Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 739-746
Received 25 April 2021; Accepted 08 May 2021.

The Effect of Stop Word Removal and Stemming In Datapreprocessing

Rama kalaivani .E[1] , Ramesh Marivendan.E[2]
1 AP , Department of CSE , Karpagam college of Engineering,Coimbatore-32.
(ramakalaivani.e@gmail.com)
2 AP,Department of ECE,Er. Perumal Manimekalai College of Engineering,Hosur-17.

Abstract: *Preprocessing is one the important technique in text mining and its application. It helps in converting the original textual data to most significant text features. The main objective of preprocessing is to split the sentences into words using whitespace as the separator. Tokenization is applied for each documents and special character is removed. The words are then filtered and stemming process is followed. Filtering is removal of words which are of less importance. The process of removing words such as prepositions, articles and conjunctions are known as stop word filtering. Stemming algorithm is used to convert different words form in to simple canonical form. It involves a set of procedure where all words with the same root are reduced with the same root to a common form. Porter stemmer method is applied to improve the efficiency of data* preprocessing. *In this paper, algorithms of stop word removal and stemming are used which mainly helps in improving the accuracy of clustering and classification techniques. In other words, the data preprocessing techniques helps in decrease of overall size of data set in storage space and time.*

Introduction:

Preprocessing is divided in to two steps namely morphological analysis and syntax and semantic analysis. In morphological analysis there are three sub categories namely tokenization, filtering and stemming. Data pre-processing step is important in data mining process mainly used for indexing of documents, where documents can be identified as transactions. This phase also effectively denotes the document in terms of space and time. Preprocessing also leads to efficient process in representing documents as a set of index terms. It enhances the accuracy between the word and document and also relevancy between category and word. It is the most critical and tedious process where data's are purified. It requires words and the ending of documents. Identifying words and separating them is known as tokenization. It includes word splitting, special character removal, It separates the input alphabet in to word character (word splitting) and word separators(special character removal).For example digits used in a document should be ignored, Punctuation marks are treated as separators and letter case are often eliminated. This leads to division of words in to nouns and verbs. This helps in analyzing the words exactly. It is important to index the text in to data vectors. A new approach called bag –of-words approach is also used in implementing the algorithms.

Literature Survey:

The process may be automated and manually done. Stop word makes the text looks heavier and it plays minor role for analysts. The main objective is to find out the word that often occurs in a document. Those words have no value for retrieval purposes. Stop wordsincludesarticles(a,an,the),preposition(in,on,of), conjunction(and,or,but,if),pronouns(I,you,them)and some more verbs,nouns,adverbs and adjectives. It increases the size of indexing structures Most frequently used words are included. In information retrieval and texts mining some words in English are found to be useless. Those words are considered as stop words. The first step is to identify the most frequently used words in English. Fix up a cut off point for the list. Identify number of occurrences and allocate ranking for the particular word. Prepare a large list such that useless words out of index items are maintained in the list. The data's in the list is also based on the parts of speech. If there are different versions of verb are counted as same word in the list. Based on the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 739-746
Received 25 April 2021; Accepted 08 May 2021.

root of contractions, the counts of words are tallied in the list. The next step is to fill up the potential index terms. They are pulled up from the indexes. Thus the important items are identified. Many of the most frequently used traditional word were also added if they have not reached the cutoff. The count of certain traditional words may not reach the cutoff value, but those words have identified and thus added a bit more fluf.

To tokenize the input stream, a recognizer named minimal DFA is also used. It also recognizes the stop words. In order to increase the efficiency, based on few criteria another set of words can also be added to the stop list. Adding the letter for which the words in the list start with the letter.This was drawn from abroad range of literature in English. The word list be classic and used abroad. In extraction of stop word list automatically the frequency is calculated. If the frequency is higher it can be related with higher noise. These are the various methods that can be related with document frequency entropy calculation.Guel suggested a different method in computing the probability where word occur in all sentence which includes word occurring in corpus.Nextincludes computing the entropy of each probabilities and selecting the stop word based on the entropy..Tsz Wai Loetal declared various approached called text based random sampling method, included by various experiments. Tzu Wai Lo et al used various approaches called text based random computing method which is based on Kull back Leibler divergence measure. Zhou et al applied program flows control to avoid the Chinese word containing English letters or symbols used in maths.

