Research Article

International Journal of Advanced Computer Research, Vol 8(34)
ISSN (Print): 2249-7277   ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2017.733030

Construction of a generic stopwords list for Hindi language without corpus statistics

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

Received: 27-September-2017; Revised: 02-December-2017; Accepted: 09-December-2017

Abstract

Most of the research in the field of information retrieval (IR) has focused on the English language, but recently there has been a considerable amount of work and effort to develop IR systems for languages other than English. Research and experimentation in the field of IR in the Hindi language are relatively new and limited compared to the research that has been done in English, which has been dominant in the field of IR for a long while. A fundamental tool in IR is the employment of stop word lists. Stop words have no retrieval value in IR. Till now, many stop word lists have been developed for English, European and Chinese languages. However, there is no standard stop word list which has been constructed for Hindi language. In this paper an approach to construct a generic stop word list for Hindi language have been presented. Our list contains more than 800 stop words.

Keywords

Stop word, Stop words list, Hindi language, Information retrieval, Text mining, Corpus statistics.

1.Introduction

Our ability to access and store information has grown tremendously since the web became popular. It is possible to store several million webpages and hundreds of thousands of text files. Unless a topic was very broad, we were unlikely to be overwhelmed with a large volume of information. With the increasing popularity of digital media, searching over the internet to gather information has also become a common task.

Although most of the research in the field of IR has focused on the English language, recently there has been a considerable amount of work and effort to develop IR systems for languages other than English. Research and experimentation in the field of IR in the Hindi language are relatively new and limited compared to the research that has been done in English, which has been dominant in the field of IR for a long while.

Text mining is a rapidly expanding field that attempts to extract important information from natural language text.

The phrase "text mining" is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information [1].

Text mining is an active research area in IR and natural language processing. In comparison to the type of data in databases, text has no such well-defined structure, any kind of information can be present anywhere in the text thus is not easy to handle algorithmically. Text mining generally deals with texts which communicate factual information or opinion, and the need to automatically extract important information from such texts is well realized, even if limited success is achieved in such attempts.

A fundamental tool in text mining techniques is the employment of stop word lists. Stop words were first introduced in 1958 by H. P. Luhn [2], who paved the way for automatic indexing and IR. In the traditional view, words in documents that frequently occur, but have no retrieval value in the IR, such as "and", "the", "of" etc. in English documents. Stop words almost can be found virtually in every sentence. These words account for a very significant fraction of all text size. The set of stop words is generally known as stop words list. Elimination of stop words is one of

[*]Author for correspondence

the first stages in typical IR systems. On English Web searching stop words are removed and they do not influence the retrieval process significantly. Till now, many stop word lists have been developed for the English language. However, there is no standard stop word list which has been constructed for Hindi language.

1.2 Hindi language

Hindi is certainly one of the most widely spoken languages of India and is written in Devanagari script. In many states of India it is a dominant language and is also used in various countries all over the world. It is, along with English, the official language of the Indian Union. Hindi is also widely used in mass media: a significant number of newspapers, magazines, films, music, radio and television programs, advertisements etc. are produced in Hindi. In North India, Hindi is often the medium of instruction in government schools. It is also extensively used in administration, the legislature and lower judiciary. Thus, Hindi is a very important language at local, regional, national and international levels.

2.Related studies

First stop word list of English language was created by Fox [3] based on word usage in English. His stop word list consists of 421words and his method of creating the list is most frequently used method. The issue with this method is that several arbitrary decisions are taken while constructing the list such as the word frequency threshold to demarcate between stop words and non-stop words. Generally the elimination and insertion of few words are based on personal judgement, which requires certain expertise in the concerned language.

Research suggests that greater than 50% of all the words in a typical small English passage are from a list of about 135 common words [4]. These words were considered to be noise words by Van Rijsbergen [5], and should be removed before any pre-processing in text analysis. Stop words contribute almost no information in the text classification task. Elimination of stop words would contribute in reducing the size of the text space considerably and help in speeding up the accuracy and efficiency of text classification [6].

Many stop word lists have been developed in English language and the most commonly used stop word lists are the Van Rijsbergen stop word list and the Brown corpus stop word list. These stop word lists were traditionally extracted from frequency analysis of the words in large corpus [7].

