sensitive to low outliers (esp. 0, which we however discount by taking means only over documents containing the target word) and less sensitive to high outliers (which is often regarded as a good thing as these are unbounded).

Considering the binary case for non-negative values (frequencies) a and b,

am(a,b) = half(a+b) >= sqrt(a*b) = gm(a,b),

with equality holding only when a=b.

The functional discrepancy thus relates to the variance, and noting that a and b both differ from the arithmetic mean by dev(a,b)=abs(a-b)/2 we see that the variance represents the deviation between the squares of the arithmetic and geometric means.

var(a,b) = dev(sqr(am(a,b)),sqr(gm(a,b))

This relationship of functional discrepancy, fd, to variance immediately leads to some understanding of the extent of the fd for different possible distributions. For a Poisson distribution applicable to our hits (the documents containing the target term) we have

am(a,b) = var(a,b)

The P-mixture essentially defines two probabilities in our model of word occurrence in documents (and implicitly assumes documents of equal size D). These can be characterized as the probability of a relevant context (α) and the probability of an additional occurrence in such a relevant context (δ) – it allows for zero occurrences in a relevant context and models relevance separately.

As we are processing only non-null contexts, and for now treat relevant contexts as being the hits for a word, we simply use a Poisson distribution to model occurrences. The Poisson model predicts the probability of a document of size N (N large and hence omitted) containing n occurrences of w when the expected number of occurrences is m as

$P_m(n) = m^n / e^m / n!$

Since we know the document contains an instance of w, we used formula 30 to predict *additional* occurrences, and investigated the relationship of the median, mode and arithmetic and geometric means computationally using this model, showing that the Poisson model with an overall the number of occurrences n is characterized by

median: n
mode: n

geometric mean: n+0.5
arithmetic mean: n+1.0

Our original observation was that a sparsity-preserving demeaning of the logarithm reflected the log of the geometric mean of the data, and that the arithmetic mean in the log-normalized space allowed proper interpretation in correlation and SVD processing. We showed that seeing zero probability as a limit, nulls correspond to the *geometric* mean of the original data for both non-null and null cases. In this model the transformation actually potentially increases sparsity as the nulls stay at zero and any documents matching corpus level expectations become additional zeros. This corresponds to the standard approach to distortion and demeaning [Powe97].

The alternative approach of subtracting the log of the arithmetic mean puts the zeros at the values corresponding to the arithmetic mean of the original data and ensures that nulls correspond to the *arithmetic* mean of the original data for both non-null and null cases. This is motivated by the traditional view that the arithmetic mean represents the expected value of the data, but then the log-distorted version is not properly demeaned and thus the correlations and SVD reductions have parametric error introduced into them.

Moreover, since mode represents the salient value in the distribution and the most probable prototype, it is possible to argue that this is a more plausible measure of central tendency and hence the closer geometric mean (expected displacement of 0.5) is more appropriate than the arithmetic mean (expected displacement of 1.0). Also note that the document groups around the mode and the median are unchanged by the application of any monotonic distortion, but that the arithmetic mean now identifies documents that were formerly identified with the geometric mean rather than the arithmetic mean, so both of

these are less resilient than mode and median. Furthermore, if the logarithmic (information or information gain) space is meaningful, and we have argued that intuitively it is an equal interval scale, then its mean is reasonable to use too. Also note that information is an absolute scale with an absolute zero (achieved at probability 1). Moreover the magnitude of the information deviation is also an absolute scale with zero indicating a totally unsurprising outcome, a non-zero magnitude specifying the minimum description length for the deviation from expectation, with the sign being an additional bit that indicates a direction in the sense of gain or loss. Thus information deviation is also an absolute ratio scale notwithstanding the availability of the sign.

Katzian analysis of sparsity preserving log model

The first observation regarding the simple splog model is that it does not distinguish between the case where a word only ever occurs a single time in a relevant document (or more generally occurs the expected number of times in a document), the case where as expected it does not occur in an irrelevant document, and the case where it does not occur in a relevant document. In the first two cases the value is 0 because it matches the empirical expectation, and the third is not distinguishable by our simplistic model. When TF=DF, there is a more serious problem as the variance is zero so $p(TF \neq DF)=0$, so that in the Katzian models, $\beta=0$, $\gamma=1$, $\delta=0$, $\varepsilon=1$ in violation of our assumptions, and α and η are consequently not well defined. There are several separate issues here arising from specific assumptions and that all cases where a probability is 0 or 1 are degenerate. In particular we want to ensure:

1. There is a non-zero probability of occurrence in an irrelevant context (α)
2. There is a non-zero probability of an extra occurrence in a relevant context (δ)
3. There is a non-zero probability of null occurrences in a relevant context (ε)

The implication of point 1, is that zero occurrences is actually an underestimate of the expected number of occurrences in an apparently irrelevant document (any term may occur incidentally or as a change of topic), so that the document information gain should actually have a negative value (a loss), the count being less than expected. To preserve sparsity, we could compensate for this by adding $\alpha>\mu$ to all non-null terms before splogging as the probability of occurrence in an irrelevant context, or μ=DF as the probability of occurrence in any context. The latter is the better choice as we specifically recognize than non-occurrence does not imply irrelevance, where as the former corresponds to the usual IR search models where non-occurrence is equated with irrelevance.

