Latent Semantic Indexing

Lecture 19, November 6, 2007
CS 674/INFO 630 Fall 2007
Lecturer: Lillian Lee
Scribes: Vladimir Barash, Stephen Purpura, Shaomei Wu

December 10, 2007

1 Background

In the previous lecture, we discussed the Singular Value Decomposition (SVD) of the term-document matrix $D \in \Re^{m \times n}$ where n is the number of documents in the corpus and m is the number of terms in the vocabulary. With the help of SVD (which is unique up to sign if the singular values are distinct), we can decompose an $m \times n$ term-document matrix into three special smaller matrices. The result is frequently abbreviated $D = U\Sigma V^T$ where:

$$U = \underbrace{\overbrace{\begin{bmatrix} \uparrow & \uparrow & \uparrow \\ | & | & | \\ \vec{u_1} & \dots & \vec{u_r} \\ \downarrow & \downarrow & \downarrow \end{bmatrix}}^{m \times r}}_{\text{left singular vectors}}, \Sigma = \underbrace{\overbrace{\begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_r \end{bmatrix}}^{r \times r}}_{\text{singular values}}, V^T = \left.\overbrace{\begin{bmatrix} \leftarrow & \vec{v_1} & \rightarrow \\ \leftarrow & \dots & \rightarrow \\ \leftarrow & \dots & \rightarrow \\ \leftarrow & \vec{v_r} & \rightarrow \end{bmatrix}}^{r \times n}\right\} \text{right singular vectors,}$$

r is the rank of D; the columns of U and V each form an orthonormal basis for their span; and the singular values are ordered such that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$.

For the n-dimensional hypersphere of linear combination coefficients $A_2 = \{\vec{\alpha} \in \Re^n \mid \|\vec{\alpha}\|_2 = 1\}$, the hyperellipse DA_2 which has semiaxes given by the

$\sigma_i\overrightarrow{u_i}$s succinctly approximates the convex hull of the document vectors ($\vec{d_i}s$) and their negatives. DA_2 represents a summary of the overall shape of the corpus and it is sensitive to the document distribution.

The advantages of SVD include (1) reduced storage space and (2) reduced computational complexity when performing calculations on the three component matrices as opposed to on D.

Assuming $\sigma_1 \neq \sigma_2$, of special interest is the first singular vector, $\vec{u_1}$, where

$$\vec{u_1} \;=\; \operatorname*{argmax}_{\{\vec{u}:\|\vec{u}\|_2=1\}} \sum_{i=1}^{n} (\|\vec{d_i}\|_2 cos(\angle(\vec{d_i},\vec{u})))^2$$

That is, $\vec{u_1}$ represents a direction of best fit for the corpus, determined in part by considering common directions among the $\vec{d_i}$ document vectors and in part by emphasizing the longest $\vec{d_i}s$.

Assuming all the singular values are distinct, each subsequent $\vec{u_i}$ represents the "next best" fit to the corpus, in the sense of the equation above, if we first subtract from each $\vec{d_i}$ its orthogonal projection onto span($\{\vec{u_1},...,\vec{u}_{i-1}\}$).

2 Improving the Representation

The representation of a document corpus via SVD is already a vast improvement over its representation via an $m \times n$ term-document matrix. But what if we could transform this representation into an even more compact (albeit lossy) one? Consider a fixed $k < rank(D)$ and the corresponding rank-k matrix

$$\hat{D} = \hat{U}\hat{\Sigma}\hat{V}^T$$

where:

$$\hat{U} = \overbrace{\begin{bmatrix} \uparrow & \uparrow & \uparrow \\ | & | & | \\ \vec{u_1} & \dots & \vec{u_k} \\ \downarrow & \downarrow & \downarrow \end{bmatrix}}^{m \times k}, \hat{\Sigma} = \overbrace{\begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_k \end{bmatrix}}^{k \times k}, \hat{V}^T = \overbrace{\begin{bmatrix} \leftarrow & \vec{v_1} & \rightarrow \\ \leftarrow & \dots & \rightarrow \\ \leftarrow & \dots & \rightarrow \\ \leftarrow & \vec{v_k} & \rightarrow \end{bmatrix}}^{k \times n}.$$

By the Eckart-Young theorem, this $\hat{D}$ is the best approximation with regard to the matrix 2-norm and the Frobenius-norm to D among all matrices of rank k.

