Data Normalization for Variation in Document Length in Exploratory Multivariate Analysis of Text Corpora

Abstract

The advent of large electronic text corpora has generated a range of technologies for their search and interpretation. Variation in document length can be a problem for these technologies, and several normalization methods for mitigating its effects have been proposed. This paper assesses the effectiveness of such methods in specific relation to exploratory multivariate analysis. The discussion is in four main parts. The first part states the problem, the second describes some normalization methods, the third identifies poor estimation of the population probability of variables as a factor that compromises the effectiveness of the normalization methods for very short documents, and the fourth proposes elimination of data matrix rows representing document which are too short to be reliably normalized and suggests ways of identifying those documents.

1. I**ntroduction**

The advent of large electronic text corpora has generated a range of technologies for their search and interpretation. Variation in document length can be a problem for these technologies, and several normalization methods for mitigating its effects have been proposed. This paper assesses the effectiveness of such methods in specific relation to exploratory multivariate analysis [5, 10]. The discussion is in four main parts. The first part states the problem, the second describes some normalization methods, the third identifies poor estimation of the population probability of variables as a factor that compromises the effectiveness of the normalization methods for very short documents, and the fourth proposes elimination of data matrix rows representing document which are too short to be reliably normalized and suggests ways of identifying those documents

2. Variation in document length: the problem

Documents in collections can and often do vary considerably in length. Where the data abstracted from such a collection is based on the frequency of some textual feature or features of interest, such length variation is a problem for exploratory multivariate analysis. The nature of the problem is exemplified using the small document collection C comprising excerpts of various lengths from historical English texts ranging from Old English to Early Modern English, shown in Figure 1.

Name	Date	Size
Sermo Lupi ad Anglos	c.1000 CE	13 kb
Beowulf	c.1000 CE	106 kb
Apollonius of Tyre	c.1000 CE	35 kb
The Owl and the Nightingale	c.1300 CE	10 kb
Chaucer, *Troilus & Criseyde*	c.1370 CE	123 kb
Malory, *Morte d'Arthur*	c.1470 CE	132 kb
Everyman	c.1500 CE	37 kb
Spenser, *Faerie Queene*	1590 CE	34 kb
King James Bible	1611 CE	11kb

Figure 1. Document collection C

1.1 Data creation

Prior to its standardization in the later 18th century, spelling in the British Isles varied considerably over time and place, reflecting on the one hand differences in phonetics, phonology and morphology at different stages of linguistic development, and on the other differences in spelling conventions. It should, therefore, be possible to categorize texts on the basis of their spelling and to correlate the resulting categorizations with chronology. The research question, therefore, is: can the documents in C be accurately categorized chronologically by their spelling?

Investigation of spelling is here based on the concept of the tuple, where a tuple is a sequence of symbols: *xx* is a pair, *xxx* a triple, *xxxx* a four-tuple, and so on. Given a collection containing *m* documents, compile a list of all letter tuples that occur in the texts. Assume that there are *n* such tuples. To each of the documents d_i in the collection (for $i = 1..m$) assign a vector of length *n* such that each vector element v_j (for $j = 1..n$) represents one of the *n* letter tuples. In each document d_i count the number of times each of the *n* letter tuples *j* occurs, and enter that frequency in the vector element v_j of the vector associated with d_i. The result is a set of vectors each of which is an occurrence frequency profile of letter tuples for one of the documents in the collection. These document profile vectors are stored as the rows of a matrix.

A letter-pair frequency matrix was abstracted from C using the foregoing procedure. 554 letter pairs were found, and since there are 9 documents, the result is a 9 x 554 matrix henceforth referred to M_C.

1.2 Exploratory multivariate analysis of M_C

From what is commonly known of the history of the English language and of spelling at various stages of its development, one expects exploratory analysis of M_C to produce no surprises: the Old English, Middle English, and Early Modern English texts will form clusters. This expectation is not fulfilled, however, as the hierarchical analysis [2] in Figure 2 shows.

