Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Agglomerative Hierarchical Clustering: An Introduction to Essentials. (3) Standardization, Normalization, and Dimensionality Reduction of a Data Matrix

Refat Aljumily[1]

[1] University of Newcastle

Received: 6 December 2015 Accepted: 2 January 2016 Published: 15 January 2016

Abstract

In a previous tutorial article I looked at a proximity coefficient and, in the light of that proximity created a vectordistance matrix and used it to construct a hierarchical tree using different hierarchical clustering methods which will be the basis for exploratory multivariate analysis. The present article deals with three topics: (i) standardization for variable scales variation, (ii) normalization for sample length variation, and (iii) dimensionality reduction or minimization of data space. These techniques reflect the author?s academic background and particular area of interest and are, by necessity, not a particular purpose and are straightforwardly applicable to other kinds of data, and thus to a wide range of analysis in Linguistics. My treatment of these techniques is, necessarily, introductory and brief. I hope that this article will provide practitioners with an introductory overview of these techniques used for cluster analysis of electronic corpora of linguistic data. The assumption is that the data is in the form of an m x n matrix D in which, may require to transform it in various ways prior to cluster analyzing it. Standardized data matrix enables practitioners to measure the variation between n-variables and to cluster the cases they describe in common scales and values, regardless of their original scales and values. Normalized data matrix enables practitioners to eliminate the effect of variation in length among n-samples and to cluster them as if they were all (about) the same length, regardless of their original length. Dimensionality-reduced space data matrix enables practitioners to select and/or extract n-most interesting variables relevant to the research question and to visualize an existing pattern, regardless of the original space. A worked example is given to illustrate the effect each transformation technique has on a given data matrix. These transformation techniques have their own strengths and weakness but are beyond the scope of

Index terms— corpus, vector, matrix, standardization, coefficient of variation, normalization, dimensionality reduction.

1 Introduction

anguage corpus typically consists of more or less numerous texts each of which is described in terms of the selected linguistic features, technically known as variables. If it is to be analyzed using clustering methods, the selected variables need to be mathematically represented. A widely used way of doing this is vector space representation. Where vector space representation is used, each text is described by a vector, and the language corpus is consequently a set of vectors. Such a set of vectors is conveniently represented as a matrix in which the rows are the texts and the columns the linguistic features (variables). Thus, language corpus consisting of m

texts each of which is described by n variables is represented by an m x n matrix D in which D i (for i = 1...m) is the i'th text, D j (for j = 1..n) is the j'th variable, and D ij the value of variable j for text i. Once the language corpus has been constructed in a matrix, it is important to consider the issues relevant to cluster analysis of texts. Three types of issues are considered: (i) variable scales variation, (ii) text length variation, and (iii) variables selection/ extraction. This article proposes ways to remove the effect of each of these issues: (i) normalization for variation in text length, (ii) standardization for variation in variable scales, and (iii) dimensionality reduction. These techniques can be used, if it is necessary, to transform a given data matrix prior to analyzing it.

2 II.

3 Transformation Techniques a) Variation of variable scales

Almost any linguistic feature in a corpus such as word-forms, sentences, grammatical sequences, parts of speech, or any other easy to count features, can be measured. We use measurements to examine these linguistic features mathematically. In general, when we measure a linguistic feature, we define or interpret its properties in relation to special scales or units of measurement, then recording its happenings. That measurement constitutes the values of the linguistic features, for example: function words usage= 3000, average word-length=3, number of punctuation marks=500, diversity of words in a text 10%, and so on. Measurement is fundamental in the creation of language data because it makes a link between a particular linguistic feature in mind and an activity that originates from an individual, and thus allows the results of cluster some scale. Scales are systems designed to tell us how much of a measurable characteristic a given variable has. Scales have different types of numerical units and ranges (scales of measurements) appropriate to them which carry different amounts of information in any given application. The variables selected for describing linguistic features involving cluster analysis may require measurement on different scalars. If variables are measured on different scales, variables with large values contribute more to the distance measure than variables with small values.