3 LSI

LSI (Latent Semantic Indexing)[Deerwester et al.1990] is based on the principle of applying the Eckart-Young theorem to a term-document matrix (i.e., creating an "Eckart-Young projection" of a term-document matrix) to represent a corpus. Similarly, LSA (Latent Semantic Analysis) refers to applications of the Eckart-Young theorem to data matrices in situations other than information retrieval.

The underlying assumption of LSI is that small singular values (and the associated $\vec{u_i}s$) represent noise in the corpus, so that the original individual document vectors ($\vec{d_i}s$) are located in too many dimensions. This is certainly plausible, as individual terms exhibit correlations, and it is quite likely that not all terms in any given document are necessary to provide an accurate (from the point of view of IR) representation thereof. Finally, proponents of LSI claim the existence of potentially better geometric relationships between the $\hat{D}_i$ vectors than between the D_i document vectors.

To re-state, the goal is to have a better representation of elements and their relationships to one another. Since the terms exhibit correlations, there may exist better document vectors in a k dimensional space, $k \ll r$. The directions of the hyperellipse DA_2 ($\vec{u_i}s$ assuming distinct singular values) that have small lengths (small singular values) might correspond to noise (noisy features). If the small singular values represent noise, $\hat{D}$ gives new (and hopefully better) geometric relationships among the document vectors.

3.1 LSI and Computational Complexity

It is very important to note that LSI does not require storage of, or operations upon, the full $\hat{D}$ matrix. Consider a common measure of similarity between two document vectors, $\hat{d_i} \cdot \hat{d_j}$. Let us first rewrite $\hat{D}$:

$$\hat{D} = \overbrace{\begin{bmatrix} \uparrow & \uparrow & \uparrow \\ | & | & | \\ \vec{u_1} & \dots & \vec{u_k} \\ \downarrow & \downarrow & \downarrow \end{bmatrix}}^{m \times k} \overbrace{\begin{bmatrix} \uparrow & \uparrow & \uparrow \\ | & | & | \\ \vec{\beta_1} & \dots & \vec{\beta_n} \\ \downarrow & \downarrow & \downarrow \end{bmatrix}}^{k \times n}$$

where $B = \hat{\Sigma}\hat{V}^T$. Then $\vec{\beta_i}$ is a vector corresponding to the coordinates of $\hat{d_i}$ in the basis $\vec{u_1}, ..., \vec{u_k}$. Then:

$$\hat{d_i} \cdot \hat{d_j} = \vec{\beta_i} \cdot \vec{\beta_j}$$

so the actual computation of $\hat{D}$ is unnecessary.

In a similar vein, computing similarity measures (for example, cosine similarity) between some query and documents in the corpus is a relatively simple matter: consider the query vector $\vec{q}$. The cosine similarity between q and some d_i in D as measured between their vector representations in $\Re^m$ is:

$$cos(\vec{d_i}, \vec{q}) = \frac{\vec{d_i} \cdot \vec{q}}{\|\vec{d_i}\|_2 \times \|\vec{q}\|_2}$$

where $\|\vec{q}\|_2$ is document independent and would not actually be computed.

Now, what about computing cosine similarity in the new k-dimensional subspace S_U, namely, the subspace with orthonormal basis $B_U \overset{def}{=} \{\vec{u_1}, \dots, \vec{u_k}\}$? The k coordinates of the orthogonal projection of $\vec{q}$ onto S_U with respect to B_U are given by the vector $\vec{\hat{q}} = (\vec{q} \cdot \vec{u_1}, ...\vec{q} \cdot \vec{u_k})^T$. So, we can assign a score between the projections of $\vec{q}$ and $\vec{d_i}$ onto S_U as follows, by analogy with the equation above:

$$\frac{\vec{\beta_i} \cdot \vec{\hat{q}}}{||\vec{\beta_i}||_2}$$

(note that one can pre-compute the L_2 norms of the n projected document vectors).

3.2 Observations About LSI

In general, LSI is a technique for (hopefully/partially) capturing term co-occurrence patterns within a corpus. These co-occurrences can correspond to conceptual similarity, e.g. between the terms "stochastic" and "probabilistic", or to collocations, e.g. the terms "humpty" and "dumpty". It is not clear whether either of these co-occurrence types is preferable to the other.