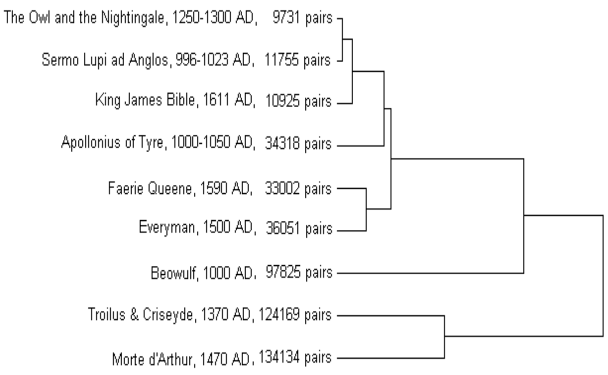
The Owl and the Nightingale, 1250-1300 AD, 9731 pairs
Sermo Lupi ad Anglos, 996-1023 AD, 11755 pairs
King James Bible, 1611 AD, 10925 pairs
Apollonius of Tyre, 1000-1050 AD, 34318 pairs
Faerie Queene, 1590 AD, 33002 pairs
Everyman, 1500 AD, 36051 pairs
Beowulf, 1000 AD, 97825 pairs
Troilus & Criseyde, 1370 AD, 124169 pairs
Morte d'Arthur, 1470 AD, 134134 pairs

Figure 2. Cluster tree of the rows of data matrix M_C

The texts do not group by chronological period, and the clustering in fact makes no obvious sense in terms of anything one knows about them and their historical context. When, however, one looks at the *Size* column in Figure 1, a correlation between cluster structure and document length is immediately clear. The texts have been grouped by their relative lengths: the short texts (*Owl, Sermo, King James*) comprise one cluster, the intermediate-length texts (*Apollonius, Faerie Queene, Everyman*) a second cluster, and the long texts (*Troilus, Morte d'Arthur*) a third, with *Beowulf* on its own commensurate with a length that falls between the intermediate-length and long texts.

1.3 Explanation of document length based clustering

When data has a vector representation, clustering by document length is explicable in terms of vector space geometry [3], in which the dimensionality *n* of the vector defines an *n*-dimensional space (here taken to be the familiar Euclidean one), the sequence of scalars comprising the vector specifies coordinates in the space, and the vector itself is a point at those coordinates. When two or more vectors exist in a space it is possible to measure the distance between them and thus to compare relative distances, so that *distance*(AB) in Figure 3a is greater than *distance*(AC). The length of a vector is the distance between itself and some reference point in the space's coordinate system; for present purposes that reference point is taken to be the origin of the coordinate axes. Like the distance between vectors, the relative lengths of vectors can be compared --in Figure 3b *length*(A) is greater than *length*(C).

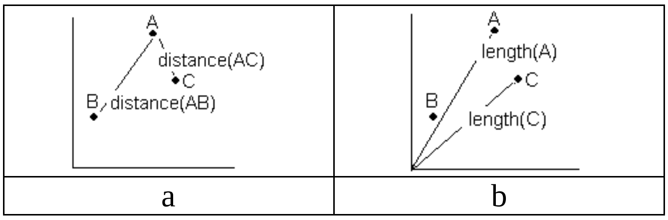

Figure 3: Distance and length in two-dimensional vector space

The distance between any two vectors in a space is jointly determined by the magnitude of the angle between the lines joining them to the origin of the space's coordinate system, and by the lengths of those lines.