Given an m x n data matrix M in which the m rows represent the m objects to be clustered, the n columns represent the n variables, and the entry at Mij (for i = 1..m, j = 1..n) represents a numerical measure of object i in terms of variable j, a clustering method has no idea what the values in the data matrix mean and calculates the degrees of similarity: variables that are measured in large values will have a greater influence on the degrees of similarity between the objects than those variables measured in smaller values, and, therefore, will affect the reliability of the cluster analysis. To see this, take a look at the following data matrix which describes nine students (A, B, C, D, E, F, G, H, I) in terms of their use of three linguistic features in the academic papers, one of which represents the total number of contractions, another one function word/content word ratio, and a third function words frequency. In Table/1the first column variable represents the total number of contractions, the second FW/CW ratio in percentage, and the third FW in frequency. A hierarchical cluster analysis of the matrix rows using Squared Euclidean distance gives the following dendrogram: In other words, the clustering analysis didn't find any significant clusters; there is a clear and very strong tendency to cluster by scale of measurement. The essence of the problem now is that we need a clustering structure that reveals the proximities among the vectors independent of the variation in scaling.However, there are many standardization methods as a technique for removing the effect variation in scaling among data and making each variable receives equal contribution in the cluster analysis. Some of these methods are:

? Standard or Z-score standardization method.
? Standardization method based on variable mean.
? Standardization method based on variable sum.
? Cosine standardization method.
? Max standardization method.
? Range standardization method.

One of the reasons for this diversity is that different standardization methods are required for different purposes; for clustering or for other purposes. No one single standardization method will be suitable for all applications. Some methods can be extremely useful even if they are mathematically limited. Other methods bring different benefits, although some bring disadvantages as well. To be suitable for cluster analysis, however, a method must preserve differences in variability among variables, thereby giving a true account of the intrinsic cluster structure of the unstandardized data matrix. The emphasis is the degree to which a method preserves the pre-standardization absolute magnitudes of variability. By the intrinsic variability, we mean the amount of variability in the values of a variable expressed independently of the scale of those values and measured in statistics by the coefficient of variation, which is defined with respect to a variable v as the ratio of v's standard deviation to its mean, and by the absolute magnitude of variability we mean the amount of variation in the values of a variable expressed in terms of the scale of those values, and is measured by the standard deviation.

A standardization method based on variable means does this in the sense that it has the effect of preserving intrinsic variability in the values of a variable, and it does that in the following way: individual numerical column vectors of unstandardized data matrix can be standardized in relation to their mean, where the value of a given numerical column vector V in the unstandardized matrix must be divided by the mean μV of column vectors: V i std = V i / μV Where:

? V i std is a standardized column vector in a data matrix, for i= 1?number of rows in matrix or, equivalently, the number of text files in a corpus.

? Vi is an unnormalized document vector, for i as above.

? µV is the column vector mean, or scalar, measured by the total number of values in each column vector.

To illustrate this, the first three students described by the total number of contractions, FW/CW ratio (in percentage), and FW (in frequency), in the data matrix of Table/1 are recalculated. In Table/2, it is clear that MEAN-standardization has made the variation magnitudes comparable and also has preserved the coefficients of variation of the unstandardized variables. This is because division by a scalar, here the column vector mean, is a linear operation that alters the scale while preserving the shape of the original value distribution. It is also clear that the standard deviations of contractions, FW/CW ratio, and FW in Table 1b are identical to the corresponding coefficients of variation. This is because, for any data vector (here representing persons), it is always the case that its coefficient of variation is identical to the standard deviation of the MEANstandardized version of vector. After standardizing the variables for the remaining persons as above, the application of a hierarchical method on the standardized data matrix in Table 1b shows sufficiently accurate clustering; the hierarchical tree in Figure/2 differs substantially, and it clusters the nine students according to the relative magnitude of values in the matrix columns, i.e. regardless of the variation in the variable scales.