An important feature of LSI is that it makes no assumptions about a particular generative model behind the data. Whether the distribution of terms in the corpus is "Gaussian", Poisson, or some other has no bearing on the effectiveness of this technique, at least with respect to its mathematical underpinnings. Thus, it is incorrect to say that use of LSI requires assuming that the attribute values are normally distributed.

Finally, the "latent semantics" of the singular vectors are not generally interpretable. There is no reason that $\vec{u_1}, ..., \vec{u_k}$ should correspond to clearly identifiable topics or memes in the corpus. First, the $\vec{u_i}s$ are not unique. Second, the singular vectors are forced to be orthogonal by construction, whereas topics (for instance, "physics", "chemistry", "math" in a corpus of high school textbooks) are not necessarily so. Third, dimensionality selection is always an issue, because there is no clear basis for picking a value of k. In general, the researcher looks for a gap in the singular values or uses empirical analysis such as cross-validation to select a value that yields the best results for his or her IR task. But no general bounds on such a value are known.

Despite these considerations and no guarantees of performance in information retrieval, the technique (along with other few-factor representations) sometimes does significantly improve performance of ad-hoc information retrieval and many other applications as well.

4 Few-factor Representations

LSI is one of a host of few-factor representation techniques for looking at corpora. Other techniques in the same category include pLSI, LDA, information-bottleneck, clustering, etc. Most of these techniques sacrifice computational simplicity in order to replace the singular vectors with more clearly interpretable alternatives. It is worth mentioning one such technique, known as the 'Chinese Restaurant Process' with an example application given in [Blei et al.2003]. The Chinese Restaurant Process takes a generative viewpoint, representing individual models as tables in a Chinese restaurant. When a new document is created ("arrives at the restaurant") it is seated at one of the existing tables with some probability proportional to the number of items already "seated" there, or assigned to a new table. The resulting model induces a rich-get-richer effect, wherein tables with many documents fill up even more. Similarly, tables with few documents remain relatively unoccupied.

The point of mentioning the Chinese Restaurant Process is that use of this model suggests alternate means of pre-specifying the number of 'factors' (models, in this case).

5 Finger Exercise

5.1 Question 1

Given a corpus C with 3 documents and 4 terms with term-count distributions as shown below:

	cat	dog	household	love
d_1	2	2	0	1
d_2	2	0	0	0
d_3	0	4	4	1

First compute the SVD for the document matrix D of this given corpus.

Can you give some intuitive descriptions of $\vec{u_i}$s(left singular vectors), σ_is(singular values) and $\vec{v_i}$s(right singular vectors)? (In the vector space, please recall the mapping from hyperpolygon to hyperellipse we did in class.)

Sol:

As given in the problem,

$$D = \begin{bmatrix} 2 & 2 & 0 \\ 2 & 0 & 4 \\ 0 & 0 & 4 \\ 1 & 0 & 1 \end{bmatrix}$$

$$U = \begin{bmatrix} -0.1202 & -0.9060 & -0.4059 \\ -0.7373 & -0.1262 & 0.5000 \\ -0.6304 & 0.3727 & -0.6450 \\ -0.2110 & -0.1563 & 0.4113 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 6.0042 & 0 & 0 \\ 0 & 2.9837 & 0 \\ 0 & 0 & 1.0232 \end{bmatrix}$$

$$V^T = \begin{bmatrix} -0.3208 & -0.7443 & 0.5858 \\ -0.0400 & -0.6073 & -0.7935 \\ -0.9463 & 0.2780 & -0.1650 \end{bmatrix}$$

Let DA_2 be the hyperellipse derived from D, thus $\vec{u_i}$s represent the directions of DA_2's axes, σ_is represent the lengths(from origin) of the hyperellipse's axes(w.r.t the axis represented by $\vec{u_i}$), and $\vec{v_i}$s are the pre-images under D of the $\sigma_i \vec{u_i}$s: $D\vec{v_i} = \sigma_i \vec{u_i}$.

5.2 Question 2

Following with previous question, if we take $k = 2$, i.e., consider the smallest singular value 1.0232 as noise, the new document matrix will be

$$\hat{D} = \begin{bmatrix} 2.2433 & 1.6704 & -0.0685 \\ 1.7003 & 0.4060 & 4.0844 \\ 0.3866 & -0.5237 & 3.8911 \\ 0.7535 & 0.3339 & 1.0694 \end{bmatrix}.$$

Obviously storing $\hat{D}$ won't save any space comparing to D. As claimed in the lecture, we only need to store B, which is the matrix corresponding to the coordinates of D in the basis $\vec{u_1}, ..., \vec{u_k}$. However, is that always better?