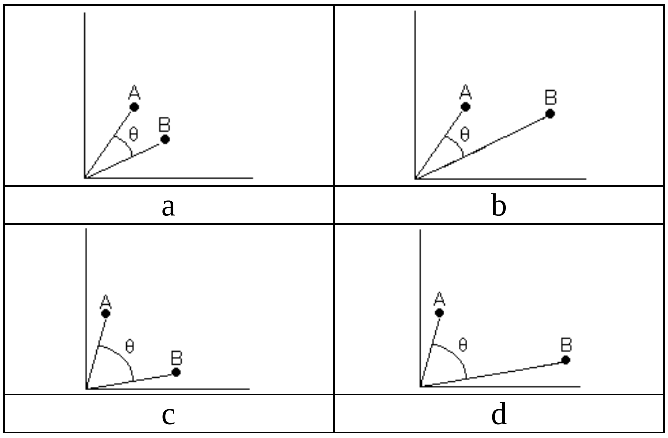

Figure 4: Relationship of vector angle and vector length to vector distance

Figure 4a shows two vectors A and B and an angle θ between them. In 4b θ remains the same and the length of B is increased, in 4c θ is increased and the vector lengths remain the same, and in 4d both the angle and the length of B are increased; in all cases (4b) - (4d) the distance between A and B increases commensurately.

It is easy to see that, as the angle decreases, length becomes increasingly dominant in determining distance. When, moreover, this observation is extended to more than two vectors, length becomes an increasingly important determinant of vector clustering in the space: where the angles between them are small, vectors of similar lengths cluster, as shown in Figure 5.

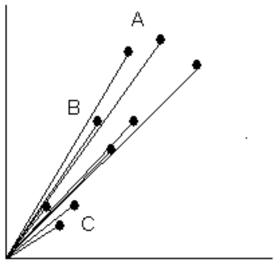

Figure 5: Clusters determined by vector length

And, because hierarchical cluster analysis groups vectors on the basis of their relative distances in space, vector length under these circumstances largely determines cluster analytical results.

This applies directly to the cluster analysis of M_C in that (i) the angles between its row vectors are relatively small, (ii) the vectors vary in length, and (iii) this length variation creates clusters in the data space. Because M_C is 554-dimensional there is no question of being to show this by plotting the row vectors directly as for the two-dimensional example in Figure 5. It is, however, possible to do so indirectly by projecting M_C into two-dimensional space using principal component analysis [6] and then plotting the rows of the projection matrix; the two largest principal components of M_C account for 70.7% of its variance, so the 9 x 2 projection matrix $M_{C(PCA)}$ is a reasonably accurate representation of

M_C. The scatter plot of the rows of $M_{C(PCA)}$ in Figure 6 shows that the angles between them are indeed relatively small and that they cluster by vector length.

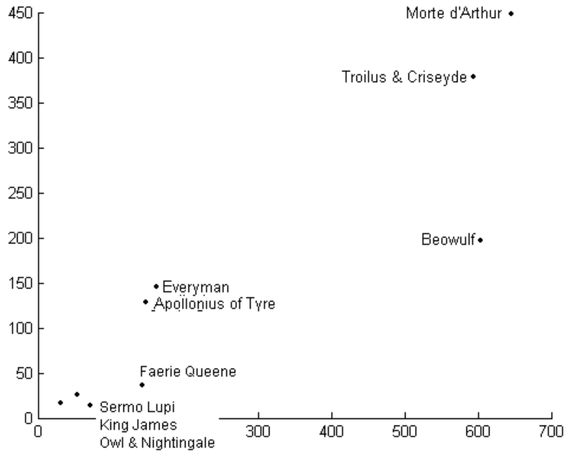

Figure 6: Scatter plot of the row vectors of $M_{C(PCA)}$

When, moreover, one observes that there is a near-linear relationship between the sizes of the documents in C (measured as the number of tuples in each) and the lengths of the vectors representing them in M_C (Figure 7), the reason for the length-based clustering of the documents in C becomes obvious.

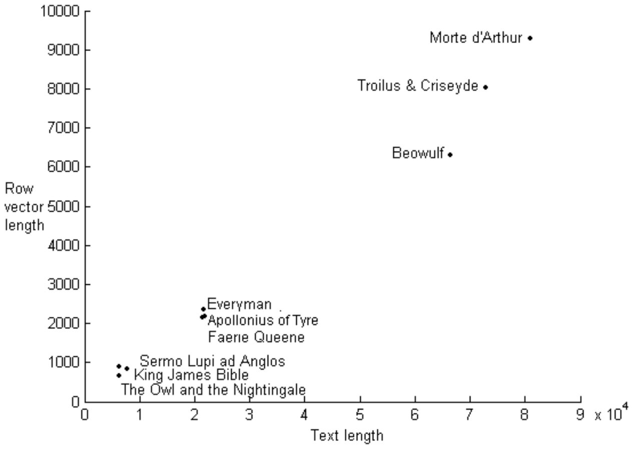

Figure 7: Plot of row vector lengths in M_C against the sizes of the corresponding documents in C

