International Journal of Database Theory and Application
Vol.9, No.8 (2016), pp.89-98
http://dx.doi.org/10.14257/ijdta.2016.9.8.10

K-Means Clustering of Shakespeare Sonnets with Selected Features

T. Senthil Selvi[1] and R. Parimala[2]

[1]*Research Scholar, PG & Research Department of Computer Science, Periyar E. V. R. College, Tiruchirapalli -23*
[2]*Research Adviser, Assistant Professor, PG & Research Department of Computer Science, Periyar E. V. R. College, Tiruchirapalli-23*
[1]*senthilselvikumar@yahoo.co.in*, [2]*rajamohanparimala@gmail.com*

Abstract

This paper focuses on clustering the lines of Shakespeare Sonnets. Sonnet Line Clustering (SLC) is the task of grouping a set of lines in such a way that lines in the same cluster are more similar to each other than to those in other clusters. K-Means clustering is a very effective clustering technique well known for its observed speed and its simplicity. Its aim is to find the best division of N lines into K groups (clusters), so that the total distance between the groups's members and corresponding centroid, is minimized. A new algorithm Sonnet Line Clustering with Random Feature Selection SLCRFS is proposed. To validate the process external validation or internal validation is done. Since, internal validation has no considerable impact in conducting research this work concentrates on the measures of external validation. Entropy and Purity are frequently used external measures of validation for K-Means. The proposed approach uses entropy as performance measure. The clusters formed are evaluated and interpreted according to the Euclidean measure between data points and cluster centers of each cluster. This paper concludes with an analysis of the results of using the proposed measure to display the clustered sonnets using K-Means algorithm with minimum entropy for different feature sets.

Keywords*: K-means Clustering, Entropy, Feature selection*

1. Introduction

Sonnets Line Clustering (SLC) can be defined as a process of organizing Sonnet lines into groups where certain group members are similar in one way and certain groups dissimilar in some other way. Clustering has a long and rich history in the field of data mining. One of the most popular and simplest clustering algorithms is the K-Means. In spite of the fact that K-Means was proposed over 50 years ago and thousands of clustering algorithms have been published since then, still K-Means is widely used algorithm and has its own stand. Hartigan introduced the K-Means algorithm where the algorithm proposed repeatedly picking a point and determining its optimal cluster assignment [10]. It starts with a random initial point and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centers until a convergence criterion is met [13]. Julie proposed that the method was useful in finding patterns and can be applied in different types of domains such as Biology, Zoology, Medicine, Psychiatry, Sociology, Criminology, Geology, Geography, Remote sensing, Pattern recognition, Marketing research and Education [15]. Recently, however researchers have begun to study the Sonnets using computational methods [11]. Hieatt studied Shakespeare's plays and his Sonnets and searched for the frequency occurrence of rarely used words and also found the links between different groups of the Sonnet [4].

ISSN: 2005-4270 IJDTA

Today's organizations are moving toward the tremendous growth of unstructured data. The Shakespeare's Sonnet is an unstructured data, comprising of 154 Sonnets of 14-line poem that rhymes in a particular pattern. To create the feature set, the 154 Sonnet collection was taken and a term-document matrix was built. The number of terms found was large. Perform preprocessing in order to remove unwanted terms before the clustering process is done. The Preprocessing steps include stopword removal, punctuation removal, and number removal. Finally stemming is performed and additional unwanted whitespaces are removed. This reduces the corpus size greatly which is considered for clustering process. Accuracy and efficiency of clustering algorithms depends greatly on the input data. Removing unimportant features from the dataset can help us to form better clusters in lesser time. Therefore, it is essential to have a proper feature selection in order to reduce the sparseness. Sparse term removal at different threshold is proposed for feature selection. In addition, the proposed work randomly selects a subset of features that best represent the entire dataset and finally K-Means clustering is performed using rest of the selected terms. In this study, K-Means algorithm is applied to generate different clusters for different runs. The rest of the paper is organized as follows: Section 2 outlines about literature review. Section 3 presents the proposed methodology and performance measure Section 4 presents a detail about Dataset used. Section 5 shows the detail about the R Environment and Libraries. Section 6 analyses the experimental results and finally the conclusion is drawn.