There exists a disparity among different stop word lists for different document corpuses and different text processing tasks. Recently [8] proposed aggregation technique for automatic construction of general stop words of Malay language using three different approaches viz. statistical, word distribution in documents using variance measure and using the entropy measure. An evolutionary technique was proposed by [9] to extract the optimal set of stop words from the critical infrastructure domain.

Two methodologies using word frequency and statistics were proposed by [10] to generate stop words for Chinese patents. An approach was proposed by [11] for generating stop word list from online social network corpus in Egyptian dialect to assess the effect of removing dialect stop words on the sentiment analysis task. Another approach was presented in [12] for stop word removal from Hindi language using a deterministic finite automata. Since generally the stop words list is constructed from the statistics of a standard large-scale corpus, which is unavailable for Hindi language, therefore the issue of Hindi stop words has not been addressed yet properly.

3.Properties of stop words

Following are some of the properties of stop words:

1. Stop words are the common words with low discriminatory power to distinguish between documents.
2. Usual elements in a stop word list are articles, prepositions and conjunctions, although specific nouns, verbs or other grammatical types could be of low importance in terms of IR in specific domains.
3. Stop words almost never have any predictive capability and they serve only a syntactic function, but do not indicate the subject matter.
4. They can affect the efficiency of the IR process because they have a very high frequency.
5. Stop words tend to diminish the impact of frequency differences among less common words, thus affecting the weighting process.
6. The removal of the stop words changes the document length and subsequently affects the weighting process.
7. They can also affect efficiency due to the fact that they carry no meaning, which results in a large amount of unproductive processing.

8. The removal of the stop words can increase the efficiency of the indexing process as 30 to 50% of the tokens in a large text collection can represent stop words.

4.Some issues with Hindi documents

There are various limitations and handicaps present while working with a language like Hindi such as:

1. Hindi or Indian languages are highly inflected and provide rich and challenging set of linguistic and statistical features.
2. Indian languages are of highly free word order.
3. Hindi/Indian language is very resource poor language. Annotated corpora, named dictionaries, POS taggers etc. are still not available in the required quality and quantity. In contrast to this for English and many other European languages we have a large set of tools and corpora available to work on.
4. Although Indian languages have a very old and rich literary history, technological developments are recent.
5. Even very basic preliminary resource such as list of stop words in Hindi is not available and whatever is available is neither complete nor authentic.
6. Another problem is that it is difficult to use stemming to reduce the size of the vocabulary.
7. Standard statistical analysis tools with support for Indian languages are simply not available for all the standard renowned tools were made in English or allied languages and hence handle ASCII characters only, whereas we require a tool which can handle the Unicode character encoding to deal with Indian languages.

5.Methodology

There is no standard document independent measure of word frequency count to remove stop words which dominate the top positions in the list. Also, there is no similar lower level threshold to remove non keywords words which form the bulk of the terms in the text. Also, quality of stop words list has an important bearing on the effectiveness of any text mining approach.

5.1Types of stop words list

There are two kinds of stop words; *generic* and *domain specific*. Generic stop words are those common stop words which are found in the texts of nearly all different domains of the language (e.g. "**the**", "**and**", "**of**" etc. in English and "**और**", "**का**", "**के**" etc. in Hindi) whereas domain specific stop words are those which are not common in general literature, but they have very low discriminatory power in a collection of domain specific documents because they are representative of that corpus. They are good candidates for inter- domain classification, but inferior candidates for intra-domain classification. For example "**algorithm**" and "**scheduling**" can be among the candidates of a domain specific stop word list if the document collection deals in scheduling algorithms. In this paper, we have focussed upon constructing a generic stop word list for Hindi language without using the corpus statistics.

5.2Stop words list construction for Hindi

We have created a generic stop word list of 800+ words of Hindi language with the help of linguistic experts. The reason for having a large number of stop words in the list is due to the characteristics of the Hindi language. The steps as outlined in *Figure 1* have been followed to create a Hindi stop words list. As stemming is difficult in Hindi and there are no standard stemming algorithms available as in English such as a porter's stemming algorithm, we have added the various possible inflected variants of a particular stop word in the list for completeness.