3. Document length normalization methods

Several ways of normalizing frequency data matrices abstracted from varying-length document collections have been proposed [1, 8, 9]. All of them work by dividing each of the values in each row of a frequency data matrix M by a constant: *k*:

$$M_{ij} = \sum_{i=1..m} \sum_{j=1..n} \left(\frac{M_{ij}}{k}\right)$$

This section mentions only two; subsequent discussion will show why an exhaustive list is unnecessary for present purposes.

- *Probability normalization*: For a given row M_i, *k* is the sum of frequencies in that row, that is, $k = \sum_j M_{ij}$ for j = 1..*n*. This replaces absolute frequency values in the matrix, whose magnitudes are dependent on document size, with probabilities, which are not; see further on the frequency-based definition of probability in Section 4 below.

- *Cosine normalization*: Any vector can be transformed so that it has length 1 by dividing it by its norm or length:

$$v_{unit} = \frac{v}{|v|}$$

In the present application v = M_i and $|M_i|$ = k. All row vectors in M are thereby made to lie on a hypersphere of radius 1 around the origin; because all vectors are equal in length, variation in the lengths of documents and, correspondingly, of the vectors that represent them cannot be a factor in analysis.

4. Effectiveness of normalization methods

M_C was normalized using the methods described in Section 3, and both the normalized matrices were cluster analyzed. In both cases the result was the same, and is shown in Figure 8.

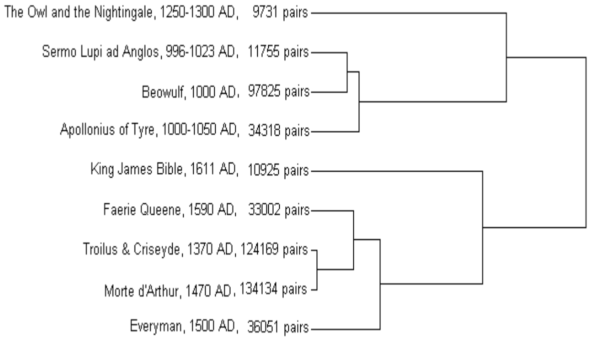
The Owl and the Nightingale, 1250-1300 AD, 9731 pairs
Sermo Lupi ad Anglos, 996-1023 AD, 11755 pairs
Beowulf, 1000 AD, 97825 pairs
Apollonius of Tyre, 1000-1050 AD, 34318 pairs
King James Bible, 1611 AD, 10925 pairs
Faerie Queene, 1590 AD, 33002 pairs
Troilus & Criseyde, 1370 AD, 124169 pairs
Morte d'Arthur, 1470 AD, 134134 pairs
Everyman, 1500 AD, 36051 pairs

Figure 8. Cluster analysis of length-normalized matrix M_C

The row vectors are now clustered by the chronological periods of the texts they represent, and make sense in terms of what is known of those texts in relation to the history of English. There are two main clusters. The upper one comprises a group of Old English texts and the single Early Middle English text irrespective of length variation. The lower one contains the later Middle English and the Early Modern English texts. Here, the most recent of the Early Modern texts, *King James*, is on its own; the *Faerie Queene*, though chonologically near to *King James*, is known deliberately to have archaized its spelling, and is thus classified with the Middle English texts.

For C, therefore, the conclusions are (i) that normalization solves the problem of variation in document length, and (ii) that the normalization methods referred to in Section 3 are equally effective. Can these conclusions be extended to document collections in general? The short answer with respect to (i) is 'no', and with respect to (ii) 'probably'; the remainder of this section deals mainly with (i), but (ii) is briefly addressed at the end.