4 b) Normalization for variation in sample length

A corpus is a collection of texts collected with a particular linguistic research project. Very often, it happens that a corpus contains texts of varying sizes; many of them can be disparate in length and not at all identical with each other. If the disparity varies greatly from text to text, a critical issue arises that must be taken into account: the data abstracted from the corpus In this figure, there is a progression from the shortest texts at the top of the tree to the longest at the bottom and this means that there is a clear and very strong tendency to cluster by length. This can easily be seen from the number to the right of each of the text names which represents the number of words in the text. The reason for this is that, in the present example, the data abstracted from a corpus is based on frequency; each vector contains frequencies of lexical types for one of the texts, and a set of vectors are stored as the rows of the data matrix. In this sense, variations in the row vector lengths are simply a result of variations in magnitudes of lexical frequencies stored on the data matrix row vectors. To understand this, assume counting the number of occurrences of some lexical type j in a corpus containing two texts, A and B. Assume that j occurs 10 times equally across those two texts. After entering the lexical frequencies into data matrix row vectors, the interpretation would obviously suggest that on the basis of their usage of j, the two texts A and B are identical and that j apparently fails to discriminate between text A from text B. If, however, one knows that text A is 5000 words long and text B 500 words long, this is no longer the case. It is clear that, although both texts have the same frequency of occurrences of j, its significance level in them is significantly different from each other. The lexical type j is relatively infrequent in text A and relatively frequent in text B and therefore this difference can be used to differentiate between those texts. If we assume again that the text B is 50000 words long instead of 500, based on its observed frequency in 500 words, then there would have been 1000 occurrences of j. In short, the longer a text, the more likely in general a given word with a specific probability of occurrence is to occur in it, and, if it occurs, the higher the frequency of occurrence is in general likely to be. These different text lengths, called variations in lengths, are inherent in all texts in collections and result in variations in the frequencies stored in the data matrix. The variation may be large or very small, but it is always present. For the cluster analysis to be accurate and reliable, weighting to compensate for variation in text length is therefore necessary to remove this effect. The common way to do so is to adjust the data matrix so that not just frequency but its significance relative to text length can be represented and thus incorporated into subsequent analysis. There are a number of normalization methods that are theoretically motivated, for example:? cosine normalization ? probability normalization

? normalization by mean term frequency within document

? normalization by maximum term frequency within document

? normalization by mean document length across collection

? normalization by maximum document length across collection.

but, the one most easy to understand is normalization by the mean document length across collection, and the reminder of discussion will concentrate on that. In this method, to adjust the lengths of each row vector of an m × n data matrix of lexical types frequencies, the frequency count for a given lexical type in a given text must be multiplied by the mean length of all texts then divided by the total number of frequency counts occurring in that text. The effect of this process: decreasing the values in the vectors that represent long texts, increasing them in vectors that represent short ones, and, for texts that are near or at the mean, to change the corresponding vectors little or not at all. This can be expressed as: where X here in relation to mean length of texts in a corpus:?? ? ?? = ?? ?? µ ?????????? ? ??

? X'i is the normalized frequency of i' th lexical type in a row vector, for i=1??.n.

? Xi is unnormalized frequency of i' th lexical type in a row vector.

? µ is the mean length of vectors across all texts (T). This obtained by dividing the sum of frequencies of matrix row vectors (T) by that of the number of texts n, for i=1?..n:µ(T) = ? ?? = 1 ? ?? ??????????? ?? ?? ?? ????

? Length (i) is the sum of frequencies of any row vector (i).

For example, let M below be a matrix having The effect of the normalization method on the data matrix shown in this example above is clear: all the values in txt.a have been substantially increased because it is significantly shorter than the mean text length: length-500 <1476 (the mean). For txt.b, the values have been slightly decreased because it is slightly longer than the average document length: length-1500 >1476. Finally, the values for txt.c have been substantially decreased because it is significantly longer than the average document length: 2430> 1476.

