Research Article

International Journal of Advanced Computer Research, Vol 8(35)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2018.835003

Spatial distribution analysis of unigrams and bigrams of hindi literary document

Sifatullah Siddiqi*
School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India

Received: 20-November-2017; Revised: 08-March-2018; Accepted: 10-March-2018

Abstract
In this paper the spatial distribution analysis of a very famous Hindi literary document "Godan" authored by the great novelist Munshi Premchand has been presented. We have attempted to perform a thorough and comprehensive spatial distribution analysis of different kinds of words (unigram) and word pairs (bigrams) in the document. Single words have been divided into stop words, keywords and non-keywords while word pairs have been divided into stop-phrases, key phrases and non-key phrases. Our proposition is that the nature of the spatial distribution pattern of different types of unigrams and bigrams in the text is different and there is a significant similarity between spatial distribution patterns for the unigrams and bigrams of same type. In this paper, we have selected a lot of example words from the text and generated their spatial distribution graphs to prove our assertion.

Keywords
Stop words, Keywords, Key phrase, Spatial distribution analysis, Hindi.

1.Introduction
Simple frequency based model proposed by Luhn in 1958 provides us a means of distinguishing between different kinds of words present in the document [1]. But it is a very crude model to estimate the importance and characterization of various words and the results are a mixture of stop words, keywords and non-keywords. Further, there is neither any threshold to remove stop words which dominate the top positions in the list nor there is any such lower level threshold to remove non keywords words which form the bulk of the terms in the text. It is intuitive that words repeated more often should have correspondingly higher importance in the text, but empirical evidence shows anomalies in the assumption, where higher frequency words are often found to be contextually less important than medium frequency words. It is also observed during word frequency analysis that even the words with similar frequencies have differing importance. The inadequacy of frequency based method compels us to look further into the word distribution pattern in the text. The word usage pattern is dominated by the current context of the text and it is observed that words important in the context are repeated more often in that section and this gives rise to the clustered occurrence of words. If this clustering could somehow could be estimated, then a better estimate of word importance can be achieved [2].

2.Related work
The frequency of a word is an indicator of the importance of that word in the document [1]. Generally more important a word is to a document, the more number of times it should occur in the document. Also in a document the words having highest frequencies are generally the stop word, words (the, and, of, it etc.), but there is no threshold which can demarcate between stop words and the rest of the words. Luhn proposed to select the middle frequency words as keywords and discard the high and low frequency words as stop words [1].

Some work has been done by observing that not only the frequency, but also the distribution of words in the document can provide further insight for estimating the importance of a word. It was demonstrated that important words of a text have a tendency to attract each other and form clusters [2]. The standard deviation of the distance between successive occurrences of a word is such a parameter to quantify this self-attraction.

The problem of finding and ranking the relevant words of a document was handled by using statistical

*Author for correspondence

information referring to the *spatial* use of the words [3]. Shannon's entropy of information was used for automatic keyword extraction. The randomly shuffled text was used as a standard and the various measures used in the original document text were normalized by corresponding measures of random text.

Automatic extraction of keywords from literary text was performed through a generalization of the level statistics analysis of quantum disordered systems [4]. Another method was proposed for ranking, the words in texts by use of non-extensive statistical mechanics [5]. The non-extensively measure can be used to classify the correlation range between word-type occurrences in a text.

A method was proposed which improved upon the entropic and clustering approaches and proposed new metrics to evaluate the performance of keyword detectors to use them to find out the best approach of the two [6]. It was observed that in general word clustering measures perform at least as well as the entropy measure, which requires a suitable partitioning of the text and word-clustering measures are also better for short texts since these measures discriminate better the degree of relevance of low frequency words than the entropy approach.

To evaluate and rank the relevant words in a text another approach was proposed which used the Shannon's entropy difference between the intrinsic and extrinsic mode, signifying the fact that the relevant words reflect the author's writing intention, and irrelevant words are randomly distributed in the text [7].

Spatial distribution was used to extract keywords by computing the fraction of the mean intermediate distance of words [8]. In another work [9], a hybrid approach was used for keyword extraction which combined spatial distribution with frequency information to extract keyword from Hindi documents. Extraction of important bigrams of a Hindi text was achieved through modification of the spatial distribution approach for keyword extraction [10].

3.Material and methods

3.1Background for keyword analysis

The idea behind the word distribution analysis is that the word occurrence pattern for keywords should be different from that of non-keywords. For a non-keyword, its occurrence pattern should be random in the text and no significant clustering should occur while for a keyword its occurrence pattern should indicate some kind of clustering since a keyword is expected to be repeated more often in specific contexts or portions of text.

3.2Background for key phrase analysis

Phrases are generally more capable of expressing structured concepts than individual words; as the length of a phrase increases beyond unity they have a smaller degree of ambiguity than their constituent words due to the mutual disambiguation effect of words. For example, while the words "*hand*" and "*drill*" both carry ambiguity (e.g. "*a **hand** of cards*" or "*shaking of **hands***"; "*oil **drilling***" or "*pronunciation **drill***"), their combination "***hand drill***" is not ambiguous, since each of its two constituent words mutually disambiguate meaning of the phrase.

A sequence of words can make a good phrase, but not necessarily an informative one, e.g. "*in spite of*". A word sequence can be informative for a particular domain, but not a phrase; "*Ford, Honda, Toyota*" is an example of a non-phrase sequence of informative words in car domain.

The word *key phrase* implies two features: *phraseness* and *informativeness*.

Phraseness can be described as the degree to which a given word sequence can be considered to be a phrase. Generally, the idea of phraseness is dependent on the user, who has his own criteria for the target application. For instance, one user might want only noun phrases while another user might be interested only in phrases describing a certain set of products. Although there is no single definition of the term *phrase*. In this paper, we focus on the cohesion of consecutive words.

Informativeness is related to how well a phrase describes the key ideas presented in the text. Since informativeness is assessed with respect to background knowledge and new knowledge, different users will have varying perceptions of informativeness.

We have selected a very famous Hindi novel "**Godan**" by great writer Munshi Premchand to generate spatial distribution graphs for many different unigrams and bigrams from the text. Some of the document statistics are:

Number of words in the document = **167707**

Number of unique words in the document = **11160**
To keep the axis easily readable and uncluttered we have marked axis at intervals of 20,000 words and last marking is at 1,80, 000 resulting in a total of 9 intervals. Below is a sample paragraph from the document.