Thus one solution is to view frequency 0 as a deviation from μ, being our estimate of the expected number of occurrences in an *arbitrary* context. Since we want to maintain sparsity and keep the expected values at 0 (and these zeros dominate the matrix and heavily influence the mean), we propose to maintain relativity by adjusting the values for non-null occurrence by subtracting μ (modification μ). Given some of our null-occurrence documents may be relevant, it would also be possible to subtract α to demean that subset of the documents (modification α). However in the absence of some independent way of

determining relevance we cannot distinguish the relevant documents and they overlap both the null and non-null occurrence documents.

The implication of point 2 is that we need to implement some kind of improved estimate of δ in at least the cases where $\lambda=TF=DF=\mu$. This could be implemented as some kind of smoothing, and in particular we have an implicit assumption that the probability of occurrence in an irrelevant context is cannot exceed the probability of occurrence in a relevant context, and both are strictly non-zero ($0<\mu<\alpha<\delta<1$). Note that if we have eliminated stopwords and are limiting our terms to topically relevant terms, the central inequalities should be strict as relevant topics should be more frequent than documents containing terms as illustrated by synonyms, paraphrases and definitions, and the Katzian cluster model is based on the idea that words are more frequent once the initial occurrence has occured. An 'add one' or 'add half' style correction would note that $E(TF)>E(DF)$ according to the model and that it is just that in a limited size corpus we are seeing specific TF and DF values from a distribution that happen not to satisfy TF>DF, and propose correcting this with an estimate $TF^*=DF+\varphi$, for $\varphi = 1.0$ or 0.5, or $\varphi/\mu = 1.0$ or 0.5.

However, this can actually lead to violations of probabilistic conditions (probabilities exceeding 1), so that a more intelligent approach is needed. We therefore introduce an increment φ that is a function of μ and a factor f that specifies the proportion (probability) of non-occurrence contexts that are in fact relevant. Our constraints give us a smoothing model of the form 'add φ to μ' (adding to empirical corpus term frequencies) or 'add φ/μ to γ' (adding to expected document term frequencies or counts, $TF_{wd} \approx TC_{wd}$) defining the reasonable values of φ as follows:

$$1 > \lambda = TF^* = \mu + \varphi > DF = \mu$$

$1 > \alpha = \mu\cdot(\mu + \varphi) / \varphi = \mu\cdot(\mu/\varphi + 1)$

Solving the quadratic inequality (55) for positive μ and φ and parameterizing with f, we obtain

$\varphi > \mu^2/(1-\mu)$

$\mu < [\varphi^2 + 4\varphi]^{-1/2} / 2$

$\varphi = \mu^2/[f\cdot(1-\mu)]$, $\alpha = f\cdot(1-\mu) + \mu$, $0 < f < 1$

Hence a minimal variant of 'add φ' smoothing would be to set documents with such expected levels of occurrence and one actual occurrence to $1+\varphi/\mu$. In this model, setting $f = 1$ violates the model and even $f = 0.5$ is not conservative enough as it states that half the null-contexts are actually relevant, however a good search term has $\mu<<0.5$ as the appropriate estimate of the probability of occurrence in a context whose relevance is not known, and $f = \mu$ leads to the definition of φ as the odds ratio ρ

$\varphi = \rho = \mu/(1-\mu)$, $\alpha = \mu + \mu\cdot(1-\mu) = 2\mu-\mu^2$, $\gamma = (2-\mu)/(1-\mu)$; $\beta = 1/(1-\mu)$; $\delta = 1/(2-\mu)$

where the increased estimate for β becomes our increased estimate of frequency when TF = DF, and all non-zero frequencies are 1. Note that $\beta > 1$, $\gamma > 2$.

When TF = DF = μ, there is no distinction between relevant and non-relevant contexts and our definitions are useful only to describe a zeroth order model G^o in which the probability of occurrence in any context, $P_0 = \mu$, is equated with the probability of occurrence in a relevant context, δ. To bootstrap a first order model that takes into account relevance of a context, we estimate the true probability of null contexts that are relevant with no occurrences as $f=\mu$, the observed proportion of contexts that are non-null and hence relevant but have no

recurrences, giving us the parameterization in (59) and estimates of α and δ that satisfy $0<\mu<\alpha<\delta<1$ for all values of TF and DF.