When a frequency matrix is abstracted from a collection containing very short documents, normalization of the vectors representing those short documents is likely to be unreliable, which in turn leads to unreliable cluster analytical results. This stems from the unlikelihood of very short texts accurately estimating the population probabilities of data variables. Given a population E of *n* events, the frequency interpretation of probability [7, pp.1-17] says that the probability $p(e_i)$ of e_i ε E (for *i* in 1..*n*) is the ratio (*frequency*(e_i) / *n)*, that is, the proportion of the number of times e_i occurs relative to the total number of occurrences of events in E. A sample of E can be used to estimate $p(e_\mathrm{i})$, as is done with, for example, human populations in social surveys. The Law of Large Numbers [4, pp. 305-320] says that, as sample size increases, so does the likelihood that the sample estimate of an event's population probability is accurate; a small sample might give an accurate estimate but is less likely to do so than a larger one, and for this reason larger samples are preferred. It has already been pointed out that, where there is variation in document length and all occurrences of some feature are counted, the sum of frequencies for a vector representing a relatively longer document is necessarily greater in magnitude than the sum of frequencies for a vector representing a relatively shorter one. The shorter the document, therefore, the less accurate its estimate of the population probabilities can be expected to be.

To see the effect of this on cluster analysis, consider first a case where the population probabilities of the data variables are known, and a data matrix where the rows represent samples of increasing size and the sample variable values have been arranged so that all give perfect estimates of those probabilities (Figure 9).

	v1 p = .067	v2 p = .133	v3 p = .200	v4 p = .267	v5 p = .333
r1 (s=15)	1	2	3	4	5
r2 (s=30)	2	4	6	8	10
r3 (s=60)	4	8	12	16	20
r4 (s=120)	8	16	24	32	40
r5 (=240)	16	32	48	64	80
r6 (=480)	32	64	96	128	160
r7 (s=960)	64	128	192	256	320
r8 (=1920)	128	256	384	512	640

Figure 9. Data matrix showing population probabilities of variables

Figure 10 shows this matrix probability-normalized.

	v1	v2	v3	v4	v5
r1	.067	.133	.200	.267	.333
r2	.067	.133	.200	.267	.333
r3	.067	.133	.200	.267	.333
r4	.067	.133	.200	.267	.333
r5	.067	.133	.200	.267	.333
r6	.067	.133	.200	.267	.333
r7	.067	.133	.200	.267	.333
r8	.067	.133	.200	.267	.333

Figure 10. The matrix of Figure 10 probability-normalized

The matrices of Figures 9 and 10 were cluster analyzed, and the results are shown in Figure 11.

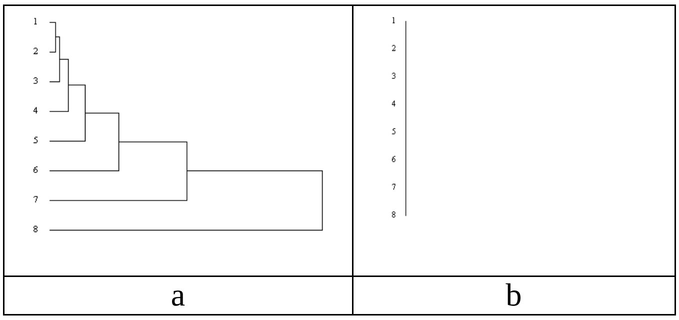

Figure 11. Cluster analyses of Figures 10 (a) and Figure 11 (b) matrices

Normalization has completely eliminated the variation in length which gives rise to the length-based clustering in Figure 11a and made the rows unclassifiable (11b), as the definition of probability normalization leads one to expect.