होरीराम ने दोनों बैलों को सानी - पानी दे कर अपनी स्त्री धनिया से कहा - गोबर को ऊख गोड़ने भेज देना | मैं न जाने कब लौटूँ | जरा मेरी लाठी दे दे |
धनिया के दोनों हाथ गोबर से भरे थे | उपले पाथ कर आई थी | बोली - अरे , कुछ रस - पानी तो कर लो | ऐसी जल्दी क्या है ?
होरी ने अपने झुर्रियों से भरे हुए माथे को सिकोड़ कर कहा - तुझे रस - पानी की पड़ी है , मुझे यह चिंता है कि अबेर हो गई तो मालिक से भेंट न होगी | असनान - पूजा करने लगेंगे , तो घंटों बैठे बीत जायगा |
'इसी से तो कहती हूँ, कुछ जलपान कर लो और आज न जाओगे तो कौन हरज होगा! अभी तो परसों गए थे | '

3.3Representing spatial distribution of words

Consider the following example text:
"Nory was a Catholic because her mother was a Catholic, and Nory's mother was a Catholic because her father was a Catholic, and her father was a Catholic because his mother was a Catholic, or had been."
In the above text, if we assign positions to the words, then:
The position of the word "***Nory***" = 1
The position of the word "***was***" = 2
The position of the word "***a***" = 3
The position of the word "***Catholic***" = 4 and so on.
For our purpose, suppose we wish to represent the distribution of word "***catholic***" in the text. The word "***catholic***" occurs at positions 4, 10, 16, 22, 28 and 34. We make a vertical mark at these positions, along the axis to represent the occurrence of words. The spatial distribution is shown in *Figure 1*.

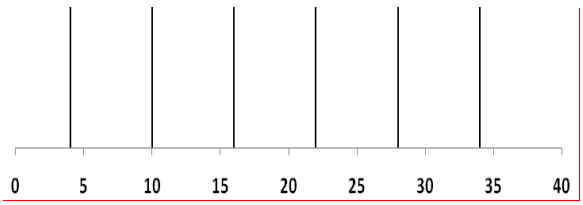

Figure 1 Representing spatial distribution of word "Catholic" in the example text

4.Results

4.1Spatial distribution of stop words

Figures 2 to 11 show the distribution pattern of 10 stop words present in the text. It can be seen that the distribution of stop words in general is random and no significant clustering is observed.

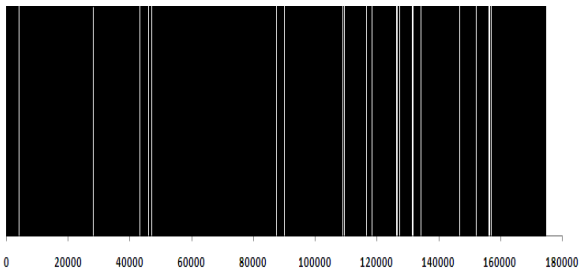

Figure 2 Stop word = "**है**", Frequency = 3917

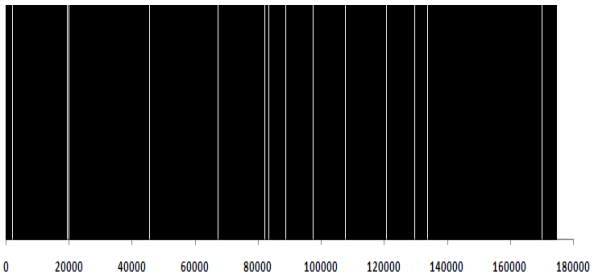

Figure 3 Stop word = "**और**", Frequency = 3114

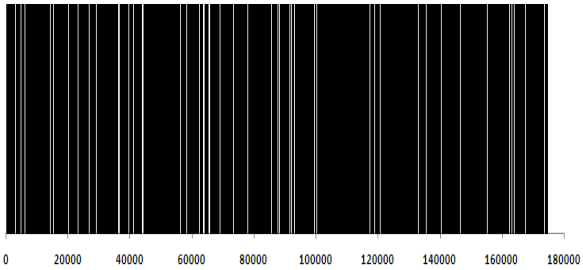

Figure 4 Stop word = "**का**", Frequency = 1994

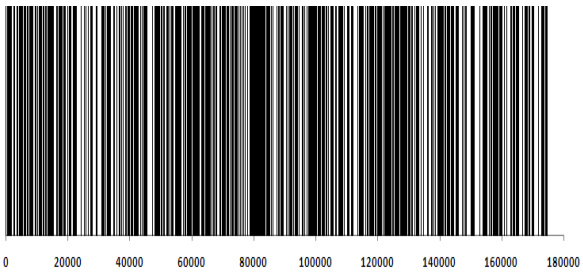

Figure 5 Stop word = "**क्या**", Frequency = 820

Figure 6 Stop word = "**गया**", Frequency = 677

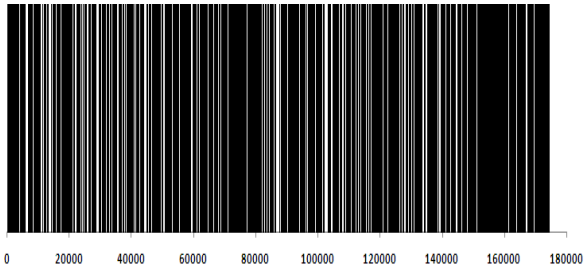

Figure 7 Stop word = "**वह**", Frequency = 1519

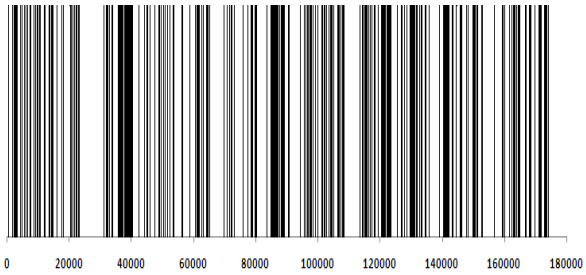

Figure 8 Stop word = "**तुम**", Frequency = 464

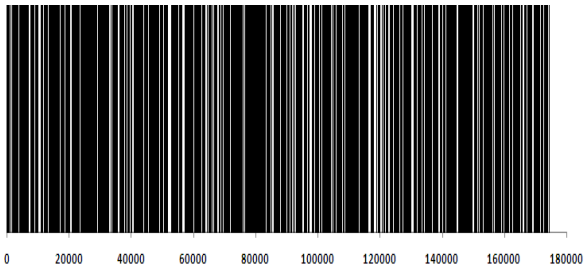

Figure 9 Stop word = "**कि**", Frequency = 1137

We have taken various stop words from the text with frequencies ranging from the lowest frequency of 467 to highest frequency of 3917. The continuous black bands seen in some stop words represent the heavy usage of those stop words in short distances.