2. Literature Review

Several methods have been proposed to solve the clustering problem. The K-Means algorithm is one of the partition clustering method [13]. In 1967 Mac Queen developed the simplest and the easiest clustering algorithm – the K-Means clustering algorithm. Bhoomi proposed that before the K-Means converges, the centroids are computed and all points are assigned to their nearest centroids [3]. Aljumily showed that the Function words, word bi-grams and character tri-grams plays role in finding the authorship style especially which distinguishes between Shakespeare and the other authors and determine the dissimilarity relations using clustering analysis. The disputed plays traditionally attributed to Shakespeare are not mathematically similar to any other of his works and, thus concludes that Shakespeare did not write them. Cluster analysis shows that Function words "and" and "to", word bi-grams "but that" and "that by", and character tri-grams "tur" and "nev" are the most important authorship style discriminators that distinguish between Shakespeare and the others and determine the dissimilarity/similarity relations among the texts examined [2]. The difference between hierarchical clustering and partition clustering algorithm have been analyzed of which the result of partition clustering of K-Means algorithm has higher efficiency and converges fast when dealing with large data sets [20]. Rakesh Chandra Balabantaray the author ascertains that the best cluster is obtained using K-Means algorithm which can be used for multi-document summarization and later can retrieve the document needed for access [22]. There has been significant work on characterizing rhyme, [5] poetry generation, case based reasoning to induce the best poetic structure [12], rhyme identification [18, 26] and also visualization of poetry[8]. Okafor stated that Entropy is a good measure for determining the quality of clustering [21]. He proposes the following measures: good data span and coverage. A dimensional space that has well defined clusters will tend to have good data span than one that is closed to random, high density; whereas for coverage two distributions can have the same data span, one may be denser and therefore qualify as a cluster. Given these criteria, a reduced dimension with good clustering should score high on this metric at some level of a threshold.

3. Methodology

3.1. Procedure for Sonnet Line Clustering

The goal of the method is to find clusters and assign labels to the objects based on the cluster that they belong to. K-Means clustering algorithm is an unsupervised algorithm and it is used to cluster Shakespeare Sonnets lines. But before applying *K* Means algorithm data is normalized using Preprocessing techniques. The aim of the K-Means algorithm is to divide M points in N dimensions into K clusters so that the sum of squares is minimized within-clusters and also find minimum entropy. A partition of a set of Sonnet's lines casually finds natural categories among objects by organizing data into clusters such that there is either high intra-cluster similarity or low inter-cluster similarity. SLC algorithm involves two stages of processing: Pre-processing and K-Means clustering as shown in Figure 1.

