IJETAE
EXPLORING RESEARCH AND INNOVATIONS

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 6, June 2017)

A Comparative Analysis of Various Stemmers in Vector Space Model for Efficient Text Retrieval

Meghna Utmal[1], Dr. R. K. Pandey[2]

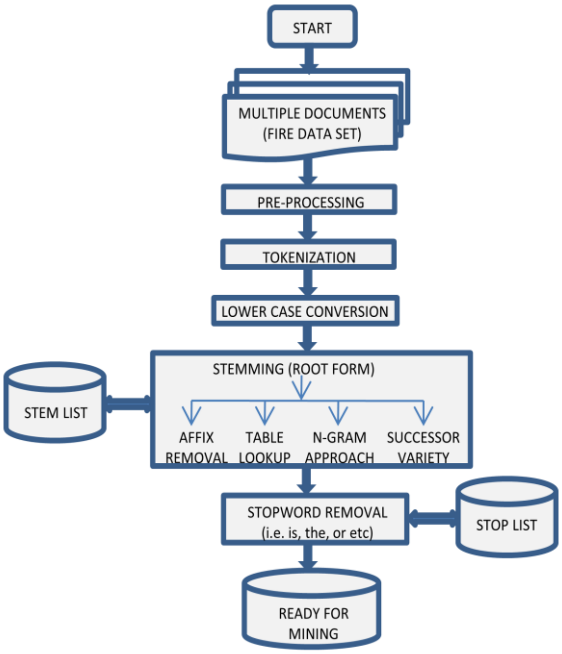
[1]Research Scholar, [2]Professor, UICSA, RDVV, Jabalpur, India

***Abstract*-- In this paper, different preprocessing steps have been performed for efficient text retrieval using RapidMiner tool for the preprocessing of unstructured dataset. Experiment has been carried out with FIRE Dataset 2007 English Collection from Sports file (an open community based retrieval dataset). Different stemming algorithms have been applied and it was found that Porter algorithm works better with the FIRE Data set. We have also experimented with n-gram and without n-gram method with the same dataset. Our present work focuses on a comparative study and its related results between n-gram and without n-gram in the various stemming algorithms. The conclusion of our work focuses on the fact that out of the four stemming algorithm Porter algorithm is the most efficient for English text.**

***Keywords*-- RapidMiner, FIRE, n-gram, stemming , data mining**

I. Introduction

The process of retrieving valuable information from an unstructured text taken from different sources is known as text mining[1]. Dealing with text data which is in unstructured format is a tedious task, hence preprocessing[2] of text data is necessary. Preprocessing is done before applying any mining technique.



The word Tokenization [3]refers to the method of breaking a sentence into chunks of data such as phrases, symbols, words, keywords, and other elements called tokens. The individual phrases, words etc are called tokens. This process discards the use of some punctuation marks and characters. The tokens become the input for another process like parsing and text mining.It is also used in the lexical analysis in computer science.

Stopwords [4] are those which occurs in 75% of the documents in the collection which is useless for the purpose of retrieval. The elements for list of stopwords are Articles, prepositions and conjunctions. An important benefit of stopwords is that it reduces the size of indexing structure.

Stemming[5] which is a preprocessing step is supported by indexing and search system in many text mining applications by improving recall by automatic handling of words ending and by reducing the words to their root word. Here the prefix and suffix are removed from indexed terms. There are 4 major English collection stemming algorithm

a) Porters
b) Lovins
c) Snowball
d) Dictionary

a) Porters stemming algorithm is the most popular stemming methods proposed in 1980 which is based on the fact that the suffixes in the English language (approximately 1200) are mostly made up of a combination of smaller and simpler suffixes. There are 5 steps and within which rules are applied. If a rule is accepted, the suffix is removed accordingly, and the next step is performed. The resultant stem at the end of the fifth step is returned.

b) Lovins Stemmer was the first popular and effective stemmer proposed in 1968 which performs a lookup on a table of 294 endings, 29 conditions and 35 transformation rules, which have been arranged on a longest match principle. This algorithm removes the longest suffix from a word. Once the ending is removed, the word is recoded using a different table that makes various adjustments to convert these stems into valid words. It always removes a maximum of one suffix from a word, due to its nature as single pass algorithm.

c) Snowball is a detailed framework of stemming designed by Porter whose main purpose is to allow programmers to develop their own stemmers for other character sets or languages. Currently there are implementations for many foreign languages.

d) Dictionary stemmers work quite differently from algorithmic stemmers. Instead of applying a standard set of rules to each word, they simply look up the word in the dictionary and it has been revealed that, they could produce much better results than an algorithmic stemmer.