5.3Result: Hindi stop words list

अंत, अंतिम, अंदर, अकेला, अकेली, अकेले, अक्सर, अगर, अगला, अगली, अगले, अच्छा, अच्छी, अच्छे, अच्छाई, अतः, अतिरिक्त, अधिक, अधिकतर, अधिकतम, अधिकांश, अनुमति, अनुरूप, अनुसार, अनेक, अन्य, अन्यत्र, अन्यथा, अपना, अपनी, अपने, अपनों, अपनापन, अपनेपन, अपेक्षाकृत, अब, अबकी, अभी, अथवा, अर्थात्, अरे, अलग

आंतरिक, आ, आई, आईं, आऊँ, आऊंगा, आऊंगी, आईए, आइएगा, आखिर, आगे, आज, आजकल, आता, आती, आते, आदि, आना, आने, आप, आपका, आपकी, आपके, आपको, आपने, आपमें, आपसे, आमतौर, आया, आयी, आए, आएंगे, आओ, आना, आइये, आएँगे, आएगा, आएगी, आओगे, आवक, आह, आहा, आहो,

इंगित, इच्छा, इच्छित, इच्छुक, इतना, इतनी, इतने, इतनों, इत्यादि, इधर, इन, इनके, इनको, इनका, इनकी, इनसे, इनमें, इन्हें, इन्हीं, इन्होंने, इस, इसका, इसकी, इसके, इसको, इसे, इसने, इसमें, इससे, इसी, इसलिए, इसीलिए

उदाहरण, उधर, उन, उनका, उनकी, उनके, उन्हें, उनसे, उनको, उनमें, उन्होंने, उन्हें, उन्हीं, उठ, उठा, उठी, उठे, उठना, उठने, उठाना, उठाओ, उठाया, उठता, उठती, उठतीं, उठते, उठवा, उठाइए, उठाई, उठाऊँ, उठाऊँगा, उठाऊंगी, उठूं, उठूँगा, उठूँगी, उठाए, उठाएगा, उठाएगी, उठाओगे, उठाता, उठाती, उठाते, उठेगा, उठेगी, उठेंगे, उठो, उतना, उतनी, उतने, उतर, उतरता, उतरती, उतरने, उतरा, उतरी, उतरे, उतरेंगे, उतरेगा, उतरेगी, उतरना, उतरने, उतारना, उतारने, उतारते, उतारा, उतारो, उसकी, उसका, उसके, उसको, उसमें, उसे, उसने, उससे, उसी, ऊई, ऊपर

एक, एकदम, एवं, ऐसा, ऐसी, ऐसे, ऐसों

ओर, और, औरों

कई, कब, कभी, कम, कमी, कर, करके, करता, करती, करते, करो, करना, करनी, करने, करें, करेंगे, करेगा, करेगी, करोगे, करोगी, किया, कीजिये, करूँ, करूंगी, करूँगा, करा, किए, कल, कह, कहा, कही, कहो, कहना, कहने, कहनी, कहते, कहता, कहती, कहेंगे, कहेगा, कहेगी, कहोगे, कहोगी, कहिये, कहके, कहलाती, कहलाते, कहलाना, कहलाने, कहूँ, कहूँगा, कहूँगी, कहाँ, कहीं, का, के, की, कि, किन्तु, कितना, कितनी, कितने, कितनों, किधर, किन, किनका, किनकी, किनके, किन्हें, किन्होंने, किस, किसी, किसे, किसका, किसकी, किसके, किसको, किसने, किसमें, किससे, किसलिए, कुछ, कृपया, केवल, कैसे, कैसा, कैसी, को, कोई, कौन, क्या, क्यों, क्योंकि

खुद, खूब

गए, गया, गयी, गिर, गिरा

चंद, चल, चला, चली, चले, चलो, चलना, चलने, चलता, चलती, चलते, चलोगे, चलोगी, चलेंगे, चलेगा, चलेगी, चलिए, चारों, चाह, चाहता, चाहती, चाहते, चाहा, चाहना, चाहिए, चाहोगे, चाहोगी, चाहेंगे, चाहे, चाहेगा, चाहेगी, चूंकि