In summary, normalization enables us to cluster and compare texts with each other irrespective of their lengths and failure to normalize for variation in text length can produce fundamentally erroneous cluster analytical results. Nevertheless, the process of normalizing data matrix column or row vectors itself has some unresolved problems and these problems are not discussed here. More on document length normalization can be found in, e.g., **??**Moisl,

5 c) Dimensionality reduction

Dimensionality is a major issue for data analysis in any given application. Where the aim is to generate a matrix M in which the rows are the data points, the column variables are lexical types, and the value at any given matrix location M ij is the frequency of lexical type j in i, dimensionality has a particular relevance to the application of cluster analysis. In dealing with highdimensional data, however, having too much is rarely a problem. Quite the opposite –the usual situation with high-dimensional data is that there is far too little. Highdimensional spaces are inherently sparse, and, to achieve adequate definition of the data manifold, the amount of data required very rapidly becomes intractably large; this phenomenon was described as the 'curse of dimensionality' by Bellman **??**1961]. The solution is that data dimensionality should be kept as low as possible consistent with the need to describe the particular research project adequately. Dimensionality reduction is the process of reducing the number of redundant variables under consideration, and can be divided into two major types: variable selection and variable extraction.

i

6 . Variable selection methods

Variable selection methods try to identify a subset of the more important user-defined variables and to remove the remainder from the analysis (given some definition of importance) without losing too much information, thereby achieving dimensionality reduction. Given that variable selection methods aim to select a subset of the more important variables, a well-defined criterion of importance is fundamental. Two of the most often used ones in the literature are variable selection based on frequency and variable selection based on variance, and these are briefly described below. Others, such as variable selection based on term frequencyinverse document frequency (TF-IDF) and measures of nonrandomness, are also available, but these give results similar to those based on frequency and variance, and the additional complexity associated with them is therefore felt not to justify their inclusion; for further information on these see [e.g. Moisl, 2015;Belew, 2000; **??**alton & McGill, 1983; **??**obertson, 2004].

7 a. Variable selection based on frequency

Frequency is the simplest criterion for selecting features from a data matrix: those variables which occur most often in the research domain -in the present domain, words in text -are judged to be the most important, and lost which occur least often are taken to be least important and can therefore be discarded. With respect to clustering, the fundamental idea is that a variable should represent something which occurs often enough for it to make a significant contribution to the clustering of the data vectors. To select variables based on frequency, given an m x n frequency data matrix D; the value at Dij is the number of times variable j, for j=1?n, occurs in text i, for i=1?m. The frequency of occurrence of variable j across the entire corpus of texts is then:

Frequencies of for all the columns data matrix D are calculated, sorted the variables in descending order of frequency, the most useful variables are selected and the less frequent variables are eliminated from D. Substantial dimensionality reduction can be achieved by applying this criterion to a data matrix D.

b

8 . Variable selection based on variance

Variability refers to the amount of variation in the values that a variable takes. Any variable x is an interpretation of some aspect of the physical world, and a value assigned to x is a measurement of the world in terms of that interpretation. If x is to describe the ages of people, it can take different values for different persons or for the same person at different times. Unless all people are exactly the same age, or the age of the same person is fixed, the values which x takes will vary substantially, and can, therefore, contribute to the distinction of people from one another, or of the age of same person at different times (i.e. the more different people groups one tests, the more variation one will see in the ages). This possibility of variability in the values assigned to variable x gives it its descriptive utility: an identical value for x tells that what x stands for in the real world does not change, moderate variability in the value tells that aspect of the world changes only a little, and widely differing values tells that it changes substantially. In general, therefore, the possibility of variability in the values assigned to

variables is necessary to the ability of variables to describe objects and thereby to represent reality. Clustering of texts or of anything else depends on there being variability in their characteristics; identical texts having the same stylistic descriptors cannot be meaningfully clustered. When the texts to be clustered are described by variables, then the variables are only useful for the purpose if there is significant variation in the values that they take. If, for example, a large number of people were described by their weights or heights, we would expect there to be logically substantial variation in values for each of them, and any cluster analysis method could legitimately be used to cluster them. On the other hand, if a large number of people were described by variables like 'eyes', 'noses', and 'legs', there would be almost no or little variation or high correlation with other features, since, with very few exceptions, everyone has two eyes and a nose, and clustering based on these variables would be effectively useless. In any clustering application, therefore, one is looking for variables with substantial variation in their values, and can ignore variables with little or no variation. Variables with no or little variation should be removed from data matrix as they contain little information and complicate cluster analysis by making the data higher-dimensionality than it needs to be [Moisl, 2015].