Thus one approach to ensuring accurate demeaning of the non-null contexts is to reestimate β and γ as above when TF=DF and then for all cases subtract β from the actual frequencies, being our estimate of the expected number of occurrences in a *relevant* context (modification β). The error introduced by modification of each non-null context is thus $\beta - \mu > 0$, while an equal error of opposite sign is introduced by modification 2 of each null context. An argument could also be made for subtracting γ, as our estimate of the expected number of occurrences in a *non-null* context (modification γ), however this is unfairly using the single occurrence to define a relevant document and then discounting it as an occurrence – the question of relevance must be decided independently for γ to be a valid estimate.

Note that applying Good-Turing smoothing or Katz-backoff across the different term types will tend to give a *decreased* estimate of the frequency of such rare terms drawn from an open class as assumed for search terms (Gale, AAAA), and its underlying exponential-form is not necessarily appropriate for smoothing across types that obey Zipf's Law (Samuelsson, 1996; Gale, AAAA), although Samuelsson's 'deceptively similar' modification that conforms to the Zipfian distrigution may be. However, the strict version of Zipf's Law implies a finite lexicon in contradiction of our notion of open class and needs to be understood as a sum of distinct distributions with different characteristics (Powers, 1998), and Zipf (1956) also noted that it was dependent on document size.

The implication of point 3 is that we need to go beyond the occurrence count to determine whether a context is relevant – there are approaches based on word similarity using techniques such as LSI or WordNet that

give an independent estimate of the probability of a specific context being relevant. If we are using SVD to calculate the LSI, then we have a recurrence problem: we could come back with a revised estimate and recalculate everything.

Note that $\log(1+x) \approx x$ for small x. Hence it is sufficient to replace the splog value for unit occurrences by $-\gamma$ to implement the modification for TF=DF, representing the lower than expected occurrence counts. More generally, $\log(k+x) \approx \log(k) + x/k$ so that to achieve modification μ resp. α, β or γ it is sufficient to decrement the splog value by μ/k resp. α/k, β/k or γ/k. We will refer to the original model as splog0, and the models implementing the respective modifications as splogμ, splogβ, etc. Model splogμ is both sparsity- and occurrence-preserving, whilst splogβ may introduce additional zeros when β is integral, and the other two are not usable in the absence of independent information about relevance.

The best choice is arguably splogμ in terms of guaranteeing preservation of both sparsity and occurrence information. Moreover, splogμ introduces considerably fewer errors and hence a much lower sum-of-squares error in the high-sparsity expected and desirable distributions for useful search terms in which the number of null contexts (splogβ errors) vastly exceeds the number of null contexts (splogμ errors). In addition, splogμ restricts the effective measure of central tendency to the range between median or mode and geometric mean (see Eqn 53).

Note that the negative values that occur when splogging a frequency that is less than expected could actually be useful in a non-positive condition for use in implementing the Boolean NOT operator – it is arguably better say the document has a lower than expected occurrence rate rather than to force it to have zero occurrences.

Parameterization of the P-mixture

We perform a similar analysis to see how TE handles the pathological cases. As noted above, the K-mixture and P-mixture are not well defined in this case as $\gamma=1$, $\beta=\delta=0$, violating our constraint that $0< \delta<1$ (and α and η are undefined as a result).

Analogous to the corrections in the previous section, we can propose two simple corrections to the K-mixture: 1. setting $\gamma_\omega=1.5$ in the specific case where $TF_w=DF_w$; or 2. incrementing γ in all cases using $\gamma_\omega=\gamma+0.5$. In both cases identities 28 and 30 to 32 can be used to define β_ω, δ_ω, ε_ω and ζ_ω, from γ_ω alone. These are clearly no longer consistent with the empirical values for TF_w and DF_w so we give priority to $DF_\omega \equiv DF_w$ and treat TF_w as being a truncated estimate of the true value $TF_\omega \equiv \gamma_\omega DF_w$ according to identity 34. Identities 27 and 33 may now be used to define α_ω and η_ω. Note that whereas previously we considered an 'add φ to μ' model, here we are parameterizing as an 'add $1/\kappa$ to β and γ' model, recalling γ estimates TF_{wd} and TC_{wd}.

Elaborating the two simple cases and asserting condition 55 as a constraint, we get

$\lambda = TF_\omega = 1.5 \cdot DF_\omega$, $\mu = DF_\omega = DF_w = 2 \cdot \varphi < ⅓$ if $DF_w=TF_w$

$\lambda = TF_\omega = [TF_w/DF_w+0.5] \cdot DF_\omega$, $\mu = DF_\omega = DF_w = 2 \cdot \varphi < ⅓$

More generally, we can reduce the increment in line with the specificity of terms κ:

$\lambda = TF_\omega = (1+1/\kappa) \cdot DF_\omega$, $\mu = DF_\omega = DF_w = \varphi\kappa < 1/(\kappa+1)$ if $DF_w=TF_w$