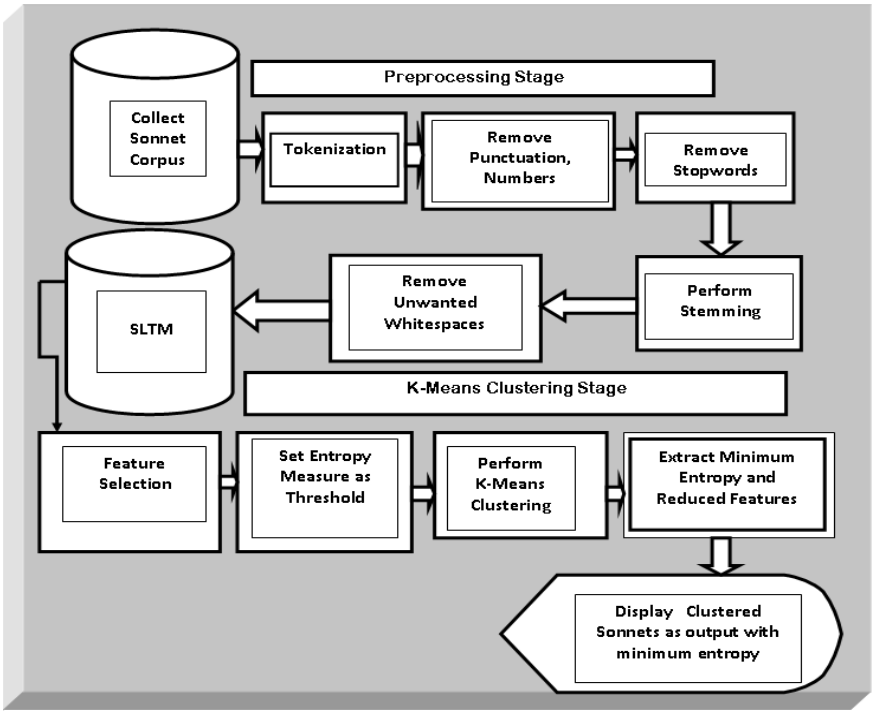

Figure 1. Two Stage Processing of SLC

Sonnet Line Corpus is in the form of unstructured text. Before the clustering process, each Sonnet lines are to be explored to extract its contents of words to be represented as terms. These are called as vector of terms or Bag of Words. Before building the term document matrix the preprocessing of Sonnet Line Corpus is to be achieved. First perform tokenization and then perform stopword removal, punctuation removal and number removal, finally perform stemming and remove all unwanted white spaces. Given an unstructured Sonnet Corpus O_i, $1 \le i \le N$, where N is the number of lines in the Sonnet collection being transformed to structured objects which results in Sonnet Line Term Matrix(SLTM). Sparse terms are removed from the SLTM. The second stage is the process of K-Means Clustering with feature selection. Feature selection is the minimum set of features that achieves maximum clustering performance. An efficient way of handling this is by selecting a subset of important features. Traditional feature selection algorithm works only for supervised data where class information is available. For unsupervised data, without class information, often Principal Components (PCs) are used. In this wok a random feature selection is obtained after performing sparse term removal.

A procedure for obtaining Sonnet line Clustering with feature selection is given in Algorithm 1.

Algorithm 1. Sonnet Line Clustering with Random Feature Selection (SLCRFS)

Begin
1. Read each lines of the Sonnet Corpus
2. Convert to VectorSource Corpus
3. Perform preprocessing steps:
Convert Corpus to Lowercase, remove punctuations, numbers and stopwords finally perform stemming. After preprocessing remove all unwanted white spaces.
4. Convert the Corpus to Sonnet Line Term Matrix (SLTM)
#Feature Selection
5. Perform sparse term removal for different sparsity from 98.5% to 99.9%
5.1. Initialize Each Object as 'M' and the number of features selected as NF
5.2. A uniform random number of sizes M chosen. A 0 or 1, at position i, indicates whether the feature i is selected (1) or not selected (0). $1 \leq i \leq M$
Perform K-Means on selected feature set and find its entropy measure.
6. Set Entropy measure to the maximum Threshold value.
7. Perform **retval** function for 100 iterations
Calculate minimum entropy value by calling **retval** function and return the optimum feature set.
8. Output the clustered output with entropy value (E) and reduced feature set (RF).
End

Function retval
Begin
1. Call **KMeans** (SLTM, K) # Calls Algorithm 2
2. Find Minimum Entropy(E) and Reduced feature Set(RF).
3. Return (E, RF)
End
End Function

3.2 K-Means Clustering Algorithm

The first step in K-Means clustering is to specify the number of clusters that will be formed in the final solution. The process begins by choosing n observations to serve as centers for the clusters. Then, the Euclidean distance is calculated for each of the K clusters and the observations are put in the cluster to which they are the closest and the center of the clusters is recalculated, and every observation is checked to see if it might be closer to a different cluster. If so record this observation and continue this process until convergent criteria is met. The process of K-Means clustering is shown in Algorithm 2.

Algorithm 2. The K-Means Clustering Algorithm

Input SLTM
For K #2,3,4,5,
Output: : A set of K clusters
Funtion KMeans(SLTM,K)
1. Arbitrarily choose K data-points from dataset SLTM as initial cluster centroids
2. Repeat

a. Calculate the distance between each data-points d_i (1<=i<=n) and all K cluster centers c_j (1<=j<=K) and assign each data item d_i to the nearest cluster (closest) centroid

b. For each cluster j, recalculate the cluster center.

U**ntil** No change in the center of clusters.

Return clusters

End Function

3.3. Performance Measure

Entropy and Purity are frequently used external measures of validation for K-Means. The proposed approach uses entropy as performance measure. Entropy is a sophisticated measure derived from the concept of information gain in the field of information theory developed by Claude Shannon in the 1940's [24]. The distance between various data points of the clusters generated by the algorithms is determined and analyzed. The clusters formed are evaluated and interpreted according to the distance between data points and cluster centers of each cluster. The determination of cluster quality is done by entropy measures. Entropy uses external information class labels in this case. The lower entropy means better clustering. So see to that every cluster should have low entropy to maintain the quality of clustering [23]. Let *CS* be a clustering solution. For each cluster, the class distribution of the data is calculated first, *i.e.*, for cluster *j* , compute p_{ij}, the "probability" that a member of cluster *j* belongs to class *i*. Then using this class distribution, the entropy of each cluster *j* is calculated using the standard formula

$$E_{ij} = -\sum_{i=1}^{L} p_{ij} \log (p_{ij}) \quad (1)$$

Where the sum is taken over all L classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster

$$E_{cs} = \sum_{j=1}^{K} \frac{K_j * E_j}{M} \quad (2)$$

where K_j is the size of cluster *j*, K is the number of clusters and M is the total number of data points [19].