Mathematically, the degree of variation in the values of a variable is described by its variance. The variance of a set of variable values is the average deviation of those values from their mean. Assume a set of n values {x 1 , x 2 ...x n } assigned to a variable x. The mean of these values µ is (x 1 + x 2 + ... + x n)/n. The amount by which any given value x i differs from µ is then x i -µ. The mean difference from µ across all values is therefore ? i=1..n (x i -µ)/n. This mean difference of variable values from their mean almost but not quite corresponds to the definition of variance. One more step is necessary, and it is technical rather than conceptual. Because µ is an average, some of the variable values will be greater than µ, and some will be less. Consequently, some of the differences (x i -µ) will be positive and some negative. When all the (x i -µ) are added up, as above, they will cancel each other out. To prevent this, the (x i -µ) are squared. The standard definition of variance for n values {x 1 , x 2 ...x n } assigned to a variable x, therefore, is:

To show how a variance is calculated, consider the following frequency counts of six variables (the, a, she, him, then, him) occurring in the corresponding five texts Given a data matrix M in which the row vectors are the texts and the column vectors are lexical type variables describing the texts, and also that the aim is to cluster analyze these texts on the basis of the differences among them, the application of variance/standard deviation to dimensionality reduction is straightforward: calculate and plot the variances of the columns and, if any have variability which is low in relation to that of the others, remove them on the grounds that they contribute little to differentiation of the texts, and decide on a threshold selection (the set of retained variables from each column of the data matrix).

9 d) Variable extraction methods

Variable extraction methods replace the set of user-defined variables with a smaller set of variables which reduces dimensionality but captures most of the variability in the original set. These methods often achieve a greater degree of dimensionality reduction, but at a cost: the newly-defined variables are generated by mathematical procedures, and their meaning relative to the research domain is typically difficult to determine reliably. There are a wide of variable extraction methods:

? Singular value decomposition (SVD)

? Principal Components Analysis (PCA)

? Factor Analysis (FA)? Multi-dimensional Scaling (MDS) ? Isomap ? Self-Organizing Map (SOM)

Each one of these methods can be used for dimensionality reduction as a feature or variable extractor, and to visualize the clusters as a clustering method. The literature on these methods is extensive and this is just a brief outline that one can follow. A more comprehensive account can be found in, for example, [Moisl, 2015;Borg and Groenen, 2005;Kohonen, 2001; **??**enenbaum, de Silva, and Langford, 2000;Gordon, 1999]. However, it will be useful to look briefly at one of these methods, that is, PCA, as a dimensionality reduction method, to see how it reduces the data down into basic components, removing any unnecessary variables.

Principal Components Analysis (PCA)is actually a dimensionality reduction method, which aims to transform a set of correlated variables into a –usually smaller–set of uncorrelated ones. PCA can also be used for clustering if the dimensionality is sufficiently reduced. The conceptual basis of PCA is elimination of variable redundancy. Specifically, given a matrix of m data items described by n variables, principal components analysis is a technique for redescribing the m items in terms of k variables, where k <n, such that most of the variability in the original n variables is retained. When k = 2 or k = 3 the m data items can be plotted in two or three dimensional space and any clusters can thereby be directly perceived. Relative to an n-dimensional data set D, the essence of PCA is this: ? An n-dimensional orthogonal basis for D is constructed, such that each axis is the leastsquares best fit to one of the n directions of variation in D. ? The axes along which there is relatively little variation are eliminated, leaving an m-dimensional basis for D, where m<n. ? The original n-dimensional data D is projected into the reduced m-dimensional space, which yields a data set D' that is dimensionality-reduced but still contains most of the variability in D.