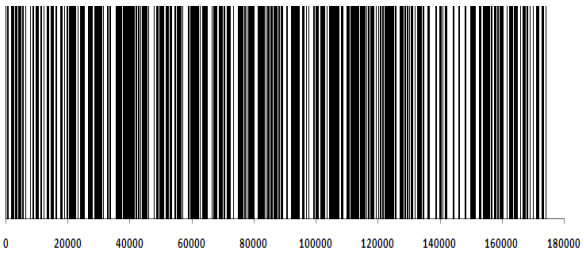

Figure 10 Stop word = "**मैं**", Frequency = 1058

Figure 11 Stop word = "**उसके**", Frequency = 538

4.2Spatial distribution of keywords in text

Figures 12 to 21 show the distribution pattern of 10 keywords from the text. It can be seen that significant clustering is present in distribution of keywords.

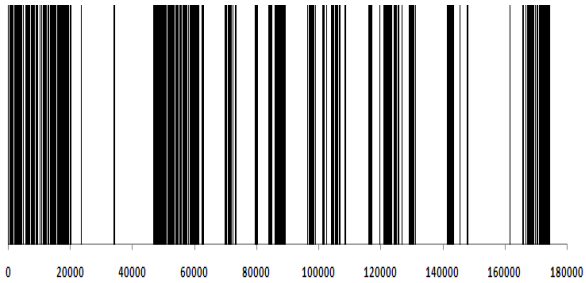

Figure 12 Keyword = "**होरी**", Frequency = 623

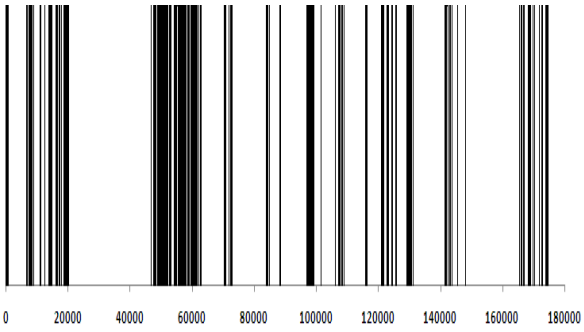

Figure 13 Keyword = "**धनिया**", Frequency = 355

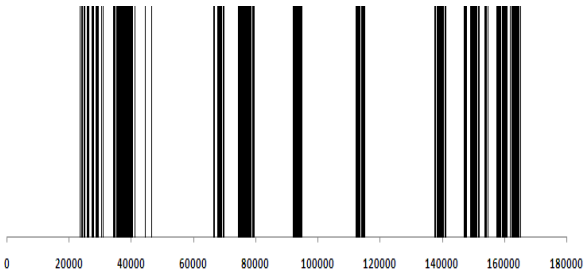

Figure 14 Keyword = "**मेहता**", Frequency = 404

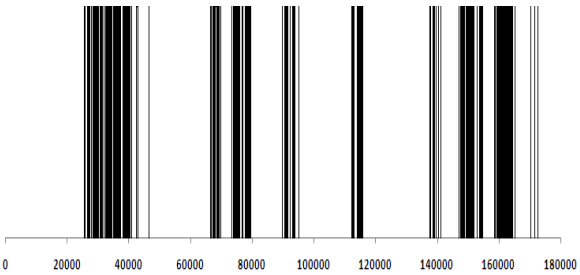

Figure 15 Keyword = "**मालती**", Frequency = 440

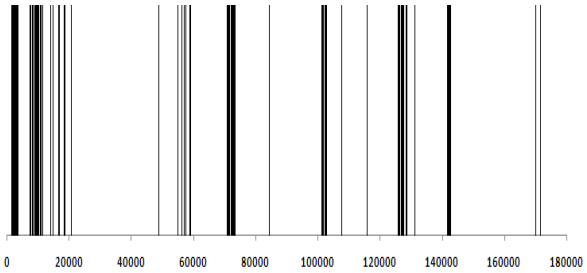

Figure 16 Keyword = "**भोला**", Frequency = 155

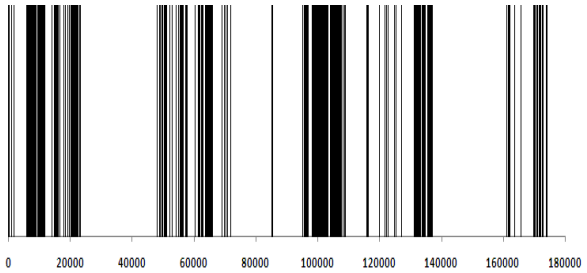

Figure 17 Keyword = "**गोबर**", Frequency = 414

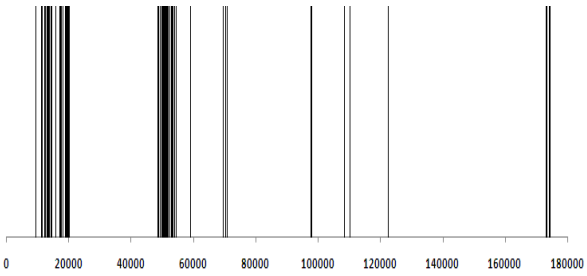

Figure 18 Keyword = "**हीरा**", Frequency = 113

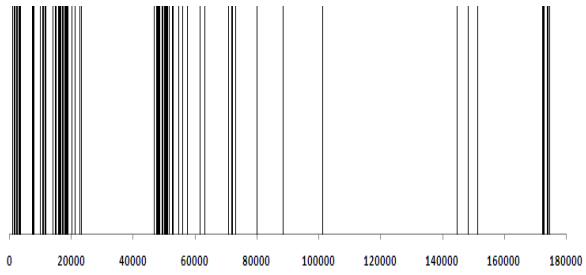

Figure 19 Keyword = "**गाय**", Frequency = 164

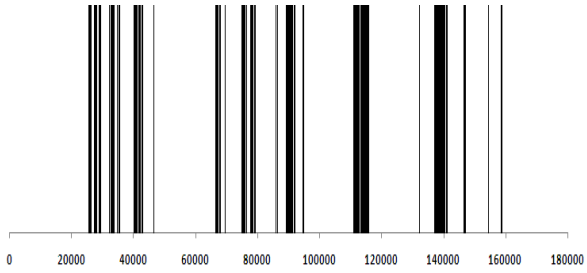

Figure 20 Keyword = "**खन्ना**", Frequency = 262

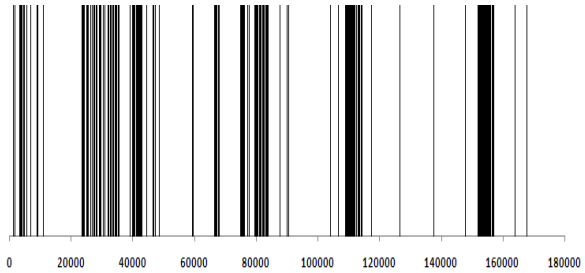

Figure 21 Keyword = "**रायसाहब**", Frequency = 271

We have taken various key words from the text with frequencies ranging from the lowest frequency of 113 to highest frequency of 623. It can be seen from the figures that there are regions of occurrence as well as regions of disappearance which gives rise to clustered distribution.