जगह, जब, जबकि, जभी, जल्द, जल्दी, जहां, जा, जाता, जाती, जाते, जाना, जाने, जाएंगे, जाओ, जाओगे, जाओगी, जाएगा, जाएगी, जाइए, जाऊँ, जाऊँगा, जाऊंगी , जान, जानता, जानती, जानते, जानना, जिधर, जिस, जिसे, जिसका, जिसकी, जिसके, जिसको, जिसमें, जिससे, जी, जैसा, जैसी, जैसे, जो, ज्यादा, जल्दी

ठीक

तक, तब, तभी, तथा, तरफ, तरह, तुम, तुम्हारा, तुम्हारी, तुम्हारे, तुमने, तुम्हें, तुम्हीं, तुमसे, तुमसा, तुमसी, तू , तूने, तेरा, तेरे, तुझे, तुझसे, तुझको, तुझमें, तुरंत, तैसे, तो

था, थी, थे, थीं, थोड़ा, थोड़ी, थोड़े

दिखाई, दिखा, दिखी, दिखे, देख, देखा, देखी, देखो, देखेगा, देखेगी, देखोगे, देखोगी, देखेंगे, देखें, देखता, देखती, देखते, देखना, देखने, देखकर, देखिये, देखूँ, देखूँगा, देखूँगी, दूर, दूँ, दूँगा, दूँगी, देता, देते, देना, देने, देती, दिया, देगा, देगी, देंगे, दिए, दी, दीजिये, दे, दें, दो, दोनों, दौरान, द्वारा

धीरे

न, नई, नया, नए, नहीं, नीचे, नीचा, ने

पर, परन्तु, परे, पर्याप्त, परसों, पहले, पहला, पहली, पहुँच, पहुँचा, पहुँची, पहुँचे, पा, पाई, पार, पास, पीछा, पीछे, पुराना, पुरानी, पुराने, पूछ, पूछना, पूछने, पूछी, पूछे, पूछा, पूछो, पूछेंगे, पूछोगे, पूछोगी, पूछेगा, पूछेगी, पूछता, पूछती, पूछते, पूछिए, पूरी, पूरा, पूरे, पूर्व

प्रकार, प्रति, प्रत्येक

फिर

बंद, बजाय, बन, बनता, बनती, बनते, बनने, बनना, बना, बनी, बने, बनो, बनाना, बनाया, बनाओ, बनेंगे, बनेगा, बनेगी, बनोगी, बनोगे, बल्कि, बहुत, बाद, बार, बारे, बावजूद, बाहर, बिना, बिल्कुल, बीच, बेहतर, बैठ, बैठा, बैठी, बैठना, बैठता, बैठती, बैठेगा, बठेगी, बैठेंगे

भर, भरा, भरो, भरें, भरना, भरने, भरता, भरती, भरते, भरेगा, भरेगी, भरेंगे, भरोगे, भरोगी, भरिये, भी, भेजा, भेजी, भेज, भेजना, भेजो, भेजोगे, भेजेंगे, भेजेगा, भेजेगी, भेजता, भेजती, भेजते

मगर, मान, मानना, माना, मानी, माने, मानने, मानो, मानता, मानती, मानते, मानेंगे, मानोगे, मानोगी, मानेगा, मानेगी, मालूम, मिल, मिलना, मिलने, मिलता, मिलती, मिलते, मिला, मिली, मिले, मिलो, मिलोगे, मिलोगी, मिलेंगे, मिलेगा, मिलेगी, मुझे, मुझसे, मुझमें, में, मैं, मेरा, मेरी, मेरे, मैंने

यदि, यह, यही, यहाँ, यहीं, या, यानी, युक्त, ये

रख, रखना, रखने, रखता, रखती, रखते, रखा, रखो, रखे, रखेंगे, रखेगा, रखेगी, रखोगे, रखोगी, रखिये, रहता, रहती, रहते, रहा, रही, रह, रहे, रहो, रहेंगे, रहेगा, रहेगी, रहेंगे,