The stop word list of artificial construction is found to be easy .It does not compare the current key which do not have pertinence to various text documents while working .The extraction of stop word list found to be more potential and flexible.

The stop word techniques are used in information retrieval .There are only limited number of words available for accounts in all texts size.Each English based documents comptrises words like IT ,AND ,THE and TO.The index terms were found to be poor.The facts were introduced by Luhn.The words with highest frequency are considered as noise word.It can also be named as common word or stop word.The difference between non keywoard terms can be identified.The words that does not have major in describing document contents is used in automatic indexing.When a search is made for these words ,the items were reached in the database, the discrimination value is low. These words are typical with around 20-30% of tokens in documnets. When these words are eliminated, the large quantity of space is saved. Information Retrieval System was not affected. The general identification of stop word is that stop word always occur in high frequency with low discrimination and can be filtered from Information Retrieval System. The most preliminary condition for knowledge discovery was to extract document feature vector, to delete all stop words.

Stop word identification-Principles and Methods

The text cannot be characterized in stop word. When such situations occur it is characterized as stop word. It cannot be directly mapped as judgment and the computer programs cannot modify the word whether it characterizes the text. There are two aspects in which the accuracy and efficiency can be improved.They are1)If more number of stop words were deleted,the accuracy of web content mining should not be reduced.2)If suppose the stop words were deleted the dimensionality of text feature space would be reduced..It means the the capability of distinguishing one article from the remaining article in dataset.The threshold may be fixed below range or above range that can be labelled.The threshold is fixed to identify the set of sample words in document collection.ManRijibergan defined a classic list of 250 stop words in English that be often used as a basic endline in text dataset

Calculation of stop word list: The stop word which already defined was calculated by combining the classical stop word list the classical stop word list with stop words depends on various domain for text

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 739-746
Received 25 April 2021; Accepted 08 May 2021.

document corpus. The following are the principles that are followed.

a) Stop word list of Fox's classic was used in SMART

b)Punctuation,Arabic numerals zero(0)-nine(9)

c)Greece Letter Mathematical Symbols

Stop word Filter algorithm:

For improving the performance of text mining it is necessary to filter the required terms If suppose when a stop word is stored in the disk it requires reading a disk for filtering each of the segment results in each text doxument.This leads to lots of time. To improve efficiency it reads stop word list in the memory. So every operation can be done in memory .Various measures are considered to improve efficiency.

Sequence Filter Algorithm:

The initial step is to retrieve the word term from linked list TERMS.With that compare stop—word in the stop-word listThe following ere the steps used insequence filter algorithm.

Input:STOP LIST(Stop word linked list)

TERMS={T_{ij}} Term Linked list
Output:TERMSFILTER ={TF_j}
Sequence filter includes both STOPLIST & TERMS

Step 1:Select any one term named T_i from TERMS

Step 2:Select any one term named Stop-word $Stop_j$ from the STOPLIST;
Step 3:CompareT_i from TERMS with $stop_j$ from STOPLIST.(T_i==$Stop_j$).If it is equal T_i is assumed as stop word and hence delete it,move to step 5.
Step 4: If STOPLIST found empty,If Yes Ti is not assumed as stop word.It is stored in TERMSFILTER if it no.move to step 2.
Step 5:If TERMS is empty ,if is false or no move to step 2.
Step 6:Return TERMSFILTER Filter ends.

Most Recently Used-Filter Algorithm:

The initial step is used to identify a term from linked list TERMS and later it is compared with

each stop word in the stop word linked list STOPLIST.If it is a stop word add or insert it to the head of STOPLIST and delete it from the previous position. The following are the steps involved in the algorithm.

INPUT :(Stop-word linked list)

OUTPUT: Results obtained in linked list TERMSFILTER={TF_i)

MRU filter comprises STOPLIST and TERMS.