4. Dataset Used

A collection of "Sonnets of Shakespeare", a Sonnet Corpus containing collection of poems from the Shakespearean era, which consists of 154 Sonnets each comprised of 14 lines. The Sonnets were collected for clustering at Gutenburg Website Shakespeare's copyright 1990-1993 by World Library, Inc., and is provided by project Gutenberg E-text of Illinois Benedictine College [28]. Recently, Scholars have begun to study the Sonnets using computational methods. In this study, A Sonnet Corpus consists of 2614 lines are considered for Clustering. Before preprocessing the size of the corpus was 470.9 KB. After preprocessing the size was reduced to 231.6 KB of storage.

5. Used Environment and Libraries

R is a programming language and software environment used for statistical computing and graphics [29]. Text Mining or Text Analytics applies analytic tools to learn any unstructured documents like books, newspapers, emails, *etc*. For the past few years "tm" package [27] in R Programming Language has gained vast interest from a variety of researchers and users of different backgrounds. R Language was used for implementation of this work.

6. Experimental Results and Discussion

A Sonnet Line clustering is high dimensional data, a standard benchmark dataset the Sonnet Corpus was taken. The unstructured corpus is transformed to structured corpus using bag of words. The unstructured Corpus has a collection of 2614 lines with 4311 terms. The size of SLTM initially was 2614 X 4311. So the dimension of the feature space is also enormous. Pre-processing steps like Tokenization, Stop word removal and Stemming were performed in SLTM and the size of SLTM was reduced to 2614 X 3058. Preprocessing Methods and its Corpus size is shown in Table 1.

Table 1. Preprocessing Methods and Corpus Size

Preprocessing Methods	Sonnet Corpus Size
Without preprocessing	470.9 KB
Tokenization + Punctuation Removal	269.5 KB
Tokenization + Stopword Removal	247.4 KB
Tokenization + Stopword Removal + Stemming + White Space Removal	231.6 KB

The high-dimensional and sparse features bring great noise to the Sonnet Line Clustering and make it difficult for clustering algorithms to effectively cluster similar lines. Sparse terms are removed from SLTM. The resultant SLTM is a object where those terms from SLTM are removed which have at least a sparse threshold percentage of empty. The K-Means clustering is applied for reduced corpus.The experiment is conducted for K=2, 3, 4 and 5. Sparse terms threshold percentage (STTP), initial number of features (NF), corpus size in Kilo Bytes (CSB), entropy measure (E) for K-Means clustering and its reduced features (RF) is summarized in Table 2.

Table 2. Entropy Measure and Reduced Features for K-Means with STTP

STTP	CSB	NF	ENTROPY (E) & REDUCED FEATURES(RF)							
			K=2		K=3		K=4		K=5	
			RF	E	RF	E	RF	E	RF	E
98.5	26.3	16	**8**	**0.0207**	9	0.0947	5	0.2035	3	0.1782
98.7	28.8	20	**8**	**0.0207**	8	0.1415	7	0.1827	3	0.2009
98.9	30.8	24	11	0.0609	12	0.1285	14	0.2248	15	0.2840
99.1	35.7	35	13	0.0522	15	0.0989	16	0.1633	16	0.2447
99.3	42.2	52	21	0.0430	24	0.1004	30	0.1942	25	0.2116
99.5	53.7	91	49	0.0333	47	0.0910	43	0.1942	**47**	**0.1563**
99.7	74.5	203	**102**	**0.0207**	91	0.0686	102	0.1392	93	0.1660
99.9	128.7	740	363	0.0229	**97**	**0.0458**	**386**	**0.1213**	356	0.1712

From Table 2, from the highlighted values it was observed that the minimum entropy for K=2 with 8 features is 0.0207, for K=3 with 97 features is 0.0458, for K=4 with 386 feature is 0.1213and for K=5 with 47 feature is 0.1563 also it is observed that for STTP =

98.5 with 8 features is 0.0207, for STTP =98.7 with 8 features is 0.0207 and for 99.7 with 102 features is 0.0207.

A sample of results for 99.1 % sparsity is given in Table 3. This table provides the list of initial features selected and features used for K-Means. The K-Means Clustering for K values of 2,3,4,5 or reduced feature set is shown in Figure 2.