रहोगे, रहोगी, रहिये, रहना, रहने, रहूँगा, रहूँगी, रहूँ, रास्ता, रास्ते

लग, लगता, लगती, लगते, लगना, लगा, लगे, लगी, लगेगा, लगेगी, लगोगे, लगोगी, लगाना, लगाइए, लगाओ, लगना, लगने, लगभग, ला, लाना, लाया, लाए, लिए, लूँ, लूँगा, लूँगी, ले, लेना, लेने, लेता, लेती, लेते, लेंगे, लेगा, लेगी, लोगे, लोगी, लिया, लीजिये, लाना, लाने, लाई, लाया, लो, लेकिन, लोग, लोगों

व, वह, वही, वो, वे, वहाँ, वहीं, वाह, वाला, वाली, वाले, वालों, विभिन्न, वैसे, वैसा, वैसी, व्यक्ति

शामिल, शायद

संभव, संभावना, संभावित, सकता, सकती, सकते, सके, सकी, सका, सकेंगे, सकेगा, सकेगी, सकोगे, सकोगी, सदैव, सब, सबसे, सबका, सबकी, सबको, सबमें, सभी, समर्थ, सराहना, सह, सही, सुन, सुना, सुनी, सुनना, सुनेंगे, सुनेगी, सुनेगा, सुनो, सुनोगे, सुनोगी, सुनता, सुनती, सुनते, सुनने, सा, साथ, सामने, सामान्य, सारा, सारी, सारे, सिवाय, सी, से, सोच, सोचना, सोचने, सोचेंगे, सोचेगा, सोचेगी, सोचोगे, सोचोगी, सोचता, सोचती, सोचते, सोचिये, स्वयं

हम, हमें, हमारा, हमारी, हमारे, हमसे, हर, हां, हाय, हाल, हालांकि, ही, हूँ, है, हैं, हो, हों, होगा, होगी, होगे, होंगे, होंगी, हूंगा, हूंगी, होता, होती, होते, होना, होनी, होने, हुआ, हुई, हुए

6.Discussion

A Generic stop word list for Hindi has been constructed (e.g. "**और**", "**का**", "**के**" etc.) and to make the list as complete as possible, we have also added the inflected variants of a stop words (for stop word "**उन**" following words are its inflected variants such as "**उनका**", "**उनकी**", "**उनके**", "**उन्हें**", "**उनसे**", "**उनको**", "**उनमें**" etc.).

Domain specific stop word list can be constructed by adding more number of common words on top of this list. For example "**नदियाँ**" and "**बाँध**" can be added to this list if the document collection related to rivers of India. This list has been constructed without using the corpus based statistics because no standard substantial dataset is available in Hindi like Brown Corpus for English language and thus frequency based analysis cannot be performed reliably.