$\lambda = TF_\omega \equiv TF_w + \varphi = [TF_w/DF_w+1/\kappa] \cdot DF_\omega$, $\mu = DF_\omega \equiv DF_w = \varphi\kappa < 1/(\kappa+1)$

Now equations 26 to 33 define the κP-mixture for these smoothed models. Note that in the DF_w=TF_w case, we have $\beta = \lambda/\mu - 1 = \kappa$ (62 or 63). Furthermore, with $\kappa = 2$ (Eqn 61 ≡ 63) ζ corresponds in value to γ in the original K-mixture and the
κ=2-parameterization of the κP-mixture leads to ζ-parameterization in terms of
$\zeta = TF_w/DF_w$, $\mu = DF_w$, referred to as the ζP-mixture, whereas the K-mixture is parameterized in terms of $\gamma = TF_w/DF_w$, $\mu = DF_w$ representing the limit as κ approaches infinity, and can be also be identified as the γP-mixture. The special case $\kappa = 1$ gives $\varphi = \mu$, $\beta = TF_w/DF_w$ and a standard 'add one' model, referred to as the βP-mixture.

With these models we are setting a fixed $\varphi = \mu/\kappa$ subject to $\mu < 1/(\kappa+1)$ and equivalently $\kappa < (1-\mu)/\mu$ (both forms following from substitution of φ in Eqn 56). This condition must hold for $\mu=DF_w$ for all terms w so we must set κ on the basis of the greatest value of DF_w observed in the corpus. We may use this parameter either for just the DF=TF case (Eqn 60 or more generally 62), or for all cases (61 and 63).

Alternatively, if we use $f = \mu$, $\varphi = \mu/(1-\mu)$ to define the parameterization (59 or 63 with $\kappa = \mu/\varphi = 1-\mu < 1$), the mixture is self-adjusting and automatically satisfies
$\kappa < (1-\mu)/\mu = 1/\varphi = \kappa/\mu$ and thus $\mu < 1/(\kappa+1)$. Observe that this also forces $\kappa < 1$ whereas in the constant κ-parameterizations we have normally tried to set $\kappa \geq 1$. In the limit this auto-parameterization approaches the $\varphi = \mu$, $\kappa=1$ ('add one') case as μ approaches 0 (corresponding precisely to those rarer terms where TF_w=DF_w is more likely, and in which limit we find TF_ω=$2DF_\omega$). In this case of small μ we can also approximate $1/\kappa = 1/(1-\mu) \approx 1+\mu$ and hence $\varphi \approx \mu(\mu+1) = \mu + \mu^2$, being early terms of the GP that defines the G^o model.

The question is whether (63) is a theoretically reasonable generalization of (62) for either the constant or the self-adjusting parameterization using $\kappa = 1-\mu$. In particular, TF_{wd} rates will occur for conditions other than $DF_w=TF_w$ and must occur for $TF_w < 2DF_w$. In these cases the TF_{wd} also represent underestimates of the actual probabilities according to our assumptions about content words recurring in relevant contexts (and so defining bursts or clusters). Assuming the validity of the $TF_w=DF_w$ reestimate (62) and now considering a case where the expected count for term w, $TF_w=DF_w+k$. According to our argument in defining this estimate, the probability of unexpressed occurrence or recurrence has decreased, and there will be exactly one document containing two occurences for the case k=1, so we still find $f=\mu$. For $\mu>k>1$ it is possible to have only documents with one or two occurrences for which $f=\mu$ again follows, but it is also possible to find larger clusters which we would however argue is consistent with the same distribution but simply further from the mode.

We now analyze the role of the probability of a document being relevant given there are zero occurrences varies with k, defining this as p_k satisfying

$$\alpha = P_D(d \text{ Я } w) = P_D(d \ni w) + P_D(d \text{ Я } w \mid d \not\ni w)\, P_D(d \not\ni w) = \mu + p_k(1-\mu)$$

By some simple arithmetic manipulations we see

$$\alpha = [(TF+\varphi)\cdot\mu] / [(TF+\varphi) - \mu] = \mu + \mu^2/[k+\varphi] = (1-\mu)/[k(1-\mu)/\mu^2 + 1/f]$$

This gives us a number of interesting and insightful relationships involving the odds ratio $\rho = \mu / (\mu-1)$ and in particular $\rho\mu$ (which is constant for a given w or $\mu = DF_w$):

$$p_k = \rho\mu/[\varphi+k],\quad \varphi = \mu^2/[f\cdot(1-\mu)] = \rho\mu/f$$

$1/p_k = 1/f + 1/f_k$, $f_k = \mu^2/[k\cdot(1-\mu)] = \rho\mu/k$, $f = f_\varphi = \mu^2/[\varphi\cdot(1-\mu)] = \rho\mu/\varphi$

Recalling that interval scales are often more appropriate and amenable to statistical manipulation than the reciprocal frequency and probability scales, we see that $f = f_\varphi$ defines the k=0 proportion of the null documents, and k specifies increments in terms of the interval φ with the smoothing technique being clear seen as of 'add one' character on this interval scale. Moreover, although (67) is apparently only well defined for k>0, because of being defined in terms of probabilities, this really only illustrates again that the fundamental events are occurrences defining intervals, and that probabilities are not always the most convenient averages to use as the basis for a model. Thus the φ and 1/f terms go to 0 in the K-mixture limit, and the k and $1/f_k$ terms go to 0 in the TF=DF limit, and when both limiting conditions hold we confirm that the K-mixture is not well defined for TF=DF.