4.3Spatial distribution of non-keywords in text

Figures 22 to 41 show the distribution pattern of 20 non-keywords from the text. It can be seen that distribution of non-keywords in general is random and higher frequency words have almost similar distribution to stop the words.

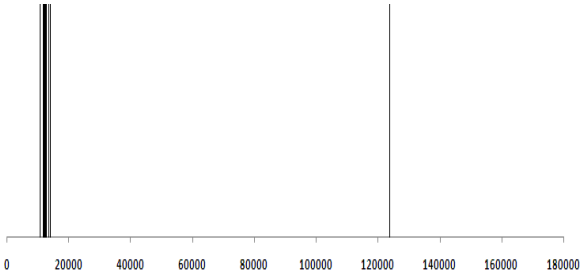

Figure 22 Non keyword = "**बाँस**", Frequency = 22

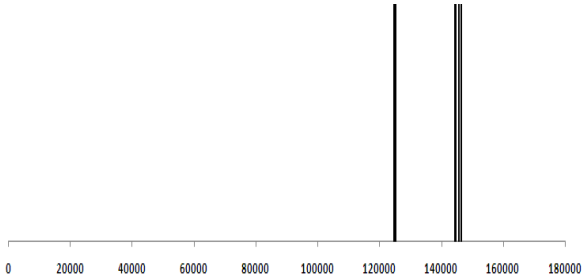

Figure 23 Non keyword = "**मथुरा**", Frequency = 26

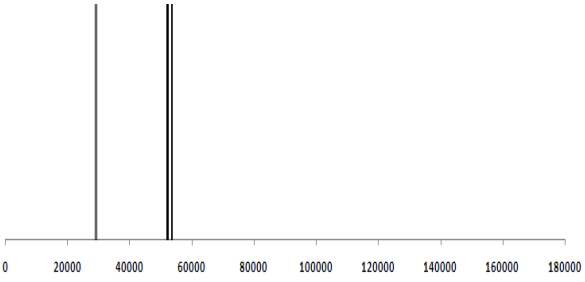

Figure 24 Non keyword = "**तलाशी**", Frequency = 18

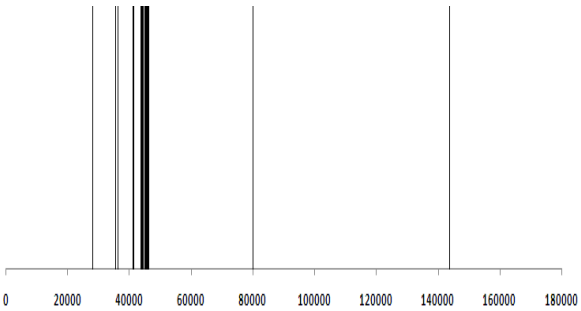

Figure 25 Non keyword = "**हिरन**", Frequency = 27

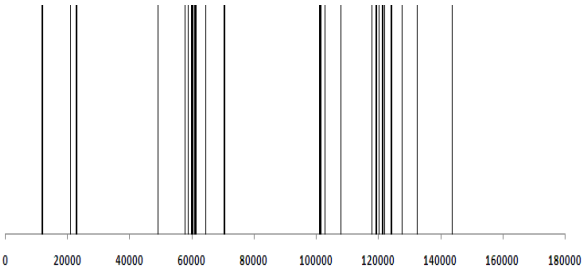

Figure 26 Non keyword = "**बिरादरी**", Frequency = 57

Figure 27 Non keyword = "**रूपए**", Frequency = 555

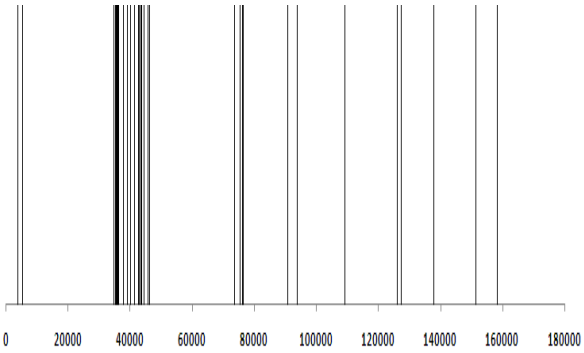

Figure 28 Non keyword = "**शिकार**", Frequency = 50

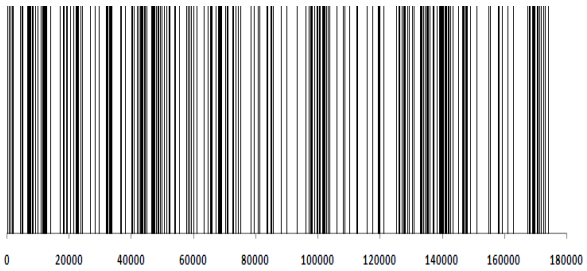

Figure 29 Non keyword = "**आदमी**", Frequency = 302

Figure 30 Non keyword = "**हाथ**", Frequency = 375

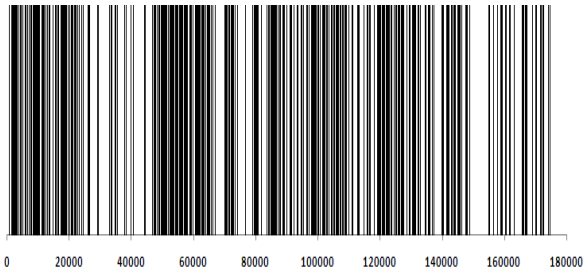

Figure 31 Non keyword = "**घर**", Frequency = 618

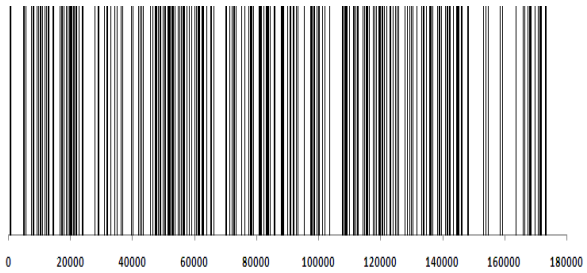

Figure 32 Non keyword = "**मुँह**", Frequency = 294

Figure 33 Non keyword = "**सिर**", Frequency = 249

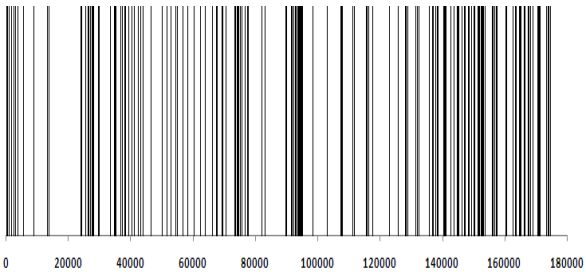

Figure 34 Non keyword = "**जीवन**", Frequency = 244

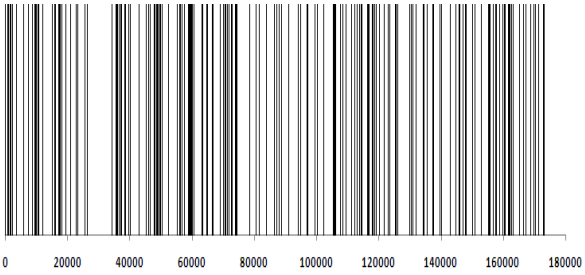

Figure 35 Non keyword = "**दोनों**", Frequency = 239

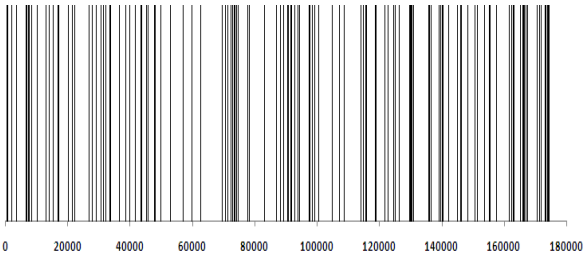