Table 3. Sample Reduced Features for Sparse Value of 99.1%

For 99.1 % sparsity the Initial Features Selected is 35			
art beauty dost even every eye eyes fair give hath heart ill let like live love loves make might mine now one praise say see self shall since still sweet time true upon will world			
Final Features selected for various K-Means			
K=2 with 13 Features	K=3 with 15 Features	K=4 with 16 Features	K=5 with 16 Features
beauty yes fair heart ill live make now one since still time will	art dost every eye give ill let like make might praise see sweet time true	dost even eye fair give heart live loves might one praise say shall since time true	even eye eyes give heart let like love one praise say see self since time upon

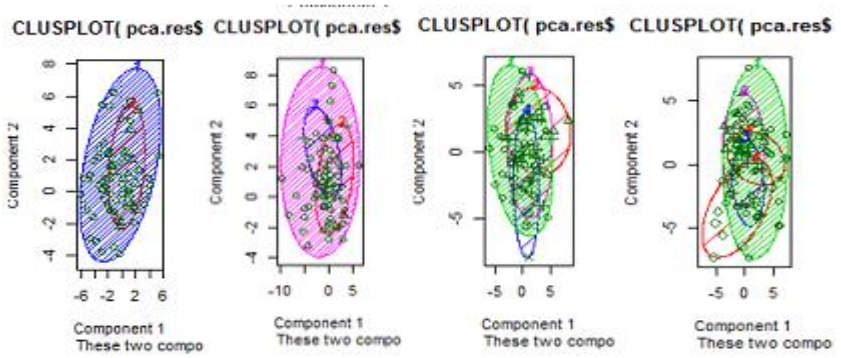

Figure 2. K-Means Clustering for Different K Values of 2,3,4,5

7. Conclusion

The Sonnet Corpus clustered into groups of size 2,3,4,5 was studied. Frequency of terms in Corpus is high. The dimensionality of the Sonnet line corpus was reduced using feature selection was proposed in this work. The Sonnet clustering was dealt with simple K-Means Clustering and different clustering results were obtained. It was found that values with minimum entropy provided better results than others. Careful choice of K results with better clustering and hence better entropy. The issue of how to deal with the values of K in K-Means clustering still remains open. In future research, clustering can be studied by considering feature set with Rhyme identification of Sonnets.

Acknowledgements

We would thank all souls who had helped us to make this paper a successful one and giving several openings in research. I also thank the R Community for providing this open source software towards successful implementation of this research work.

