Stop Word Removal of English Text Documents Based on Finite Automata

Shraddha K. Bhirud[1] Komal D. Bhagvat[2] Atul P. Marathe[3] Jaypal A. Rajput[4] Harshal R. Kotwal[5]
[1,2,3,4]Research Scholar [5]Assistant Professor
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]SSBT Collage of Engineering and Technology Jalgaon, India

Abstract— In IR(information retrieval systems), Web Mining, Artificial Intelligence, Natural Language Processing, Text Summarization, Text and Data Analytic systems, optimization of text data becomes very important. One of the preprocessing step is stop word removal. Some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded. These words are called stop words. In the Information era, optimization of processes for Information Retrieval, Text Summarization, Text and Data Analytic systems becomes utmost important. Therefore in order to achieve accuracy, extraction of redundant words with low or no semantic meaning must be filtered out. Such words are known as stopwords. Stopwords list has been developed for languages like Sanskrit, Chinese, Arabic, Hindi, etc. Stopword list is also available for English language. A large number of available works on stop word removal techniques are based on manual stop word lists. An efficient stop word removal technique is required. In this paper, we are proposing a stop word removal algorithm for English Languages. Which is using the concept of a Finite Automata (DFA). Then pattern matching technique is applied and the matched patterns, which is a stop word, is removed from the document.

Keywords: Information Retrieval (IR), Natural Language Processing (NLP), English, Stopword, Tokenization

I. INTRODUCTION

Stop Words: A stop word is a commonly used word (such as the, a, an, in) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.We would not want these words taking up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to be stop words. In IR (information retrieval systems), Web Mining, Artificial Intelligence, Natural Language Processing, Text Summarization, Text and Data Analytic systems, optimization of text data becomes very important. One of the preprocessing step is stop word removal. Some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded. These words are called stop words.

Preprocessing of textual information leads to prepare data for core text mining operations. It filters out noise data from text. Stop words removal is one such method of preprocessing where frequently appearing words conveying little or no meaning are eliminated. Stopwords are words which frequently appear in text do not possess any important semantic relations. For e.g. in English language words like the, in, that, those, for, of, and are considered as stopwords as they does not account for any key role apart from grammatical formations. Stopwords are also known as function words. Stop word removal techniques are required in many NLP activities like Information Retrieval systems wherein the words are indexed which on removal of stopwords decreases indexing space. Removal of stopwords from corpus also leads to its decreased size which increases efficiency of any NLP activities. The English stopword list generated from this implementation will serve different NLP systems developed in future. Like other natural languages, English due to its rich grammatical features and being mother of most Indian languages, it enjoys distinguished place in research domain like machine translation.

II. RELATED WORK

Stop words are most common words found in any natural language which carries very little or no significant semantic context in a sentence. It just carry syntactic importance which aid in formation of sentence. As a preprocessing operation it must be removed to ease further task and speedup core task in text processing.

Preprocessing is an important step in the data mining process. It transforms the raw data from original data source to a format that will be more easily and effectively processed further. This preprocessed format is suitable for employing different types of feature extraction methods. It includes cleaning, normalization, transformation, feature extraction and selection etc. Text Processing methods are based on the recognition and derivation of depictive features for documents in natural languages. In order to identify and extract the features, it is necessary to filter out the word or phrases that is not important from the important text. These not so important words, which are evenly distributed in a document, are the most frequent words of any language. These are known as functional words or stopwords. For example, Some of the stopwords are whe, you, yours, do. Stopwords have been identified as not important since the earliest days in Text Processing tasks. These words are very frequent and have no additional meaning to the actual content and meaning of a document. These words can be categorized into several word-groups like prepositions, conjunctions, adverbs etc. Removal of these words saves a lot of processing time andmemory space and does not damage information retrieval effectiveness. Stopword removal comes under data cleaning part of preprocessing. These removal can reduce text count in the document corpus by 35-45%.

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Fig. 1: Stop Word Removal

Pattern matching method for removal of stopwords. This method uses a dictionary to store stopword lists. The document corpus is searched for the words present in this dictionary and when words are matched, they are removed from the document corpus. This method requires a lot of space to store the dictionary and too much time in searching the word in the document corpus. Our method does the same task of stopword removal in much more efficient way by the implementation of DFA.