Figure 36 Non keyword = "**आँखों**", Frequency = 145

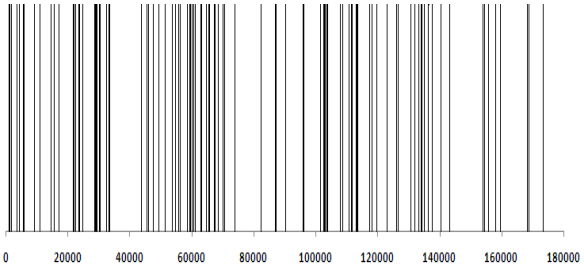

Figure 37 Non keyword = "**पाँच**", Frequency = 129

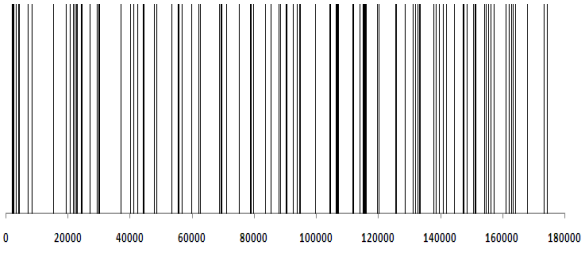

Figure 38 Non keyword = "**समझ**", Frequency = 124

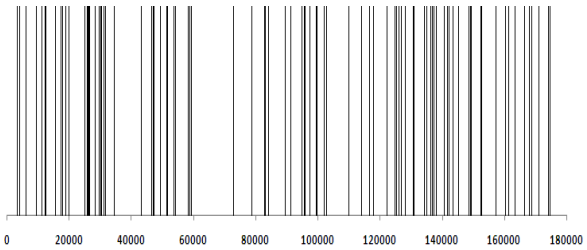

Figure 39 Non keyword = "**समय**", Frequency = 113

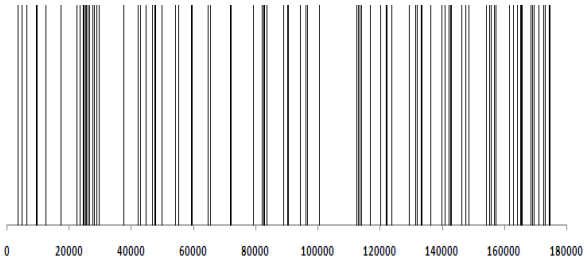

Figure 40 Non keyword = "**नाम**", Frequency = 103

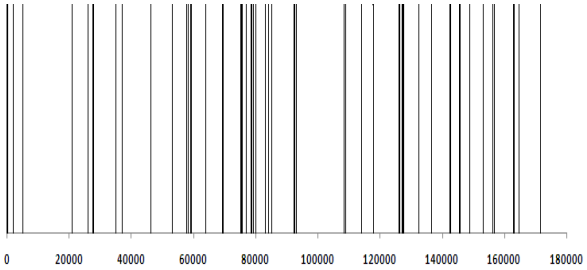

Figure 41 Non keyword = "**स्त्री**", Frequency = 85

We have taken various non-key words from the text with frequencies ranging from the lowest frequency of 18 to highest frequency of 618. It can be seen from the figures that no appropriate clustering is observed in spatial distribution of non-keywords. The distribution of many high frequency non keywords was observed to be similar to those of stop words.

4.4Spatial distribution of stop phrases in text

It was observed that most of the topmost bigrams ranked in frequency were of "***stop word-stop word***" type. We define such bigrams to be stop phrases of the text. In this paper a stop phrase is a bigram which comprises of 2 stop words (for example "*of the*", "*to the*"). The spatial distribution of non-relevant phrases follows a similar pattern as in the case of stop words and the stop-phrases of the text are also the most frequent as in the case of stop words. *Figures 42 to 51* show the distribution pattern of some of the top bigrams in our datasets.