What this shows is that whatever justification we have for the k=0 case (TF=DF), it carries over sensibly to k>0 cases in a manner highly reminiscent of Zipfian smoothing methods rather than Good-Turing or the original Katz (1987) backoff scheme, where Good-Turing/Katz-backoff smooths towards an infinite geometric distribution over types that is similar to G^o but Zipf's Law implies a finite inverse linear distribution and both of these appear to define bounds on the true distribution rather than reflecting the distributions directly (Samuelsson,1996; Powers,1998). Thus we conclude that the generalization of (62) to (63) is appropriate for any reasonable definition of κ, and precit that f=μ, κ=1-μ will be the better than constant κ-parameterizations.

Note that Katz (1996) both defined and tested models using the same corpus and aggregations of terms with similar statistics, whereas the

correct way to test the model is to parameterize it using empirical probabilities drawn from a training set and to validate the model and compare models using an independent validation set and then publish results for the selected optimum model for a third independent test set. The partitioning may be done on the basis of documents or terms, or preferably both.

We called the $\kappa=1$ and $\kappa=2$ parameterization according to equations 61 and 63 the β- and ζ-parameterizations defining the β- and ζP-mixtures as discussed above. We define the $\kappa=1-\mu$ parameterization according to equation 63 the μP-mixture, and refer to other parameterizations with a constant κ as κP-mixtures. Equation 62 doesn't define a true parameterization but rather a correction of the one specific degenerate case of the K-mixture, and we thus refer to these as κP-correction with the $\kappa=1-\mu$,
μP-correction being the recommended correction. Note that the constant κ parameterizations and corrections are subject to the constraint $\mu < 1/(\kappa+1)$, and introduces an additional parameter, whilst the $\kappa=1-\mu$ constraint of the canonical P-mixture or P-correction does not introduce an additional parameter. It is now a matter for empirical investigation as to which of these four models is best, and in the case of the κ-parameterization or κ-correction, which value of κ, as well as demonstrating that they are better than the degenerate zeroth order G^o model that arises when TF=DF (this must be demonstrated in application to unseen text where this may or may not hold). The autoparameterized P-mixture may be expressed directly as

$$\lambda = TF_\omega = [TF_w/DF_w+1/[1-\mu]] \cdot DF_\omega,\ \mu = DF_\omega = DF_w = \varphi[1-\mu] < 1/(\kappa+1)$$

The discussed variants of the κP-mixture defined by equations 26 to 33 and 63 (replacing 34 and 35) may be summarized as follows, noting that ρ is the odds ratio $\mu/(1-\mu) = f/\kappa$, that $\varphi = \mu/\kappa = \mu\rho/f$, and that $0 < f <$

1 is required for a valid model so that the γP-mixture is the ill-defined unsmoothed K-mixture defined by $\mu = DF_w$ and one of γ as empirical ratio $\gamma_o = TF_w/DF_w$, δ as $\delta_o = 1 - 1/\gamma_o$ or ε as $\varepsilon_o = 1 - \delta_o = 1/\gamma_o$:

Name	**φ**	**κ**	**f**	**γ**	**$\gamma_o = TF_w/DF_w$**
κP-mixture	μ/κ	κ	κρ	$\gamma_o + \varphi/\mu$	$(\lambda-\varphi)/\mu$
K-/γP-mixture	0	∞	∞	γ_o	**γ**
βP-mixture	**μ**	1	**ρ**	$\gamma_o + 1$	**β**
ζP-mixture	μ/2	2	2ρ	$\gamma_o + ½$	**ζ**
μP-mixture	**ρ**	1-μ	**μ**	$\gamma_o + 1 + \rho$	$(\lambda-\rho)/\mu$

These models and the corresponding corrections to just the TF=DF case may all be tested empirically by modeling on one (sub) corpus and validating on another. Such validation may include direct term fitting or practical application of the corrections to some application, such as the information retrieval search domain.

The true Katz models, G= G″ and G′, and our generalization, $G^{\{m\}}$, depend on a concept of relevance or topicality that requires bursts of more than one occurrence (viz. $k \geq 2$) to define topicality in the G model, and in $G^{\{m\}}$ that generalizes to $k \geq m$. In our smoothing process for first order models, $P_0 = \delta_0 = \mu$ doesn't change and $P_1 = \delta_0$ can be directly defined from the above table as $\delta = 1 - 1/\gamma$. To generalize this to higher order models we distinguish two possibilities.