Step 1:Select an term T_i from TERMS

Step 2:Select a term Stop j from stop list

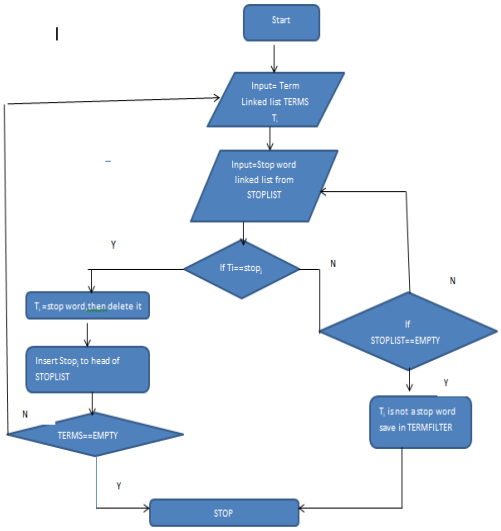
Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 739-746
Received 25 April 2021; Accepted 08 May 2021.

Step 3:Compare if T_i from TERMS and Stop j from stoplist are equal .If it is true or comparison is Yes,$T_i$is stop word. Delete T_i ,insert tor add stop j into the head of STOPLIST and delete stopj if ,then up going position move to step 5.

Step 4:If STOPLIST is empty, if it is true T_i is not assumed as to 2.

Step 5:If TERMS are found to be empty, move to step 6 or if it is not empty move to step 1.

Step 6:Return TERMS FILTER .Filtering ends.



Flowchart for MRU filter Algorithm

Hash Filter Algorithm:

Identify hash function F.It is helpful to compute the hash code .It calculates for every stop word available in the list, which helps in returning the remainder that is considered as the index of the stop word. Using hash function F, the index of the term is calculated. The search is done using terms index in the hash filter. If there exists the TERM already it is a stopword,if it does not exists NULL is returned. If hash collision occurs stop word is put in to the head of the linked list.

HASHTABLE Construction Algorithm:

INPUT: Stop word linked list STOPLIST,SEED,SIZE of hashtable)
OUTPUT: Hashtable

Hashstoplist (STOPLIST, SEED, SIZE);
Step 1:Get a stop from STOLIST;
Step2:Compute the Hashcode;$Stop_j[0]$*SEED^(n-1)+$Stop_j[1]$*SEED ^ (NH2)+……+$Stop_j[n-1]$=Hashcode and a value V is mapped to $stop_j$;

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 739-746
Received 25 April 2021; Accepted 08 May 2021.

Step 3: Evaluate Hashcode mod size and retrieve index.

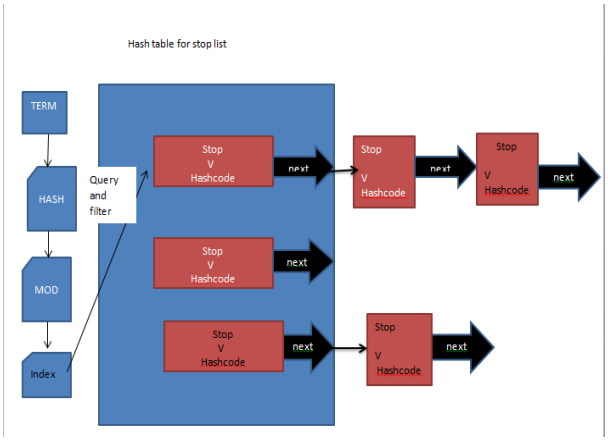
Step 4:Check whether the Hashtable at Index;

If it is NULL,it can be Saved as ($Stop_j$,V,Hashcode,Next=Previous Node) at the linked list which is started from HASHTABLE(Index);

Step 5:If STOPLIST is found empty?If STOPLIST not empty move to switch (1);

Step 6: Return Hash Table
The next algorithm is Hash Filter Algorithm which is used to retrieve TERM FILTER.