Now consider what happens with a matrix empirically derived from a collection of, say, 16 documents where accuracy of the population probability estimates cannot be guaranteed. For comparability with Figures 9 and 10, each document in the collection is twice as long as the preceding one, giving the same progression of relative sample lengths as in Figure 9. The documents are increasing-length excerpts from a randomly-selected text, Dickens' *Dombey & Son*, and the variables are again letter pairs: the first document contains the first 10 letter pairs in the text, the second the first 20 pairs, and so on up to the sixteenth at 327680 pairs. There are 560 letter-pair types, which yields a 16 x 560 frequency matrix M_{560}. Figure 12a shows a cluster analysis of M_{560}, and Figure 12b of the probability normalized matrix $M_{560(norm).}$

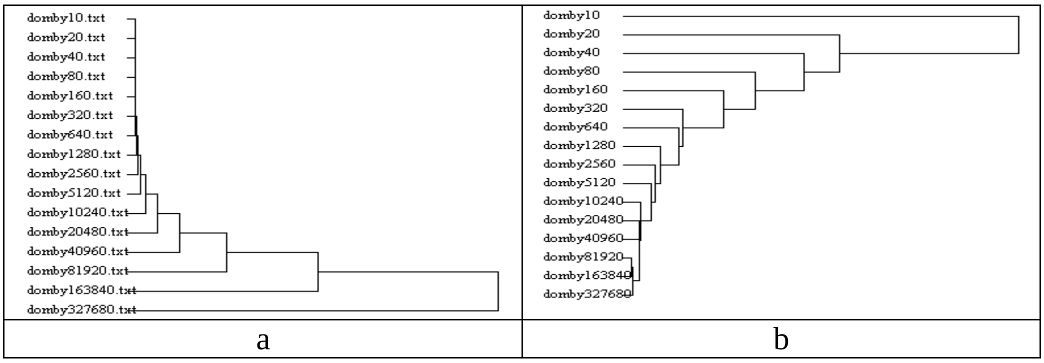

Figure 12. Euclidean distance / single link cluster analysis of M_{560} and $M_{560(norm)}$

Like Figure 11a, 12a shows length-based clustering. Unlike Figure 11b, however, 12b is not flat, that is, the matrix rows have not been normalized to uniform values. The reason for this emerges from an examination of the distributions of individual variable probabilities. Figure 13 shows the distributions for the three most frequent letter pairs in the collection, *th*, *in*, and *he*, across all 16 documents; the remaining columns are similar.

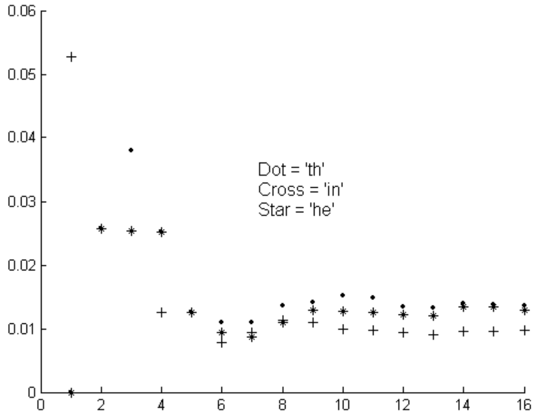

Figure 13. Probability distributions of the letter pairs *he*, *in*, and *th*

The horizontal axis represents the 16 documents with the shortest on the left and the vertical axis the population probability estimates for *he*, *th*, and *in*. In each distribution, the probabilities fluctuate for the shorter documents and then settle down to a much more restricted range of values corresponding to the increasingly-accurate estimate of the population probability as one moves to the longer documents on the right, which is what one expects from the Law of Large Numbers. The fluctuations on the left are caused by frequency values that are too large or too small relative to the length of the segment to estimate the population probability accurately. In other words, frequency values for variables in short texts can be and in the present instance are unreliable estimators of population probabilities.

Finally, it remains to note that the unreliability of normalization with respect to very short documents discussed above affects any method that divides row vector values by a constant, such as the cosine normalization mentioned in Section 3. These methods are all linear vector transformations, and, as such, affect the scaling of the row values but not their distribution.

5. Dealing with very short documents

The only obvious solution to the problem described in Section 4 is to identify which documents in a collection are too short to provide reasonably reliable estimates of population probabilities, and to eliminate the corresponding rows from the data matrix. But how short is too short?