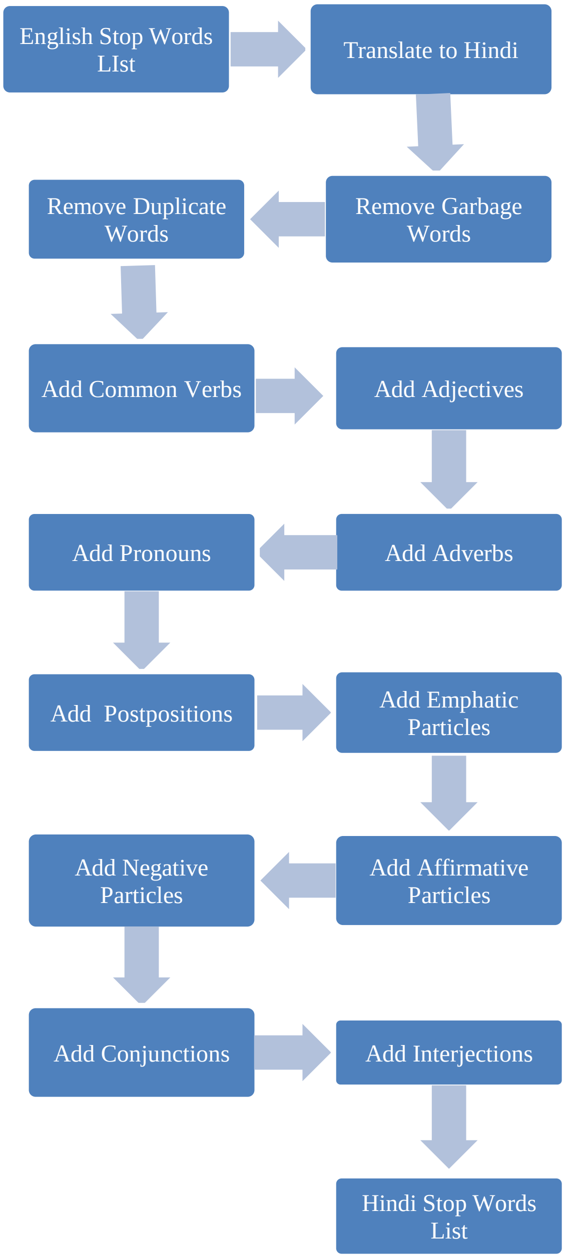

Figure 1 Details of stop word list construction for Hindi

6.1Evaluation of list

A very small list of 97 stop words is available at forum for information retrieval evaluation (FIRE) (http://www.isical.ac.in/~fire/data/stopwords_list_hin .txt).

Another list of 225 stop words is available on the website (https://github.com/stopwords-iso/stopwords-hi/blob/master/stopwords-hi.tx)

A list of 299 stop words is available on website of Indian language technology proliferation and deployment centre (http://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=1637&lang=en).

These lists lack in quantity of stop words as well as in the quality of stop words due to unavailability of various possible inflected forms of stop words. In contrast to these lists, we have created a generic stop word list of 800+ words of Hindi language and we have also added the various inflected variants of stop words for completeness.

7.Conclusion

In recent times there has been a considerable amount of effort to develop IR systems for languages other than English. A frequently used resource in IR is the stop word list. Till now there is no standard stop word list which has been constructed for Hindi language and whatever list is available is very limited and incomplete. In this paper, we have presented an approach to construct a generic stop word list for Hindi language. We have constructed a large stop word list for Hindi language and our list contains more than 800 stop words. We hope that this list would be beneficial to researchers working with Hindi documents.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Sebastiani F. Machine learning in automated text categorization. ACM Computing Surveys. 2002; 34(1):1-47.

[2] Luhn HP. The automatic creation of literature abstracts. IBM Journal of Research and Development. 1958; 2(2):159-65.

[3] Fox C. A stop list for general text. In SIGIR forum 1989 (pp. 19-21). ACM.

[4] Hart GW. To decode short cryptograms. Communications of the ACM. 1994; 37(9):102-8.

[5] Rijsbergen CJV. Information retrieval. London: Butterworths; 1979.

[6] Fox C. Information retrieval data structures and algorithms. Lexical analysis and stoplists. Prentice Hall; 1992, p.102-30.

[7] Yang Y. Noise reduction in a statistical approach to text categorization. In proceedings of the international ACM SIGIR conference on research and development in information retrieval 1995 (pp. 256-63). ACM.

[8] Chekima K, Alfred R. An automatic construction of Malay stop words based on aggregation method. In international conference on soft computing in data science 2016 (pp. 180-9). Springer Singapore.

[9] Amarasinghe K, Manic M, Hruska R. Optimal stop word selection for text mining in critical infrastructure domain. In resilience week 2015 (pp. 1-6). IEEE.

[10] Na D, Xu C. Automatically generation and evaluation of stop words list for Chinese patents. TELKOMNIKA (Telecommunication Computing Electronics and Control). 2015; 13(4):1414-21.

[11] Medhat W, Yousef AH, Korashy H. Egyptian dialect stopword list generation from social network data. Egyptian Journal of Language Engineering. 2015; 2(1):43-55.

[12] Jha V, Manjunath N, Shenoy PD, Venugopal KR. HSRA: Hindi stopword removal algorithm. In international conference on microelectronics, computing and communications 2016 (pp. 1-5). 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 in unsupervised and statistical keyword/keyphrase 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

Aditi Sharan is an Assistant Professor at the School of Computer and Systems Sciences, Jawaharlal Nehru University. Her research interests include Text mining, IR and NLP.

Email: aditisharan@gmail.com