INPUT:Term linked list TERMS={Ti), HASHTABLE
OUTPUT:Output can be linked list TERMSLIST={TFi}
HASHFILTER (TERMS,HASHTABLE);
Step 1:ChooseTifromTERMS;
Step 2:Compute the hashcode :Ti[0]*Seed^(n-1)+Ti[1]*Seed^(n-2)+……+Ti[n-1]=hashcode;
Step 3:Evaluate Hash code and size then retrieve index
Step 4:The linked list is searched in HASHTABE at Index: If it exists Ti is stop word and delete it.
Step 5 :If TERMS are empty move to step 6 if not empty move to step (2)
Step 6:Return TERMSFILTER



Hash Filter Diagram

The following are the methods used in stop word removal
Classic Method :Methods based on zipfs law.(removing most frequent words).It includes two types .First type is removing words that occur most frequently(TF-High) and removing words that occur once (TFL). It is called as low inverse document frequency.(IDF).First step is ranking the terms in each database based on the frequency using IDF method. The next step is involves plotting of rank frequency distribution for the ranked terms.
Mutual Information method :It is termed as a feature selection routine. Features do not contribute on correct classification decision, it is considered as stop word and removed from the feature space consequently. It also gives information about the term that tells about a given class. Since it has low discrimination power, it is removed easily.Term based random sampling (Iterate separate chunks of data and performs ranking).This method identifies stop words automatically form web documents

Stemming:

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 739-746
Received 25 April 2021; Accepted 08 May 2021.

Stemming is used to reduce the number of different terms and reduce words to their stem.. In morphological analysis the root of a word can be used as clues to grammatical structure. It is one of the important features in word indexing and search system. The main objective is to increase the recall through automatically handling of word endings. Stemming process helps in increasing the word purity. Thus it contains plain and actual meaning. The attached suffixes and prefixes are removed in stemming. The exact meaning of the word remains same, although it gets an addition to suffix.Thoses words are considered as new word, there is a change in spelling. This process helps to improve the retrieval system.It is also called over stemming. If two words that are stemmed to the same root a false negative is occurred. It is known as under stemming. It is proved that heavy stemmers increases over stemming errors in turn it reduces the under stemming errors.

Methods in stemming:

Truncating Methods: It includes the removal of suffixes or prefixes of a word. In stemmers if a word is truncated at n^{th}symbol,"n" letters are maintained and others are removed. The words less than are not modified. In S-stemmers, the suffixes in plural are removed to convert in singular forms.

Levin Stemmer: It handles removing longest suffix, it also frames tables for converting the stemmers in to valid words. It applies single pass algorithm in removing a suffix from word. It performs faster and data consuming. It has vast amount of space for time and removal of suffix.

Corpus Based Method:

Automatic modification of conflation classes may be performed in this method. The words that result in common stem suits the characteristics of text corpus .It is found that words that are conflated for corpus may co occur in documents from that corpus Inflectional and Derivational Thus in stemming the morphological part of a word is converted in to stems. It is not necessary that the stemming should be a word in dictionary, but it should it should map to all variants. Words with different meaning are placed separately. Morphological words always have base meaning and can be mapped to the same stem. It is known as language processing application. It can be referred as recall-enhancing device. It is heuristic process where it includes in removing the derivational affixes. The example of stemming is Derivation, Deriving, Derives, and Derive.

Porter Stemmer Algorithm:

The algorithm includes

Step(1)i)Mapping with rule SSES→ SS.Similar to CARESSESS→ CARESS.
Mapping IES → I ,This example can be related to Ponies → Pon
Mapping SS → SS .This example can be related to caress → caress.
Mapping S→ .This example can be related to cats→ cat.
Step(1)ii) If (m>0) ,EED → EE.The example is Feed → feed
(*V*)ED → Plastered → Plastered
(*V*)ING → Motoring → motor.
If these rules are successful following examples can also be considered.
AT → ATE ,For This type of mappingthe example is Estimat(ed) → Estimate
BL → BLE,In this type of mapping,the example is Bub(bl)ed → Bubble
IZ→ IZE, The example is Siz(ed) → size.
(*d and not(*L or *S or *Z) → single letter.The examples are
drink(ing)→drink

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 739-746
Received 25 April 2021; Accepted 08 May 2021.