10 III.

11 Conclusion

In this article, I discussed three techniques to adjust a data matrix before applying cluster analytical methods to take account of the variation in scales among the variables, the variation in length among the texts, and any superfluous variables in it using standardization, normalization, and dimensionality reduction techniques. A full and detailed consideration of each of these techniques addressed in this article would require several articles. My treatment of them is, necessarily, introductory and brief. Therefore, I urge interested computational linguists to follow the more in depth sources cited in the references. The application of these techniques for cluster analysis with specific reference to corpus linguistics is only one of many possibilities. The data items/matrix rows might be students in a second language learning (L2) survey and the variable/matrix columns motivational factors like learning experience, attitudes, cultural interest, and so on. n formants in a sociolinguistic or dialectological survey and the variables/matrix columns phonetic features like voicing, and so on. The lexical frequency example was selected because it is generic with respect to a wide range of possible applications.

IV. [1]

1

Figure 1: Figure 1 :

2

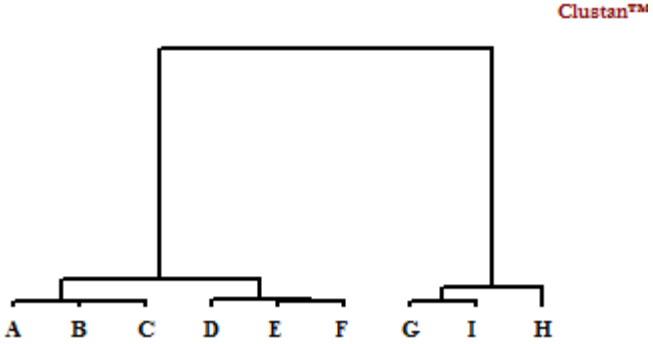

Figure 2: Figure 2 :

[1] © 2016 Global Journals Inc. (US)

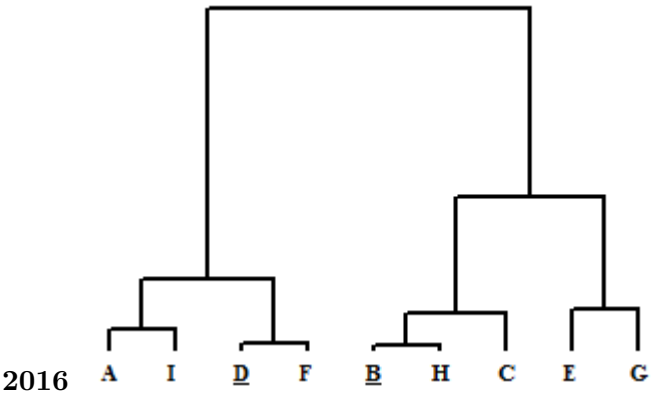

Figure 3: Year 2016 Agglomerative

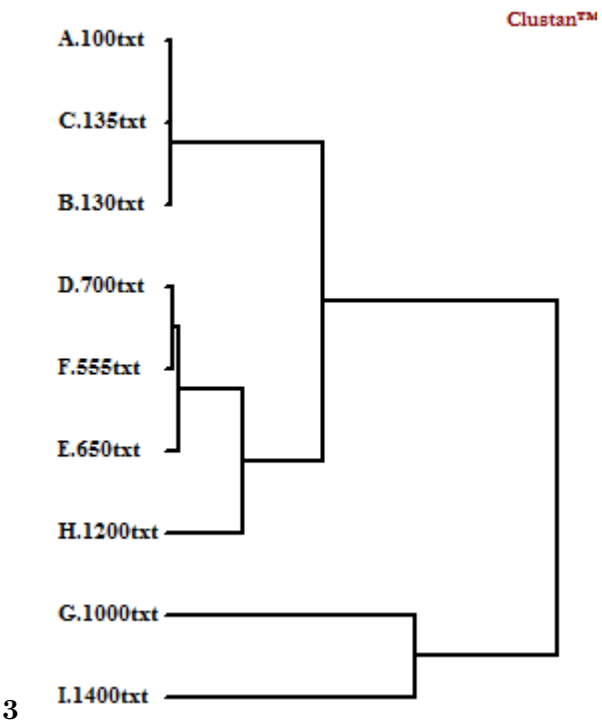

Figure 4: Figure 3 :