Sol:

The storage problem for LSI is mentioned in [Langville and Meyer2006] (Page 5). As we can see, both $\hat{D}$ and B seem more complicated when compared to the original D. If we store B in this problem as a matrix consisting of *double*-typed numbers, the space needed to store B will be $64 \times k \times n = 384$bits. Meanwhile, we can store D as a matrix consisting of *int*-typed numbers and hence the space needed for D will be at most $32 \times m \times n = 384$bits as well; however, since D is very sparse, we don't actually need to store all $m \times n$ *int*-typed numbers, so the space to store D can be further reduced. In this case, picking $k = 2$ actually doesn't help at saving storage. Actually, if we store B as a *double*-typed matrix and D as an *int*-typed matrix, then k needs to be less than $m/2$ to have LSI use less space, when D and B have the same density. When the density difference of D and B is taken into consideration, k should be further reduced for LSI to use less storage than regular VSM, which might not be possible without losing too much information from the original corpus. Hence, even though we pointed out in previous notes that “reduced storage space” is one of the advantages of SVD, it is not the main motivation for this technique. The main motivation for this technique is to build a better term-document matrix representation of the corpus.

6 Deeper Thought Exercise

6.1 Question

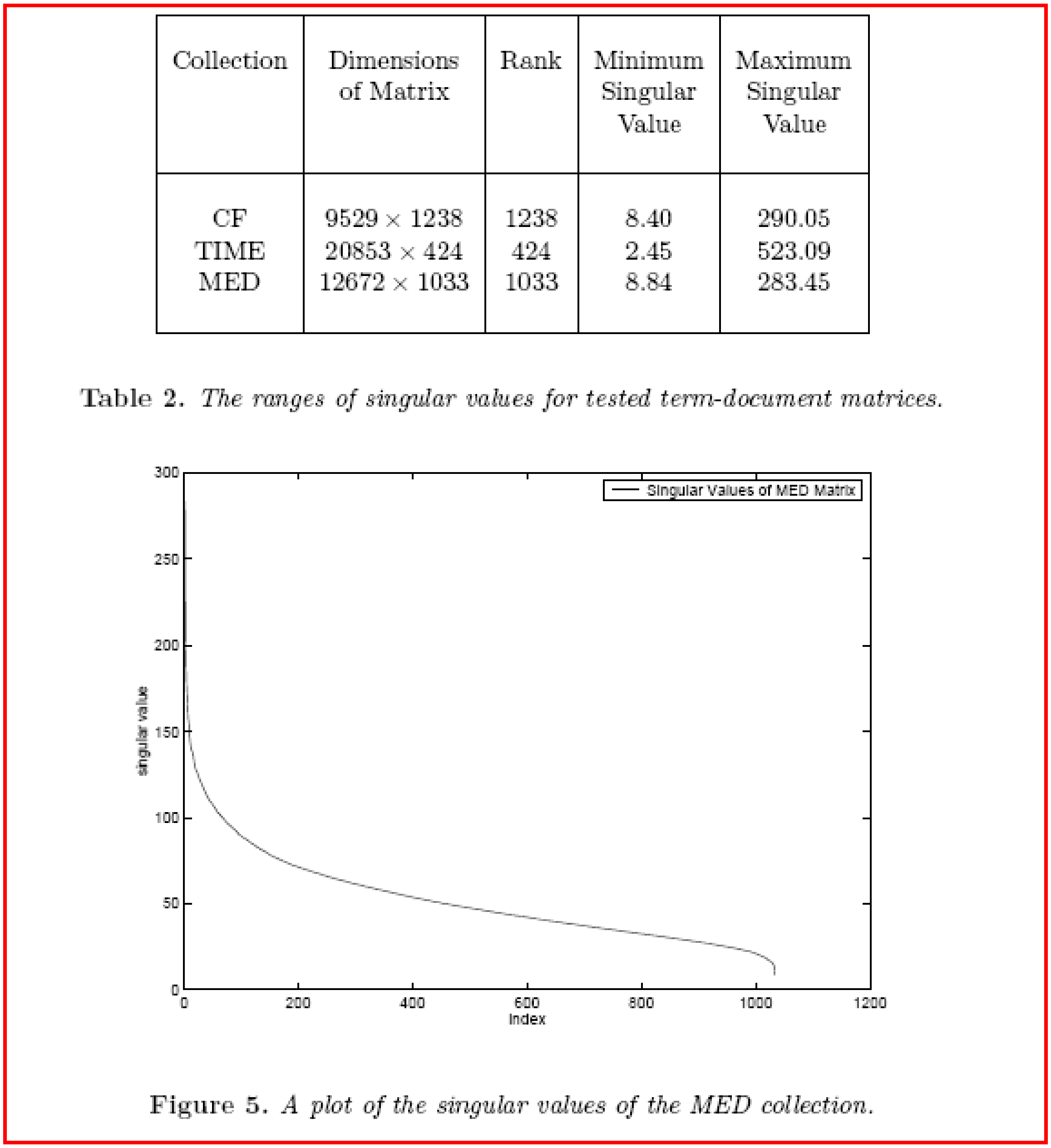
Consider the relative values of some set of σ_is, and corresponding $\vec{u}_i$s produced by singular value decomposition of some term-document matrix D. Are all of these values necessary to represent the corpus? Or are the small singular values just “noise in the corpus”? If they are noise, can we make any observations about what values are too small to be considered relevant?