tanned→tan
giving→ giv
missing→ miss
ruling ->rule
The rule for mapping a single letter includes removing one of the double letter pair.
Step (1)iii) (*V*) Y → I ,happy → happi
Sky → sky.
The above all dealt with plurals and past participles.
Step 2:If (m>o) ATIONAL → ATE, example can be Motivational → Motivate.
(m>0) TIONAL→ TION, Operational → operation
(m>0), ANCI→ ANCE, Balanci→ Balance
(m>0),ABLI → ABLE,Knowledgabli → Knowledgable
(m>0),OUSLI → OUS,Ambiguosli→ ambiguous.
(m>0),IZATION → IZE,Organization→ Organize
(m>0),ATION→ ATE,Animation → Animate.
(m>0),ATOR→ ATE,Modulator → Modulate
(m>0),IVENESS → IVE,Forgiveness→ forgive
(m>0),ALITI → AL,Casualiti → Casualiti
(m>0),BILITI → BLE,Eligibliti → Eligible
The testing of the string S1 was fast by programming on the penultimate letter for the word.
Step 3: If (m>o) ICATE → IC,replicate → replic
(m>0) ACTIVE → ,Formative → form
(m>0) ALIZE → AL, Modularize → Modul
(m>0)FUL → ,Joyful→ Joy
Step 4: (m>1) AL → revival → reviv
(m>1) ANCE → Allowance→ allow
(m>1) ENCE → Inference → Infer
(m>1)ATE → Activate → acti
(m>1)IVE → Effective→ Effect.
The suffixes can be removed.
Step 5)i) (m>1)E → Probate → Probat
rate→ rate(m=1 and not *0)E → Cease → Ceas
Step 5)ii) (m>1 and *d and *L) → Single Letter. Controll→ Control
roll→ roll

Stemmer Strength:

Stemmer strength can be defined as a degree where stemmer changes its words to stem is called stemmer strength. Stemmer the handling with few suffixes and merging with highly related words are termed as weak or lightstemmer.

There are various criteria for measuring the strength of the stemmer.

The mean size of words of conflation group. It identifies theaverage count of words in conflation group that are transferred to the same stem. Connected can be stemmed as words connect, connected and connecting .The conflation class size is assumed as three. if the stemmer is found to be stronger stemmers will have higher words for conflation.

Index compression factor(ICF)-

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 6, 2021, Pages. 739-746
Received 25 April 2021; Accepted 08 May 2021.

Index Compression factor can be defined as ICF=n-s / s n' is the number of words in the corpus and 's' assumed as the number of stems. Index Compression Factor found to be the fractional reduction in index size that can be gained through stemming**.** The index compression factor for 50,000 words(n) and 40,000 stem(s) was found to be 20%.The stemmers which are stronger will have higher index compression factors

The number of words and stems which differ:

Stemmers always make remain unchanged.eg: the root words like "engineer" cannot be altered but often stronger stemmer words will have changes than weaker stemmers words.eg:wander to wand ,authority to author

Conclusion:

The techniques used in data preprocessing method helps in removing noisy, unwanted data. Thus, it increases efficiency, accuracy in various clustering and classification process. Thus it reduces the volume of database and helps in removing unnecessary data.Hence the data mining techniques also increases the recall rate.

REFERENCES:

[1] "A Query Formulation Language for the Data Web" - IEEE Transactions on Knowledge And Data Engineering, Mustafa Jarrar and Marios D. Dikaiakos, Member, IEEE Computer Society-May-12.
[2] The History of Information Retrieval Research"-Proceedings of the IEEE,Mark Sanderson and W. Bruce Croft-May-12.
[3] "Concept-Based Indexing In Text Information Retrieval" International Journal of Computer Science & Information Technology (IJCSIT), FatihaBoubekeur and Wassila AzzougFeb-13.
[4] "Design and Development of a Stemmer for Punjabi" International Journal of Computer Applications, Dinesh Kumar, Prince Rana-Dec-10.
[5] "Stemming Algorithm to Classify Arabic Documents" Symposium on Progress in Information & Communication Technology, Marwan Ali.H. Omer, Ma shi long-2009.
[6] Jennifer .P, Kannan Subramanian., "Retrieving the Personal Photos in Web Data" in International Journal of P2P Network Trends and Technology (IJPTT) – Volume2 Issue3 Number1 May 2012.