III. PROBLEM STATEMENT

The document containing English text is parsed at sentence level and for each sentence in the document, it is parsed at word level. For each word, DFA STOPWORD() function is called. The output returned by the function is checked for true or false, boolean values. If it returns false then the word is not a stopword and appended in the document. If it returns true then the word is a stopword and removed from the document.

Function DFA STOPWORD() illustrates a faster method for stopword removal using DFA implementation. It considers json object which acts as input to algorithm. Json object consists of ve parameters namely states, character, transitions,start state and accepting state. For each character, initial conditions are checked. If current state is not in dened transitions then it returns false. At any point of time, if DFA encounters a character that is not in transitions then it returns false. After all the above mentioned conditions are met, DFA transits to the respective state and updates the current state to new state. If current state is an accepting state then DFA returns true else returns false. True indicates that the word is a stopword and false indicates that it is not a stopword.

An efficient stop-word removal technique is needed in many natural languages processing application such as: spelling normalization, stemming and stem weighting, and in Information Retrieval systems (IR). Most TC systems remove the stop- words, and many systems perform a much more aggressive filtering, removing 90 to 99 percent of all features. The elimination of stop words also reduces the corpus size typically by 20 to 30 percent which leads to higher efficiency. The general trend in IR systems has been the use of quite large stop lists (200300 terms) - due to morphological richness of the language; the list contains all possible morphological variants of each stop-word- to very small stop lists (712 terms) to no stop list whatsoever. For example, in English articles the propositions such as the, on, and with are usually stop words. Stop-words may also be document-collection specific, for example, the word blood would probably be a stop word in a collection of articles addressing blood infections, but certainly not in a collection describing the events of World Cup. Subsequently, many words that occur frequently are eliminated. Eliminating such words from consideration early in automatic indexing speeds processing, saves huge amounts of space in indexes, and does not damage retrieval effectiveness. Two related facts were noticed in the early days of IR. First, a relatively small number of words account for a very significant fraction of all texts size. Words like ”IT”, ”AND”,”THE” and ”To” can be found in virtually every sentence in English-based documents. Secondly, these words make very poor index terms, with which users are indeed unlikely to ask for documents.

- Stopwords are the words with low discrimination power.
- The specific nouns, verbs or other grammatical types could be having less candidature for being stopwords and the elements like articles, prepositions, and conjunctions are usually present in a stopword list.
- Stopwords serve only a syntactic function and never have any predictive capability. They do not indicate the subject matter.
- They have a very high frequency so they can affect the efficiency of the information retrieval process.
- They affect the weighting process as stopwords are tend to diminish the impact of frequency differences between less common words.
- The document length can be changed by the removal of the stopwords and affects the weighting process.
- The fact that if they carry no meaning, they can also affect the efficiency, resulting in a large amount of unproductive processing.

IV. PROPOSED ENGLISH STOP-WORD LIST

English is very rich in lexical tokens that means stop-words are available in big quantities. Stop-words in English have certain properties.

- They have no meaning if they are used separately
- Appear many times in a text.
- Necessary for the construction of the language.
- Mostly adjectives.
- General words and not particularly used in a certain field.
- Not used as a search keyword.
- Never form a full sentence when used alone.
- Stop-words in English include some of grammatical links such as the definite article (AL)(the), attached and separate prepositions, conjunctions, interrogative words, negative words, exclamations and calling letters, adverbs of time and place, also they include all the pronouns, demonstratives, subject and object pronouns, the Five Distinctive Nouns, some numbers, additions and verbs. Stopwords may be separate or attached ones in a form of prefixes or suffixes.

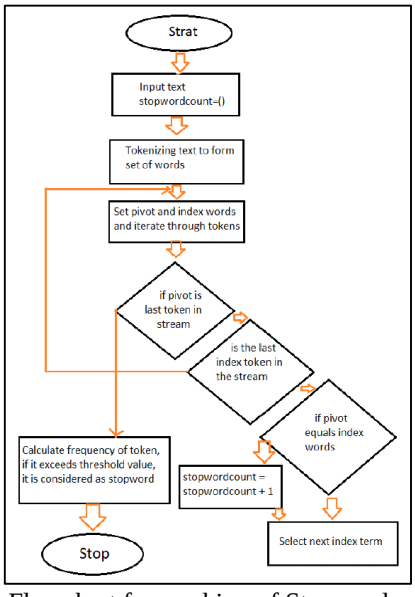

Fig. 2: Flowchart for working of Stopword removal

There exists a general English stop-words list; however, due to the highly inflectional nature of English language those words may come in different forms according to prefixes and suffixes attached to them.