6.2 Answer

To some extent, justifying that the small singular values are just noise requires determining the optimal rank, k, of $\hat{D}$ for use in LSI. For example, in [Jessup and Martin2001], the optimum k of three corpora were examined. [Jessup and Martin2001] used a a collection of articles from TIME magazine (TIME) from 1963, a collection of Medline (MED) articles on various medical topics, and the Cystic Fibrosis collection (CF) which consists of a set of Medline articles containing the phrase Cystic Fibrosis. Figure 1 shows the minimum and maximum singular values for the three corpora and a graph of all of the singular values for the MED corpus. Notice that (1) all the singular values for the corpora are greater than one and (2) despite the strong performance of LSI on the MED corpus, there is no non-trivial phase transition in the graph of the singular values of the MED corpus.

Rather than examining the singular values themselves, it is more informative to examine the impact of reducing the dimensionality on the error in the representation of the term-document matrix. Consider Figure 2 from Ando's PhD thesis[Ando2001], which illustrates the shape of the error bounds as the dimensionality of the approximated term-document matrix increases from 1 to rank(D).

The three error equations in Figure 2 are equations 3.1, 3.2, and 3.3 of Ando (2001) (see Ando's PhD thesis for the full definitions and theorems).



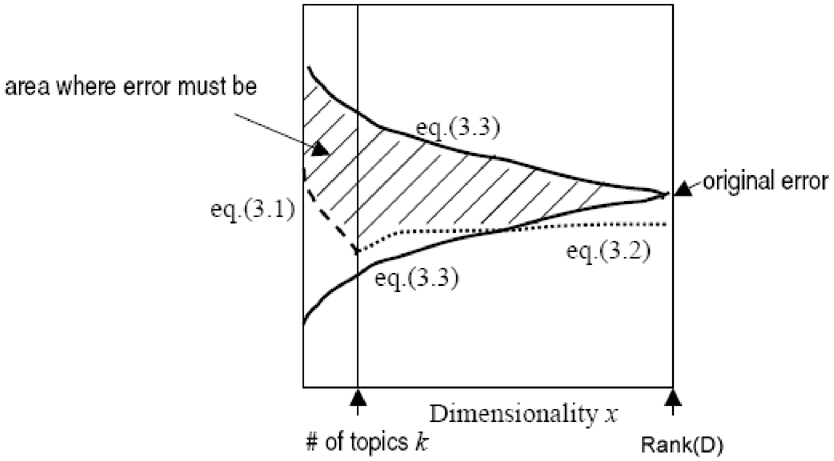
Collection	Dimensions of Matrix	Rank	Minimum Singular Value	Maximum Singular Value
CF	9529 × 1238	1238	8.40	290.05
TIME	20853 × 424	424	2.45	523.09
MED	12672 × 1033	1033	8.84	283.45

Table 2. *The ranges of singular values for tested term-document matrices.*

Figure 5. *A plot of the singular values of the MED collection.*

Figure 1: The range of singular values in TIME, MED, and CF corpora from page 12 of Jessup and Martin (2005)