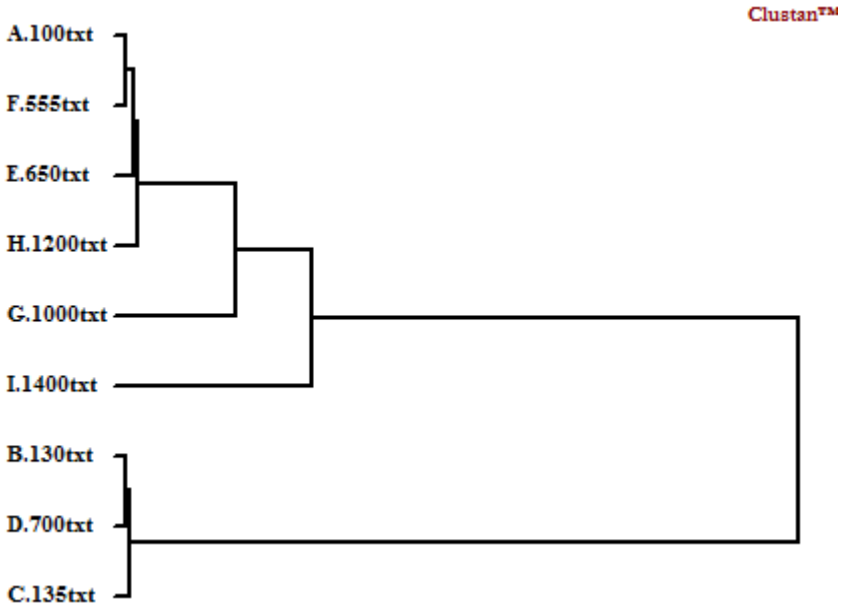

Figure 5:

$$freq(F_j) = \sum_{i=1..m} F_{i,j}$$

Figure 6:

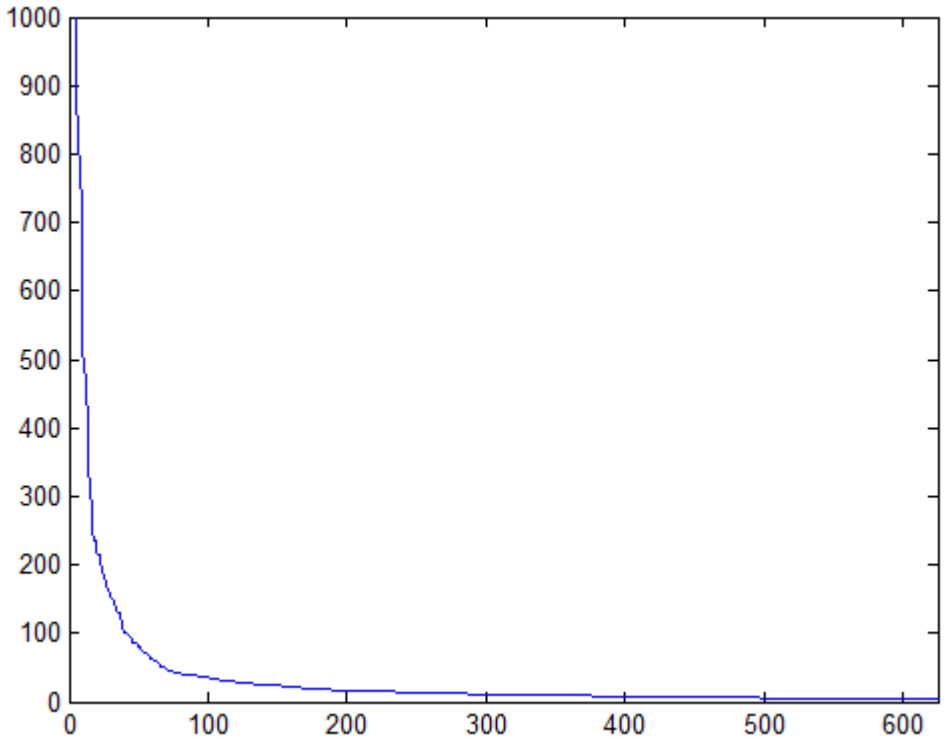

Figure 7:

1

Students	Number of contractions	FW/CW (percentage)	FW (frequency)
A	187	40	27000
B	185	35	25000
C	184	33	26000
D	170	29	23500
E	166	25	22000
F	164	26	21000
G	160	60	15000
H	150	53	10000
I	159	61	14500

Figure 8: Table 1 :

2

students	Contraction	FW/CW	FW	Contraction	FW/CW	FW
A	187	40	27000	1.01	1.11	1.03
B	185	35	25000	1	0.97	0.96
C	184	33	26000	0.99	0.91	1
Std	1.247	2.943	816.496	0.084	0.022	0.028
CV	0.006	0.081	0.0314	0.084	0.022	0.028

[Note: a. unSTD matrix of Table(1) b. Mean STD matrix of Table(1)]

Figure 9: Table 2 :

.1 Acknowledgments

The author wishes to thank all those who dedicated their time answering my queries and providing me with valuable comments during the preparation of this study.

.2 Conflicts of Interest

The author declares no conflict of interest.

[Tenenbaum and Langford ()] 'A global geometric framework for nonlinear dimensionality reduction'. J Tenenbaum , V , J Langford . *Science* 2000. 290 p. .

[Milligan and Cooper ()] 'An examination of procedures for determining the number of clusters in a data set'. G Milligan , Cooper . *Psychometrika* 1985. 50 p. .

[Priddy and Keller ()] *Artificial Neural Networks: An Introduction*, K L Priddy , P E Keller . 2005. USA: Spie Press.

[Belew ()] R Belew . *Finding Out About: A Cognitive Perspective in Search Engine Technology and the WWW*, (Cambridge) 2000. Cambridge University Press.

[Borg and Groenen ()] I Borg , P Groenen . *Modern Multidimensional Scaling. 2 nd*, (Berlin) 2005. Springer.

[Moisl ()] *Cluster Analysis for Corpus Linguistics*, Hermann Moisl . 2015. Berlin: De Gruyter Mouton.

[Gordon et al. ()] 'Data clustering: a review'. A Gordon , Chapman , A Halljain , M Murty , P Flynn . *ACM Computing Surveys* 1999. 1999. 31 p. . (Classification 2 nd)

[Singhal et al. ()] 'Document Length Normalization'. A Singhal , G Salton , C Buckley . *Information Processing and Management* 1995. 32 p. .

[Chu et al. ()] 'Effect of data standardization on chemical clustering and similarity searching'. C Chu , J Holliday , P Willett . *Journal of Chemical Information and Modeling* 2009. 49 p. .

[Dy and Bodley ()] 'Feature selection for unsupervised learning'. J Dy , C Bodley . *Journal of Machine Learning Research* 2004. 5 p. .

[Dy (ed.) ()] *In: Computational Methods of Feature Selection*, J Dy . H. Liu and H. Motada. London: Chapman and Hall CRC (ed.) 2008. p. . (Unsupervised feature selection)

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

[Kohonen ()] *Self-Organizing Maps*, T Kohonen . 2001. Berlin: Springer. (3rd ed)

[Gnanandesikan et al. ()] 'Weighting and selection of variables for cluster analysis'. R Gnanandesikan , S Tsao , J Kettenring . *Journal of Classification* 1995. 12 p. .