Rank	Word	Rank	Word	Rank	Word	Rank	Word	Rank	Word
1	the	21	this	41	so	61	people	81	back
2	be	22	but	42	up	62	into	82	after
3	to	23	his	43	out	63	year	83	use
4	of	24	by	44	if	64	your	84	two
5	and	25	from	45	about	65	good	85	how
6	a	26	they	46	who	66	some	86	our
7	in	27	we	47	get	67	could	87	work
8	that	28	say	48	which	68	them	88	first
9	have	29	her	49	go	69	see	89	well
10	I	30	she	50	me	70	other	90	way
11	it	31	or	51	when	71	than	91	even
12	for	32	an	52	make	72	then	92	new
13	not	33	will	53	can	73	now	93	want
14	on	34	my	54	like	74	look	94	because
15	with	35	one	55	time	75	only	95	any
16	he	36	all	56	no	76	come	96	these
17	as	37	would	57	just	77	its	97	give
18	you	38	there	58	him	78	over	98	day
19	do	39	their	59	know	79	think	99	most
20	at	40	what	60	take	80	also	100	us

Fig. 3: English stop-word list

V. METHODS OF STOPWORD REMOVAL

Following are the most commonly used techniques for the removal of stopwords from a text: The Classic Method: This method is used to remove stopwords obtained from pre compiled lists.

Methods based on Zipf's Law (Z-Methods): Three stopword creation methods are used in addition to the classic stoplist. This includes removing most frequent words (TF-High), removing words that occur once, i.e., singleton words (TF1), and removing words with low inverse document frequency (IDF).

The Mutual Information Method (MI): It is a supervised method that is used by computing the mutual information between a given term and a document class (e.g., positive, negative), providing a solution of how much information the term can tell about a given class. Low mutual information suggests that the term has a low discrimination power and it should be removed consequently.

Term based Random Sampling(TBRS): This method was first proposed by Lo et al. in which the stopwords are detected manually from web documents. This method is used by iterating over randomly selected separate chunks of data and ranks terms in each chunk based on their in-format values using the Kullback-Leibler
divergence measure as shown in the following equation:
$dx(t) = Px(t).log2Px(t)/P(t)$
where,
Px(t) is the normalized term frequency of a term t within a mass x, and
P(t) is the normalized term frequency of t in the entire collection. The final stop list is constructed by taking the least informative terms in all chunks by removing all possible duplications.

VI. APPROACH USED TO REMOVAL STOPWORD

A dictionary based approach is been utilized to remove stopwords from document.
The algorithm is implemented as below given steps.

1) Step 1: The target document text is tokenized and individual words are stored in array.
2) Step 2: A single stop word is read from stopword list.
3) Step 3: The stop word is compared to target text in form of array using sequential search technique.
4) Step 4: If it matches, the word in array is removed, and the comparison is continued till length of array.
5) Step 5: After removal of stopword completely, another stopword is read from stopword list and again algorithm follows step 2. The algorithm runs continuously until all the stopwords are compared.
6) Step 6: Resultant text devoid of stopwords is displayed, also required statistics like stopword removed, no. of stopwords removed from target text, total count of words in target text, count of words in resultant text, individual stop word count found in target text is displayed.

VII. WORKING OF DFA

For each word it start from current state ,it changes the states to intermediate state to accepting states ,if it is a stop word it will return true otherwise false. If it is a stop word it will remove it from main document and display the result with a count of number of stop words removed. It considers json object which acts as input to algorithm. Json object con-sists of five parameters namely states, character, transitions.

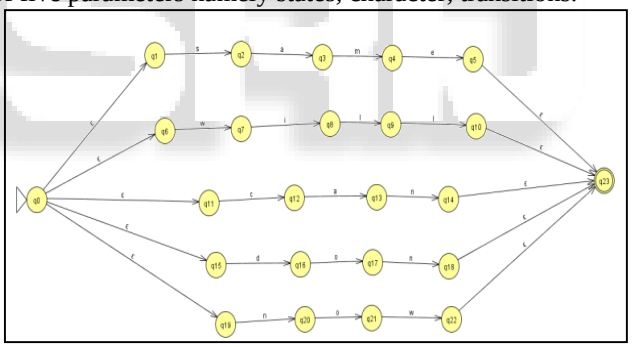

Fig. 4: Sample DFA for stop word removal

VIII. SIMULATION RESULTS

A. User

First registration is performed in user module.If the registration is completed Successfully ,then the login process is performed by the user with the user id and password,if it is valid then the user id and password,if it is valid then the user is login successfully,if not the error will occur.