The first possibility is to adjust only $P_m = \delta_m \approx C_m$ on the basis that all $k<m$ are equally indeterminate in defining relevance although some may nonetheless be topical but not have the expected additional recurrences. This may be done analogously to the above using μ_m as the

probability of topicality ($k \geq m$ occurrences) in place of μ and δ_m as the conditional probability of recurrence in a topical context (with $k \geq m$ occurrences) in place of δ. Note that the geometric series defined by $(1-\delta_m)\delta_m^{k-m}$ sums to unity for $k \geq m$ and thus sum over $k \geq m$ of the contributions of the final term of $G^{\{m\}}$ are independent of k so that adjustment of δ_m does not require any concomitant adjustment of δ_k for $k<m$ to ensure the sum over all $P_w(k)$ is unity as required for a probability distribution. This is then the first approach to consider.

The second possibility is to adjust all P_k for $0<k \leq m$. This requires additional assumptions, and potentially a whole family of m constants κ_k and totally separate parameterizations $\delta_k^{(m)}$ for each $G^{\{m\}}$. The obvious way of avoiding the first of these problems, the explosion of new assumptions, is to assume that the probability of non-occurrence in a relevant context is the same whether or not the $k \geq m$ condition for topicality has formally been met, and independent of k. This is a generalization of the assumption that led us to define the μP-mixture using $f = \mu$, and hence $\varphi = \rho$ and $\kappa = 1-\mu$.

The second of these problems, with its potential for overfitting, can be dealt with by defining the parameters δ_n sequentially according to approach one for the desired $n=m$ as well as for all $n<m$, and incorporating them into a variant of our second approach. This may however tip the balance to the opposite extreme and lead to a parameterization that is too smooth, but both the first approach to smoothing and the original $G^{\{m\}}$ model are more likely to be overfitted to the training data even though the definition of P_m is already an average over $k \geq m$. Note that empirically $P_w(k)$ is rarely strictly monotonic for a given w.

The optimal value of m may be a function of D, the typical document size, as this imposes a limit to k, viz. $k << D$. Generally, we prefer a

model that does not tune κ, f or φ to specific w and where m is a function of D or selected by observing when the order n>m fit leads to $\delta_m \approx \delta_n$.

The assumptions and notations underlying the unsmoothed $G^{\{m\}}$ model may be summarized as

$$d \ni w \leftrightarrow k \geq m \rightarrow w \text{ Я } d;\ \ C_m = P_m;$$

That is, we define membership as having a threshold equal to the order of the model and implying relevance without relevance implying membership, and also treat the average conditional probability of additional occurences P_k as being constant C_m for k≥m.

The additional assumptions generalizing the μP-mixture approach to $G^{\{m\}}$ are

$$P_w(w \text{ Я } d \mid k=0) \equiv P_w(k=m \mid k \geq m \ \&\ w \text{ Я } d) = \mu_m$$

That is the probability of a context being relevant despite seeing no occurrences is equated with the probability of finding no further occurrences after seeing m. The first approach uses this only when m is the order of the model, whilst the second approach uses it for any m≤n where n is the order of the model desired. These are thus applied only to determine P_m in our first level approach to smoothing, giving $G^{[m]}$, but in our final variant we coopt the P_m (m<n) values of the low order models with first level smoothing as the P_k used to define our fully smoothed model $G^{(n)}$. Variants corresponding to the κP-mixture may be used in place of assumption of identity 70, leading to models we denote $G_\kappa^{[n]}$ and $G_\kappa^{(n)}$, but as discussed above the latter may require a whole family of additional parameters κ_k dependent on μ and possibly m.

Modifications to term entropy model

The TFIDF and TE models involve a log factor so we investigate if it can be made sparsity preserving by applying an orthogonal demeaning (the log factor and log term depend on w only, whilst the linear factor depends on both w and d). IDF is not appropriate for sparse orthogonal demeaning as it is constant for a given word and thus simply weights binary occurrence vectors in the word vectors of the dual formulation. We will also see that TF can be sparsity-preserved during demeaning.

Orthogonal demeaning in IR means taking the mean across documents for each word, notwithstanding that we are primarily interested in using words as signature vectors to characterize documents, and thus would normally demean those vectors. We will thus swap to thinking in terms of comparing the similarity of words, with their similarity of sense being measured by the similarity of their signature vectors of document occurrence information. The first step is to note that the -log(η) term in the TE model is constant for a given word and is thus preserved in the mean of the corresponding signature vector. This is similar to the IDF binary occurrence weighting and should similarly be retained rather than demeaned in order to distinguish typical occurrences from non-occurrences. This distinguishing term is missing from the TFIDF model, and some such term would thus need to be introduced, so the TFIDF model will become more like the TE model if we make it sparsity and occurrence preserving, with the obvious modification being to add in an IDF term (cf. equation 37). This is equivalent to using an ‘add one’ correction on TF: TF1IDF = (TF+1)*IDF.