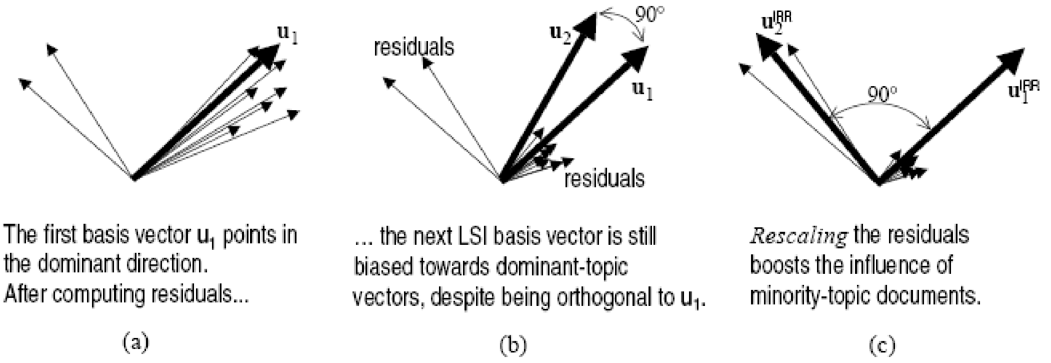
The error functions conceptually represent the bounds where the error must be located as the dimensionality is changed from the original term-document matrix. Ando's thesis suggests that non-uniformity in the distribution of topics among the documents (coupled with the presence of "topic mixing" of multiple topics in a single document) significantly impacts where the actual error in the representation lies for a given dimensionality. A very lopsided distribution with a lot of topic mixing produces the greatest upper bound on the error for LSI. A perfectly uniform distribution of topics (k) over documents with little inter-mixing produces the best upper bound on the error for the LSI representation for IR at around k topics, where k is the number of actual topics in the corpus.



,

Figure 2: Schematic illustration of LSI error bounds and dimensionality; Figure 3.3 from Ando PhD thesis (2001) (modified for display)

Now consider Figure 3, which shows that when topics are distributed non-uniformly among documents, LSI tends to be biased toward representing the sub-structure of larger topics. [Ando and Lee2001] demonstrates a method (LSI with Iterative Residual Rescaling) to adjust for the problems caused by non-uniformity of the topic distribution by producing a representation which can perform better than LSI and VSM (in the tested models).



The first basis vector $\mathbf{u}_1$ points in the dominant direction. After computing residuals...

(a)

... the next LSI basis vector is still biased towards dominant-topic vectors, despite being orthogonal to $\mathbf{u}_1$.

(b)

Rescaling the residuals boosts the influence of minority-topic documents.

(c)

Figure 3: Effect of non-uniformity of topic-document distribution on LSI and how IRR compensates; Figure 4.2 from from Ando PhD thesis (2001)

Returning to the motivating questions, Figure 2 can help us construct responses. When the actual distribution of topics in the corpus is more lopsided or inter-mingled, it is less likely that decreasing rank-k results in a better representation of the corpus (i.e. noise reduction). So, considering the relative size of the singular values can be misleading. However, in practice, the true topic-document distribution is unknown, so methods like IRR may be more effective than the original LSI because IRR adjusts for non-uniformity by changing the construction of the representation to account for it.

References

[Ando and Lee2001] R. Ando and L. Lee. 2001. Iterative Residual Rescaling: an analysis and generalization of LSI.

[Ando2001] Rie Ando. 2001. *The Document Representation Problem: An Analysis of LSI and Iterative Residual Rescaling.* Ph.D. thesis, Cornell University, Ithaca, NY.

[Blei et al.2003] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. 2003. Hierarchical topic models and the nested chinese restaurant process. *Neural*

Information Processing Systems (NIPS), 16. http://www.cs.princeton.edu/~blei/papers/BleiGriffithsJordanTenenbaum2003.pdf.

[Deerwester et al.1990] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407. http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf.

[Jessup and Martin2001] Elizabeth R. Jessup and James H. Martin, 2001. *Taking a Closer Look at the Latent Semantic Analysis Approach to Information Retrieval*. SIAM Press.

[Langville and Meyer2006] Amy N. Langville and Carl D. Meyer. 2006. Information Retrieval and Web Search.