B. Manage Text

In manage text,first user add the word. Second he add the paragraph. Third view paragraph then in first block user see the main paragraph, second block show the find stop word and third block show the word after removing the stop word.

C. Report

It shows the status about the activities or operations which are performed by the users. It shows the status such as original

paragraph, removed and stop words which are performed by the user.

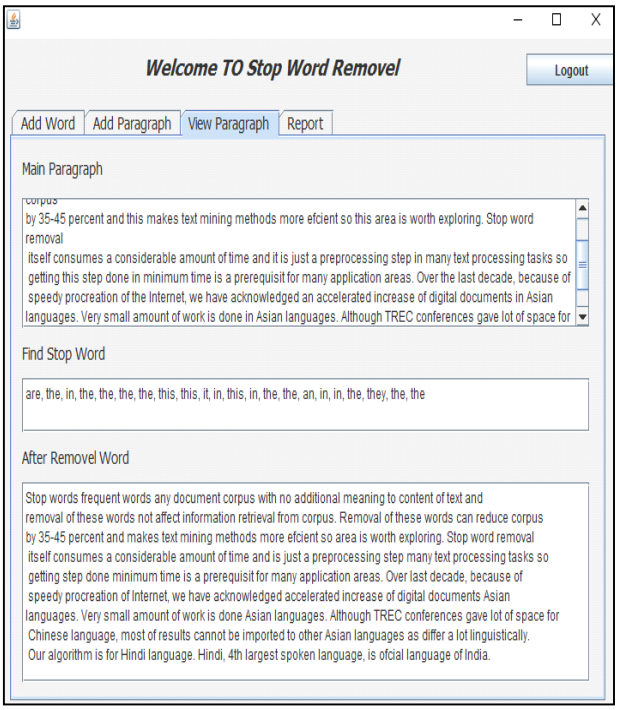

Fig. 5: showing results after stop word removal

IX. Conclusion

A stopword removal algorithm is implemented for English Language which is based on DFA. It takes English documents as an input and returns the document with stopwords removed from them. Most of the existing stopword removal techniques uses the pattern matching method for removal of stopwords. Pattern matching method uses a dictionary to store stopword lists. The document corpus is searched for the words present in this dictionary and when words are matched, they are removed from the document corpus. This process requires a lot of space to store the dictionary and too much time in searching the word in the document corpus.

References

[1] F. Zou, F. L. Wang, X. Deng, S. Han, and L. S. Wang, Automatic construction of chinese stop word list, in Proceedings of the 5th WSEAS international conference on Applied computer science, 2006,pp. 1010 1015.
[2] Jaideepsinh K. Raulji, Jatinderkumar R. Saini, Stop-Word Removal Algorithm and its Implementation for Sanskrit Language International Journal of Computer Applications (0975 8887) Volume 150 No.2, September 2016.
[3] Basim A and Mohammad A, Hybrid Stop-Word Removal Technique for Arabic Language, Egyptian Computer Science Journal, Vol-30 No-1, Jan 2008.
[4] NLTK Edward Loper and Ewan Klein(2009),NLTK Book.
[5] J. Savoy, "A Stemming Procedure And Stopword List For General French Corpora", Journal of the American Society for Information Science, 50(10), 1999, 944-952.
[6] K.T. Lua, and G.W. Gan, An application of information theory in Chinese word segmentation, Computer Processing of Chinese and Oriental Languages, 8(1):115-124. 1994.
[7] Z. Yao and C. Ze-wen, Research on the construction and filter method of stop-word list in text preprocessing, in Intelligent Computation Technology and Automation (ICICTA), 2011 International Conference on, vol. 1. IEEE, 2011, pp. 217221B. F. William and R. Baeza-Yates, Information retrieval: Data struc-tures and algorithms, ISBN-10, vol. 134638379, 1992.