Figure 42 Stop phrase "**आ कर**", frequency = 164

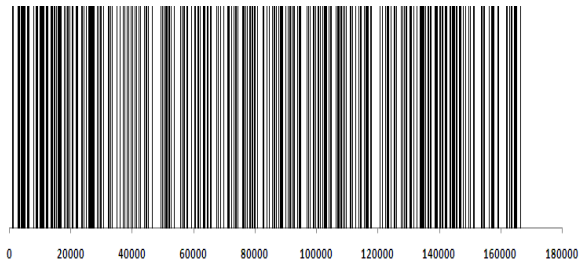

Figure 43 Stop phrase = "**के लिए**", Frequency = 432

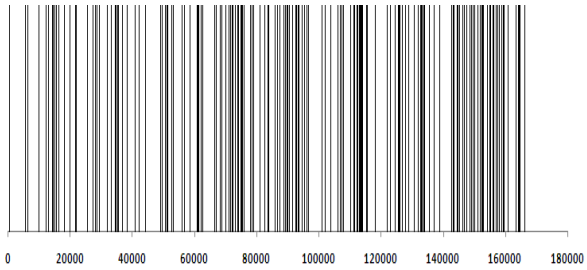

Figure 44 Stop phrase = "**हो कर**", Frequency = 186

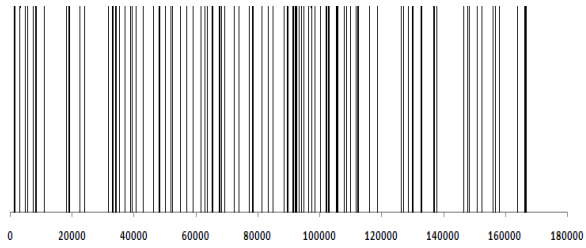

Figure 45 Stop phrase = "**रहा है**", Frequency = 120

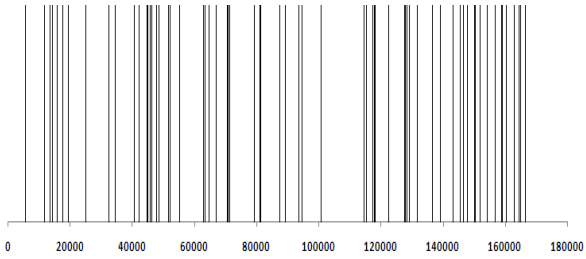

Figure 46 Stop phrase = "**हुआ था**", Frequency = 69

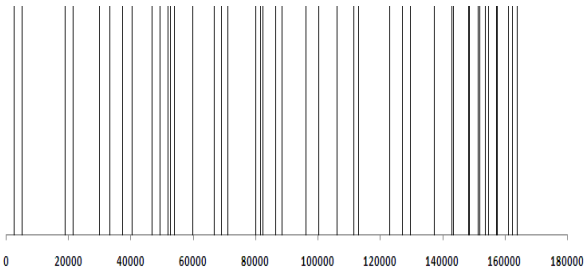

Figure 47 Stop phrase = **"वह तो**", Frequency = 46

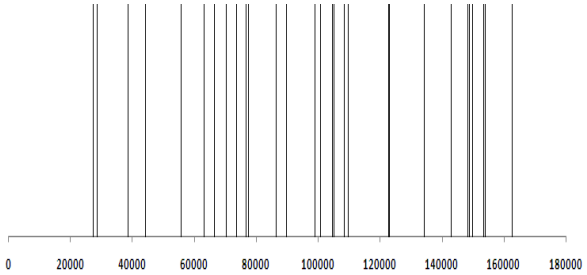

Figure 48 Stop phrase = "**उन पर**", Frequency = 30

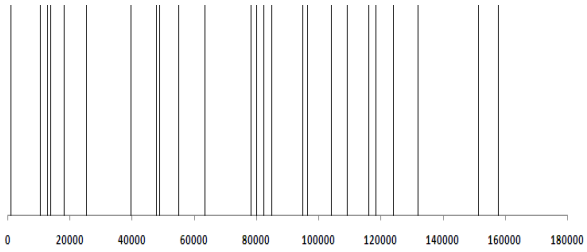

Figure 49 Stop phrase = "**हो जाये**", Frequency = 28

Figure 50 Stop phrase = "**नहीं है**", Frequency = 302

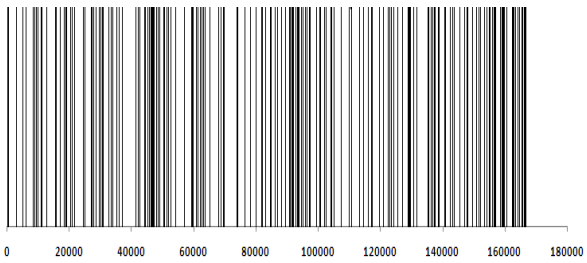

Figure 51 Stop phrase = "**हो गया**", Frequency = 239

We have taken various stop phrases from the text with frequencies ranging from the lowest frequency of 28 to highest frequency of 432. It can be seen from the figures that no appropriate clustering is observed in spatial distribution of stop phrases. Their distribution is almost similar in nature to those of stop words observed in the earlier figures.

4.5The spatial distribution of the actual key phrases in text

Key phrases have a much smaller frequencies compared to keywords in a text. *Figures 52 to 59* show the spatial distribution of 8 key phrases from the text.