One approach is to sort the row vectors of the unnormalized matrix in ascending order of document length and the column vectors in descending order of variable frequency, and then to create probability plots for the most frequent variables starting at the left of the matrix, as in Figure 13. Documents that are too short will show up as large vertical fluctuations; the point on the document axis where the fluctuations settle down is the required length threshold. In Figure 13, for example, documents 1-4 would be regarded as too short for inclusion in analysis, though one would want to examine more variables before drawing that conclusion.

The other approach is to determine how well the rows of the matrix have converged on some criterion and to remove those rows which have converged insufficiently well. Using the centroid vector of M_{560} as the criterion, the least squares distances of rows 1-16 of the normalized matrix $M_{560(norm)}$ from it were calculated and plotted in Figure 14; to compensate for differences in scaling between M_{560} and $M_{560(norm)}$ all vectors were converted to standard or z-scores prior to the distance calculation.

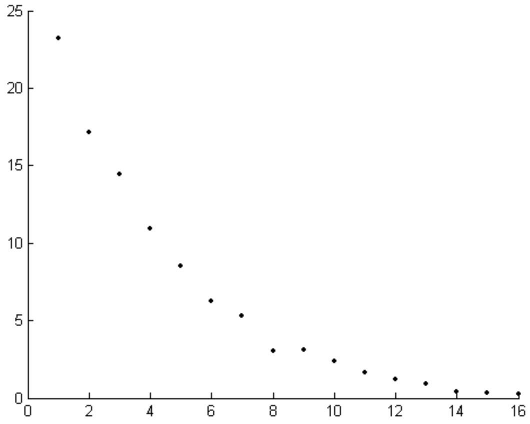

Figure 14. Least squares distances of row vectors of $M_{560(norm)}$ from the centroid vector of M_{560}

The indication is that the convergence is acceptable from vector 8 onwards, and that rows 1-7 representing the shorter documents should be removed from M_{560}.

With both approaches, the decision on where exactly the watershed should go is a matter of judgement, and will presumably be clearer in some applications than in others.

Conclusions

The discussion began with the observation that variation in the length of documents in electronic text corpora can be a problem for a range of interpretative technologies, and undertook to address that problem with reference to exploratory multivariate analysis of frequency data. The discussion was in four main parts. The first part stated the nature of the problem, the second described some normalization methods designed to mitigate or eliminate it, the third identified poor estimation of variable population probability as a factor that compromises the effectiveness of the normalization methods for very short documents, and the fourth proposed elimination of data matrix rows representing document which are too short to be reliably normalized and suggested ways of identifying those documents.

References

[1] C. Buckley. The importance of proper weighting methods. In: M. Bates (ed.). *Human Language Technology.* San Mateo, CA: Morgan Kaufmann, 1993.

[2] B. Everitt, S. Landau, M. Leese.. *Cluster Analysis,* 4th ed. London: Arnold., 2001.

[3] J. Fraleigh, R. Beauregard. *Linear Algebra*. 2nd ed. Menlo Park, CA: Addison-Wesley, 1995.

[4] C. Grinstead, J. Snell. *Introduction to Probability*, 2nd ed. American Mathematical Society, 1997.

[5] J. Hair, W. Black, B. Babin, R. Anderson, R. Tatham. *Multivariate Data Analysis*, 6th ed. New Jersey: Prentice-Hall, 2005.

[6] I. Jolliffe. *Principal Component Analysis*, 2nd ed. Berlin and Heidelberg: Springer Verlag, 2002.

[7] J. Milton, J. Arnold. *Introduction to Probability and Statistics*, 4th ed. Boston:McGraw-Hill, 2003.

[8] A. Singhal, G. Salton, M. Mitra, C. Buckley. Document Length Normalization. *Information Processing and Management* 32: 619-633, 1996.

[9] A. Singhal, C. Buckley, M. Mitra. Pivoted document length normalization. *Proceedings of the 19th ACM Conference on Research and Development in Information Retrieval (SIGIR-96)*, 21-29, 1996.

[10] B. Tabachnik, L. Fidell. *Using Multivariate Statistics,* 5th ed.. Boston: Allyn & Bacon, 2006.