The IDF factor in $(TF_{wd}+1)*IDF_w$ and the -log(δ) factor in TE are constants (given w) and the mean over the full set of documents of TF_{wd} is TF_w. Thus the orthogonal means are $(TF_w+1)*IDF_w$ and -

$\log(\delta)*TF_w - \log(\eta)$ resp. These means represent the average document in the sense that its distribution of terms resembles the corpus as a whole. However, such a document would typically be encyclopaedic and thus not be representative in size (number of words in the sense of tokens irrespective of type) or lexical coverage (number of distinct terms or types). But note that we have actually assumed all documents are the same size or are normalized to equivalent size (in tokens), which also limits its lexical coverage L (in types). In fact, the fact that most actual documents are missing many terms (types) indicates they are not representative and the sign of the deviations from the orthogonal means tells us how much more or less than expected a word occurred in the document.

If we average only over the documents that contain a term, then the average of TF_{wd} is TF_w/DF_w since it occurs in only DF_w/N of the corpus documents. This we recognize as the pivotal ratio γ in the K-mixture or ζ in the $\kappa=2$ ζP-mixture model or β in the $\kappa=1$ βP-mixture. This leads to orthogonal means of $(TF_w/DF_w+1)*IDF_w$ and $-\log(\delta)*TF_w/DF_w - \log(\eta)$ for our respective models.

The vector mean, TF_d, corresponds to the typical (but non-specific) word frequencies of the documents for non-stopwords, and is not likely to vary much with the topic of the document, but rather reflects the nature of the document (encyclopaedia, dictionary, paper, story, letter, etc.) We thus drop the subscript and treat TF as a corpus or language constant. Words that occur more or less frequently than this represent words that are expected to be more or less significant in the implausible TF model. In TFIDF or TE (Eqn 37) it is harder to characterize as both factors are dependent on w and one is dependent on d as well. However, if we do a double demean to avoid the dependence on d, we get an average value for TF_d that we define as the constant TF which

for all variants of the TFIDF and TE models is a simple document-independent average of a function of our model parameters.

Intuitions about document topicality

Suppose all our documents are of size D and we have a document that topically features word 1 (train, say) at some high level $f_{11} = D*p_{11}$ and a second document that topically features word 2 (dog, say) at some high level $f_{22} = D*p_{22}$. We will further assume no occurrences of word 1 in document 2 or vice versa. If we were simply to append the documents the new document would have size 2D, $f_{s1} = f_{11} = D*p_{11} = 2D*p_{s1}$ etc. The absolute frequency of a term in the combined document is the same as in the respective original documents, but its probability is halved, and the expected interval between occurrences is doubled (although this is misleading as all instances of word 1 will be in the first half and word 2 in the second half). This combined document is moreover no more use to us than the separate documents – this we call encyclopaedic combination where words are topical in separate entries, but the overlaps and relationships are not in focus.

If we now summarized the combined document down to size D, it is unlikely that we would halve the frequencies of the two terms given these remain topical as intended, but nor is it likely that we could maintain the full frequency of occurrence given the independence of the topical domains as discussed in the documents. The act of summarization will furthermore not make the document more relevant which would require overlaps to be discussed, although its lesser size may make it more useful – and even though size is factored out this additional density may be emergent and this is one reason why normalizing by the frequency of the most frequent term (content word) in a document can be usefully used for TFIDF rather than mere

document size (whether in words or terms). On the other hand, such a measure will also reflect changes in style, genre and register (e.g. age or vocabulary or author or audience, length of sentences, use of anaphora, use of hyponyms, hypernyms, technical terms and circumlocutions).

We could make the documents more relevant by adding new material that discussed relationships between the focus terms, but to do that we would have to drop existing material, perhaps replacing irrelevant examples for one term with examples that are in the topic domain of the other (e.g. taking a dog on the train instead of a bicycle, training a dog instead of rat). If there is implicit topical relevance in the other document, but it is not obvious due to use of a paraphrase, synonym, hyponym or hypernym, the summarization could make the relevance more apparent without changing its actually relevance (e.g. taking a dog on the train instead of an animal, training a dog instead of an Alsatian). This problem can be accommodated by using either implicit (e.g. LSA/LSI) or explicit (e.g. WordNet) semantic matching.