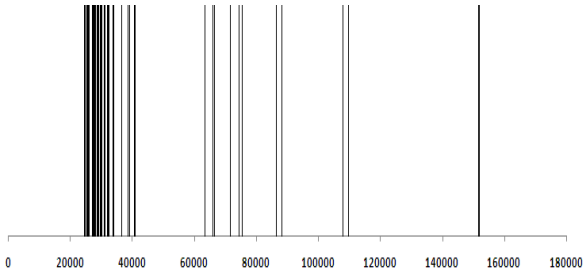

Figure 52 Key phrase = "**मिस मालती**", Frequency = 59

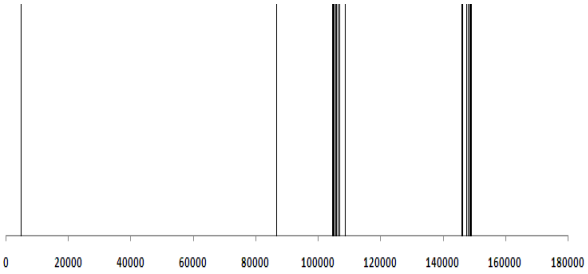

Figure 53 Key phrase = **"राजा साहब**", Frequency = 32

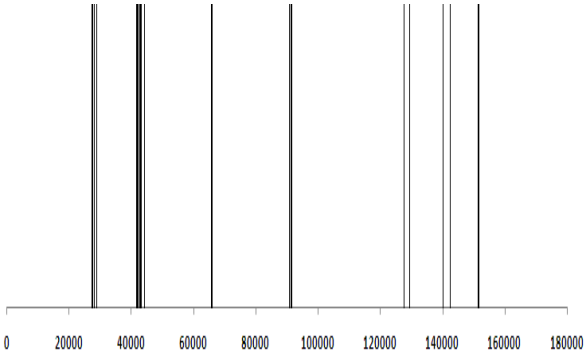

Figure 54 Key phrase = "**मिर्ज़ा जी**", Frequency = 32

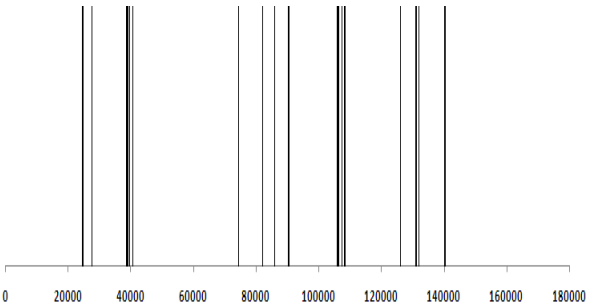

Figure 55 Key phrase = "**मिस्टर खन्ना**", Frequency = 30

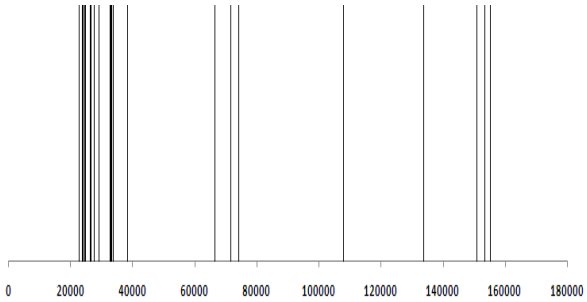

Figure 56 Key phrase = "**मिस्टर मेहता**", Frequency = 29

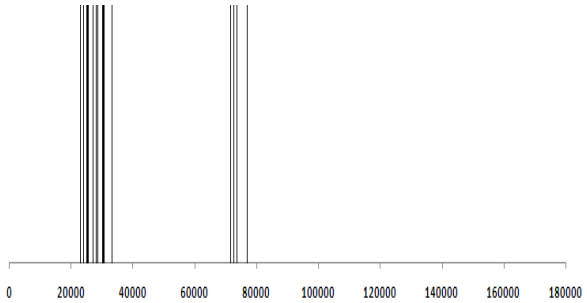

Figure 57 Key phrase = "**संपादक जी**", Frequency = 28

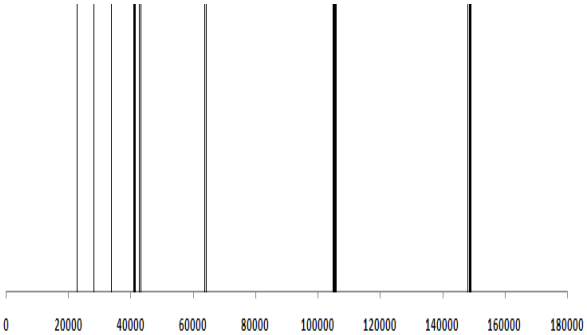

Figure 58 Key phrase = "**मिस्टर तंखा**", Frequency = 25

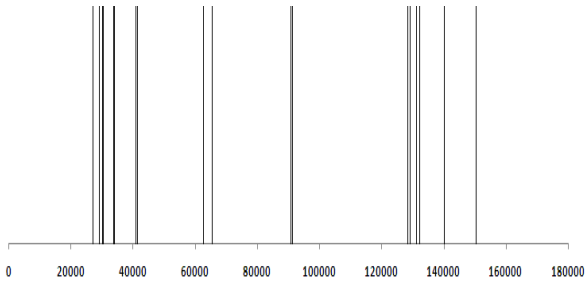

Figure 59 Key phrase = "**मिर्ज़ा खुर्शेद**", Frequency = 24

We have taken various key phrases from the text with frequencies ranging from the lowest frequency of 24 to highest frequency of 59. It can be seen that the spatial distribution of actual key phrases is not similar to those of keywords as observed in earlier figures. Their distribution doesn't give regions of clustered occurrence like keywords.

4.6Spatial distribution of non-key phrases having distribution similar to keywords

Some bigrams were found to have their spatial distributions similar to distribution of keywords. These non-key phrases were found to be of "***keyword stopword***" type or of "***stopword keyword***" type. *Figures 60 to 69* show the spatial distribution of 10 bigrams which are not key phrases of text.