References

[1] A. M. Reddy and R. Gauth, "An optimum method for Enhancing the Computational Complexity of K-Means Clustering Algorthim with Improved Initial Centers", International Journal of Science and Research(IJSR), vol. 3, no. 6, **(2014)**.
[2] A. Refat, "Hierarchical and Non-Hierarchical Linear and Non-Linear Clusterng Methods to Shakespeare Authorship Question", Social Sciences, vol. 4, **(2015)**, pp. 758-799.
[3] M. B. Bangoria, "Enhanced K-Means Clustering Algorithm to ReduceTime Complexity for Numeric Values", International Journal of Computer Science and Information Technologies, vol. 5, no. 1, **(2014)**, pp. 876-879.
[4] S. B. Everitt, S. Landau, M. Lesse and D. Stahl, "Cluster Analysis", 5th Edition, Wiley, **(2011)**.
[5] J. ByrdRoy and M. S. Chodorow, "Using an online dictionary to find rhyming words and pronunciations for unknown words", Proceedings of the 23rd Annual Meeting of ACL, **(1987)**, pp. 277-283.
[6] C. Burrow, "William Shakespeare: The Complete Sonnets and Poems", Oxford University Press, **(2002)**.
[7] D. Rodolfo, "Exploring Shakespeare's Sonnets with SPARSAR", Linguistics and Literature Studies, , (http://www.hrpub.org DOI: 10.13189/lls.2016.040110), vol. 4, no. 1, **(2016)**, pp. 61-95.
[8] A. A. Rahman, J. Lein, K. Coles, E. Magurie, M. Meyer, M. Wynne, C. Johnson, A. E. Trefethen and M. Chen, "Rule based Visual Mappings-with a case study on Poetry Visualization", In Compute Graphics Form, **(2013)**, pp. 32.
[9] H. Jiawei and K. Michelin, "DataMining :Concepts and Techniques", Morgan Kaufmaan Publishers, (**2005)**.
[10] J. A. Hartigan, "Clustering Algorithms (Porbability & Mathematical Statistics)", John Wiley & Sons Inc, **(1975)**.
[11] H. A. Kent, W. H. Charles and P. A. Lake, "When Did Shakespeare Write Sonnets 1609", Studies in Philology 88.1, **(1991)**, pp. 69-109.
[12] P. Gervas, "Engineering linguistic creativty: Bird flight and jet planes", In proceedings of the NAACL HLT 2010 Second workshop on Computational Approaches to Linguistic Creativity, **(2010)**, pp. 23-30.
[13] J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, **(1967)**, pp. 281-297.
[14] B. A. J. Fisk and D. Bruster, "Clustering and Classifying the Sonnets' Dateless Night", Department of English, The University of Texas at Austin, **(2015)**.
[15] S. Julie, "A survey of the literature of clustev analysis",Comput. J., vol. 25, no. 1, **(1982)**, pp. 130-134.
[16] K. A. A. Nazeer and M. P. Sebastin, "Improving the accuracy and efficiency of the K-Means Clustering Algorithm", International Conference on Data Mining and Knowledge Engineering(ICDMKE), London, UK, Proceedings of the World Congress on Engineering (WCE2009), **(2009)**.
[17] M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, "An Efficient Enhanced K-Means Clustering Algorithm", Journal of Zhejiang University, vol. 10, no. 7, **(2006)**.
[18] M. Chaturvedi, G. Gannod, L. Mandell, H. Armstrong and E. Hodgson, "Rhyme's Challenge: Hip Hop, Poetry, and Contemporary Rhyming Culture", Oxford University Press, Literary Criticism, **(2012)**.
[19] M. Steinbach, G. Karypis and V. Kumar, "A Comparison of Document Clustering techniques", Technical Report, Department of Computer Science and Engineering, University of Minnesota, http://www.cs.umn.edu/tech_reports/, **(2000)**.
[20] N. Singh and D. Singh, "Performance Evaluation of K-means and hierarichical clustering in terms of accuracy and running time", Interantional Journal of Computer Science and Information Technologies(IJCSIT), vol. 3, no. 3, **(2012)**.
[21] Okafor and Anthony, "Entropy based techniques with applications in data Mining", Florida, University of Florida, **(2005)**.
[22] R. C. Balabantaray, C. Sarma and M. Jha, "Document Clustering using K-Means and K-Medoids", htttp://www.publishingindia.com.
[23] S. C. Sripada and M. S. Rao, "Comparison of Purity and Entropy of K-Means Clustering and Fuzzy C-Means Clustering", Indian Journal of Computer science and Engineering (IJCSE), vol. 2, no. 3, **(2011)**, pp. 343.
[24] E. S. Claude, "A mathematical theory of communication", Bell System Technical Journal, vol. 27, **(1948)**, pp. 379-423 and 623-655.
[25] T. S. Eliot, "Hamlet and his Problems", In The Sacred wood and Major early essays, 58, New York: Dover, **(1998)**.

[26] W. K. Addanki and Dekai, "Unsupervised Rhyme Scheme Identification in Hip Hop Lyrics using Hidden Markov Models", Proceedings of the 1st International Conference on Statistical Language and Speech Processing, Tarragona, Spain, **(2013)**.
[27] I. Feinerer, K. Hornik and D. Meyer, "Text Mining Infrastructure in R", Journal of Statistical Software, http://www.jstatsoft.org/v25/i05/, vol. 25, no. 5, **(2008)**, pp. 1-54.
[28] http://www.gutenburg.org/cache/epub/100/pg100.txt.
[29] http://www.R-project.org/.

Authors

T. Senthil Selvi, is a Reseach scholar and currently working as Assistant Professor in Periyar E.V.R. College, Tiruchirapalli. Her area of research is in the field of Web Mining. Other research activity include in the field of Artificial Intelligence and Information Retrieval.

R. Parimala graduated with M.Sc., Applied Science at the National Institute of Technology, (formerly Regional Engineering College) Tiruchirapalli in 1990. She received her M. Phil., Computer Science at Mother Teresa University, Kodaikanal in 1999. She started teaching in 1999 at National Institute of Technology and is currently working as Assistant Professor in Department of Computer Science, Periyar E. V. R. College (Autonomous), Tiruchirapalli. She completed her Ph.D. at National Institute of Technology, Tiruchirappalli. Her area of research interests include Neural Networks, Data mining and Optimization Techniques.