The Frequency and Time domains

So far we have assumed that we are dealing with frequencies to the exclusion of consideration of time (viz. the order of words) and using a fairly large window to determine the frequencies (a document of size D which is typically of the order of 2000 words or more, and sometimes as large as entire chapters or even books). For many purposes, a smaller context is more appropriate, such as a clause, sentence or paragraph. In inducing syntactic information some knowledge of the word order is essential and often this information is determined over small contexts of a dozen words or less. For speech processing, the contextual information is often very limited and stored in the form of a Hidden

Markov Model derived from Ngram statistics regarding sequences of N units (classically words). Models of the same order are also used in various compression schemes, and these models have also been used for statistical or semantic modelling since the 1960s (Wolff) as well as in the more general Minimum Message Length and Minimum Description Length models of theory evaluation. These have the advantage that sufficient information is retained to reconstruct a message indistinguishable from the original whilst maximizing the information that is implicitly stored in the compression model. In speech processing this level of physiological indistinguishability includes retaining all the information a human hearer would use to identify the speaker and distinguish characteristics such as gender, accent, emotion, etc.

In recording and analysing real world situations and data, these techniques can be extended to include multiple sensors and modalities, including multiple microphones (for noise localization and cancelling), cameras (for lip-reading and gaze tracking) and biosensors (for brain computer interface and monitoring of attention, stress, etc.) This raises further issues about the signal processing, compression and fusion of massive amounts of information from diverse sources. Again a standard approach is to define frames, windows or epochs of a relatively small size, and possibly on multiple scales and across multiple dimensions, and to perform spatial, temporal or spatiotemporal frequency analysis.

However, a common early step after frequency analysis is to compress using SVD, and this is a common precursor to more advanced techniques signal separation techniques such as ICA. Typically multiple sensors give rise to multiples streams of time-varying signal for data with dimensions sensors x time or sensors x frames x times,

where the frames are arbitrary or psychologically plausible time segments. The standard technique for converting into the frequency domain is to convert the time information in to frequency information using an FFT or similar. This gives the data the dimensions sensors x frequency (spectrum) or sensors x frames x frequency (spectrogram). Performing an SVD and/or an ICA transformation can now be used to reduce this to sensors x sources and sources x frequency or sources x frames x frequency.

However, since SVD and ICA are linear operations, and FFT and IFFT are also linear operations, the SVD and ICA will recover the same sources when applied direct to the time domain data as when applied to the frequency domain data, and apart from rounding and training error the frequency data extracted is the same whether the FFT is performed first or last. These leads to two conclusions: there is no point in performing an IFFT after SVD to take the sources back into the time domain, and it is more efficient to perform the SVD dimension reduction before the FFT than after. Note that whilst the SVD and ICA matrices need to be calculated based on the data (though theoretically they are then stable for stationary sources), the FFT and IFFT are fixed matrix multiplications, and a Symmetric Fourier Transform (SFT) actually expresses all the sin (imaginary) and cosine (real) components of the FFT in a rectangular real-valued matrix whose transpose acts as a pseudoinverse (viz. ISFT = SFT′).

To the extent that other forms of frequency analysis are linear and invertible, this would also be true for those techniques. However the other techniques do not in general have these properties, and even the windowing techniques applied to FFT represent deviations from these properties. DCT for example throws away phase information to double the frequency resolution. Wavelets sacrifice linearity for time

specificity. Linear Predictive Coding (LPC) and Auto-Regression (AR) sacrifice the ability to distinguish phase, frequency and reverberation for a smoother frequency envelope. Cepstrum takes the logarithm of power between the FFT and IFFT operations, thus sacrificing linearity for increased frequency specificity and rejection of reverberation whilst counting the frequency of frequencies, which is particularly useful for establishing the fundamental where many harmonics are present.

Schur

Tensor

References

Church, K.W. and Gale, W.A. (1995) Poisson mixtures. *Natural Language Engineering* **1**(2):163-190.

Church, K.W. and Gale, W.A. (BBBB) Inverse Document Frequency (IDF): A measure of deviations from Poisson.

Gale, W.A. (AAAA) Good-Turing smoothing without tears.

Gao, Jianfeng and Lee, Kai-Fu (XXXX) Distribution-Based Pruning of Backoff Language Models

Gao, Jianfeng, Li, Mingjing and Lee, Kai-Fu (YYYY) N-gram Distribution-Based Language Model Adaptation

Katz, Slava M. (1987). Estimation of Probabilities from sparse data for the language model component of a speech recognizer, *IEEE Trans. On Acoustics, Speech and Signal Processing* **35**(3), 400-401.

Katz, Slava M. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering* **2**(1):15-59.

Powers, David M. W. (1998), "*Applications and Explanations of Zipf?s Law*", pp151-160, **NeMLaP3/CoNLL98 Joint Conference**, Sydney, January 1998. (Research/AI/papers/199801c-CoNLL-Zipf.pdf)

Samuelsson, C. (1996) Relating Turing's formula and Zipf's law. WVLC'96

Umemura, K., Takeda, Y., Tanaka, M. Feng, L., Yamamoto, E. (ZZZZ) Empirical Term Weighting

Zipf (1956), The principle of least effort.