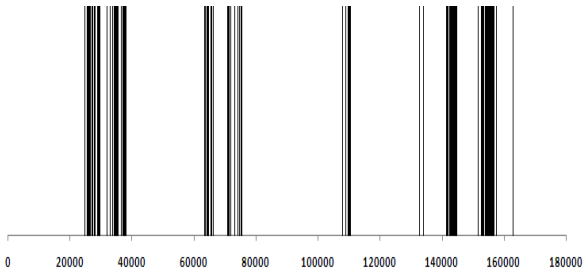

Figure 60 Non Key phrase = "**मालती ने**", Frequency = 131

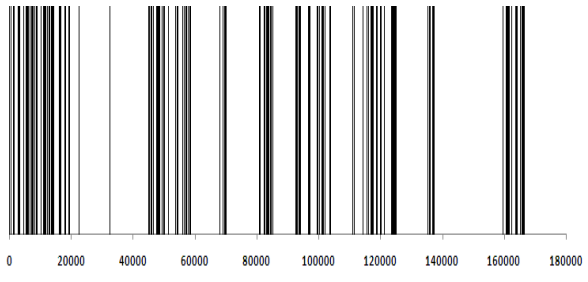

Figure 61 Non Key phrase = "**होरी ने**", Frequency = 190

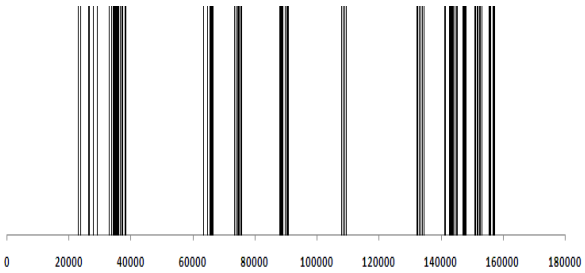

Figure 62 Non Key phrase = "**मेहता ने**", Frequency= 130

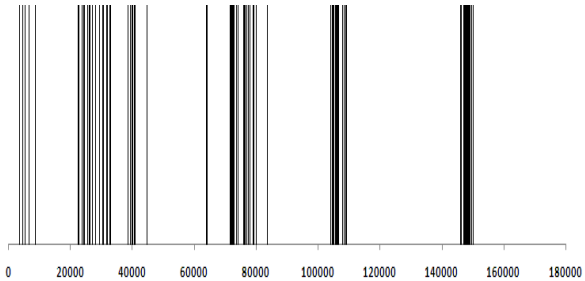

Figure 63 Non Key phrase = "**रायसाहब ने**", Frequency = 97

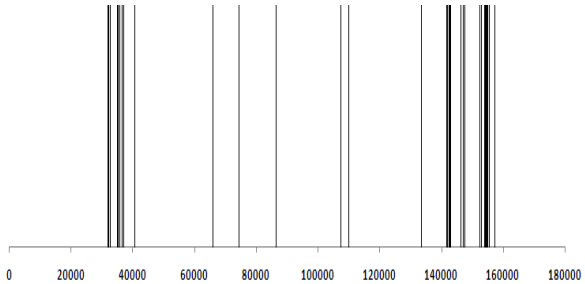

Figure 64 Non Key phrase = "**मालती को**", Frequency = 41

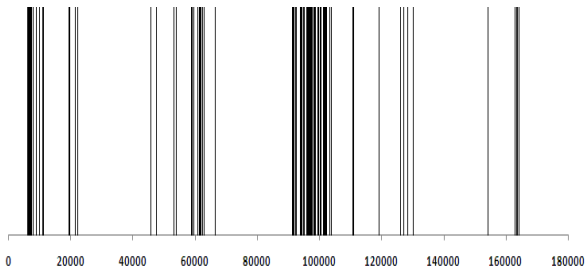

Figure 65 Non Key phrase = "**गोबर ने**", Frequency = 101

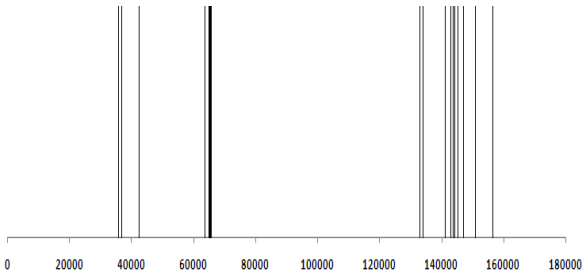

Figure 66 Non Key phrase = "**और मेहता**", Frequency = 21

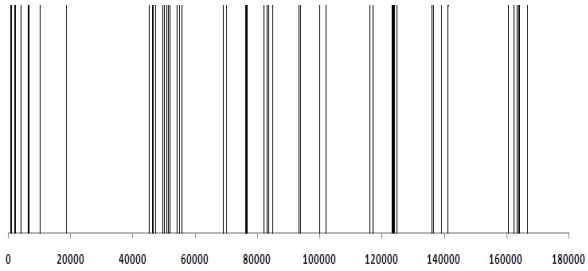

Figure 67 Non Key phrase = "**होरी को**", Frequency = 61

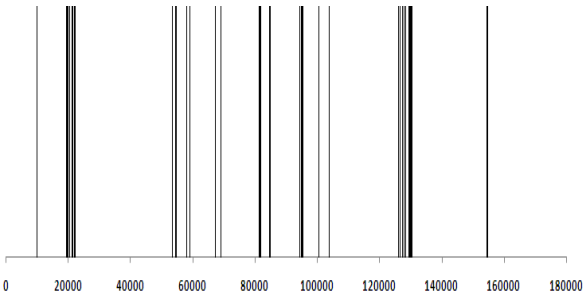

Figure 68 Non Key phrase = "**झुनिया ने**", Frequency = 52

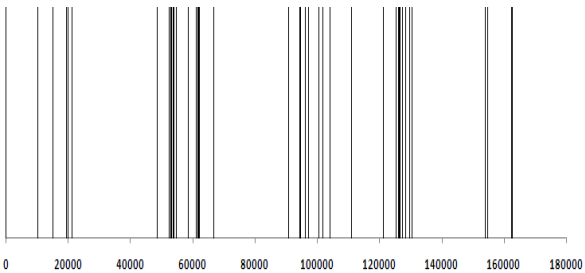

Figure 69 Non Key phrase = "**गोबर को**", Frequency = 49

We have taken various non-key phrases from the text which showed spatial distributions almost similar to those of keywords with frequencies ranging from the lowest frequency of 21 to highest frequency of 190. The distribution is not dense as was in case of keywords due to the smaller frequencies of bigrams compared to unigrams.

5.Discussion

Analysis of spatial distribution of various stop words of the text with varying frequencies reveals the fact that their distributions are quite similar and random in nature. The heavy usage of stop words across the text results in almost band like structure.

Analysis of various known keywords from the text with varying frequencies also resulted in distributions of similar nature. Regions of frequent occurrence as well as regions of disappearance were observed, which results in clustered distribution. Various non-key words with differing frequencies showed no significant clustering in their spatial distribution. The distributions of many non-keywords of high frequency were seen to be similar to those of stop words suggesting that non keyword also follows the random distribution.

Many stop phrases were also analyzed and their distribution was also found to be similar in nature to those of stop words.

The analysis of key phrases revealed that their distribution is not similar to those of keywords. Regions of clustered occurrence were not observed for key phrases.

Non key phrases of medium to high frequency showed the spatial distribution almost similar in nature to those keywords. Since the frequencies of bigrams as compared to unigrams are quite smaller the distribution was not found to be as dense as was in case of keywords.

6.Conclusion

In this paper a thorough and comprehensive spatial distribution analysis of different types of unigrams and bigrams in a famous Hindi literary document has been presented. Unigrams and bigrams were divided into 3 categories (stop words, keywords and non-keywords) and into (stop-phrases, key phrases and non-key phrases) respectively. The idea behind the spatial distribution analysis is that word occurrence pattern for different types of word units should be different. In our paper we have selected a lot of example unigrams and bigrams from the text and generated their spatial distribution graphs to study the various distribution types of words. The distribution of stop words was found to be random while those of keyword showed clustering and non-keywords also showed random distribution. In case of bigrams the stop phrases (bigrams composed of two stop words) showed similar distribution as stop words. Actual key phrases were not found to occur in clusters while many bigrams were found to cluster distribution, but they were not key phrases of the text. In future this work can be extended to other works of the same author as well as many other authors and can even be performed on texts of different languages to understand the basic cognitive structure of human written language.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Luhn HP. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development. 1957; 1(4):309-17.

[2] Ortuno M, Carpena P, Bernaola-Galván P, Munoz E, Somoza AM. Keyword detection in natural languages and DNA. Europhysics Letters. 2002; 57(5):759-64.

[3] Herrera JP, Pury PA. Statistical keyword detection in literary corpora. The European Physical Journal B. 2008; 63(1):135-46.

[4] Carpena P, Bernaola-Galván P, Hackenberg M, Coronado AV, Oliver JL. Level statistics of words: finding keywords in literary texts and symbolic sequences. Physical Review E. 2009; 79(3):1-4.

[5] Mehri A, Darooneh AH. Keyword extraction by nonextensivity measure. Physical Review E. 2011; 83(5): 1-6.

[6] Carretero-Campos C, Bernaola-Galván P, Coronado AV, Carpena P. Improving statistical keyword detection in short texts: entropic and clustering approaches. Physica A: Statistical Mechanics and its Applications. 2013; 392(6):1481-92.

[7] Yang Z, Lei J, Fan K, Lai Y. Keyword extraction by entropy difference between the intrinsic and extrinsic mode. Physica A: Statistical Mechanics and its Applications. 2013; 392(19):4523-31.

[8] Siddiqi S, Sharan A. Keyword extraction from single documents using mean word intermediate distance. International Journal of Advanced Computer Research. 2016; 6(25):138-45.

[9] Sharan A, Siddiqi S, Singh J. Keyword extraction from Hindi documents using statistical approach. In intelligent computing, communication and devices 2015 (pp. 507-13). Springer, New Delhi.

[10] Siddiqi S, Sharan A. Keyword and keyphrase extraction from single Hindi document using statistical approach. In international conference on signal processing and integrated networks 2015 (pp. 713-8). IEEE.

Sifatullah Siddiqi is a Research Scholar at School of Computer and Systems Sciences at Jawaharlal Nehru University (JNU), New Delhi. His current research interests are unsupervised and statistical keyword/key phrase extraction techniques for documents. He did his M. Tech. in computer science from JNU and completed his B.Tech. in Computer Engineering from Zakir Hussain College of Engineering & Technology, Aligarh Muslim University (AMU), Aligarh.
Email: sifatullah.siddiqi@gmail.com