In this paper we focuses on a comparative study between n-gram and without n-gram [6] in the various stemming algorithms for efficient retrieval in English text collection. The experiments are carried out in the FIRE 2007 dataset collection for English language.

This Paper is comprise of the following sections: Section 2 discuss about the related work, section 3 introduces the basic function of rapid miner moreover it also explains the purpose of using this tool, section 4 discusses the datasets and their file formats, section 5 reports the experiments and analysis of our evaluation results. The paper concludes by explaining why the use of porter stemming algorithm is best in comparison with other algorithms.

II. Vector Space Model

The basic idea in the vector space model is to represent each document as a vector of certain weighted word frequencies. In order to do so, the following parsing and extraction steps are needed [7].

1. Ignoring case, extract all unique words from the entire set of documents.
2. Eliminate non-content-bearing “stopwords” such as “a”, “and”, “the”, etc. For sample list of stop words.
3. For each document, count the number of occurrences of each word.
4. Using heuristic or information-theoretic criteria, eliminate non-content-bearing “high- frequency” and “low-frequency” words.
5. After the above elimination, suppose w unique words remain. Assign a unique identifier between 1 and w to each remaining word, and a unique identifier between 1 and d to each document.

III. Related Work

Marti A. Hearst in his paper discussed some important properties of text and its analysis process moreover he also focused on the difference between text data mining and information retrieval [1].

Zdenek Ceska explores the influence of text preprocessing techniques on plagiarism detection moreover it examines stop-word removal and word generalization. His work also contributed into the influence of punctuation and word-order within N-grams [2].

Ms. Anjali Ganesh Jivani discussed the different methods of stemming and their comparisons in terms of usage, advantages as well as limitation. Also she demonstrated that the main purpose of stemming is to reduce different grammatical forms / word forms of a word like its noun, adjective, verb, adverb etc. to its root form [5].

D’Amore and Mah’s introduced the concept of n-grams by replacing the whole terms with n-grams in vector space model moreover computed the weight for n-gram using the number of occurrences in a document [6].

1. Rapid Miner:

Rapid Miner [8] is a tool written in Java that provides an environment for data mining, text mining , predictive analysis, business analytics and machine learning. It uses a client/server architecture which uses Software as a service on cloud infrastructures.

Features of Rapid miner:

a) Open source community & market place
b) Fully integrated platform focused on predictive analytics
c) User friendly with no programming requirement
d) Support for wide range of structured and unstructured database
e) Functionalities like data preprocessing, visualization, statistical modeling, evaluation and deployment

2. Dataset:

The experiments has been carried out in the data set of FIRE 2007 (http://www.isical.ac.in/~fire/) for English. The stuff contains English sports documents in large scale. The FIRE 2007 sports file comprise of 7615 data item files of which 100 sample data is collected for experiment and analysis.

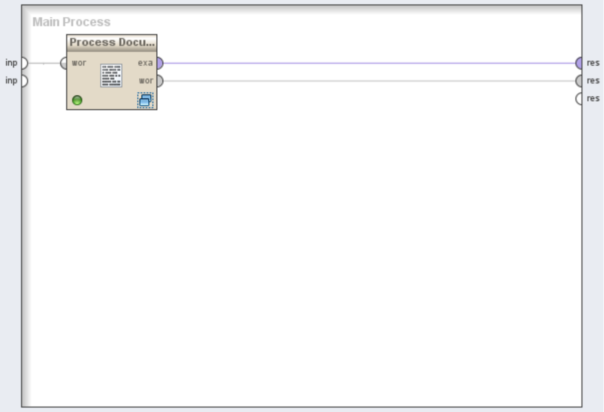
2.1 Document Format

FIRE dataset adapt TREC document style format (http://trec.nist.gov) .Each text document is stored in a separate file for English text collection the document supports the UTF8 encoding system. The document has 3 fields DOC,DOCNO and TEXT. DOCNO is a unique identifier which is assigned to every document in the stuff. TEXT field contains the actual news articles in plain text. The example odf a text file is shown below:

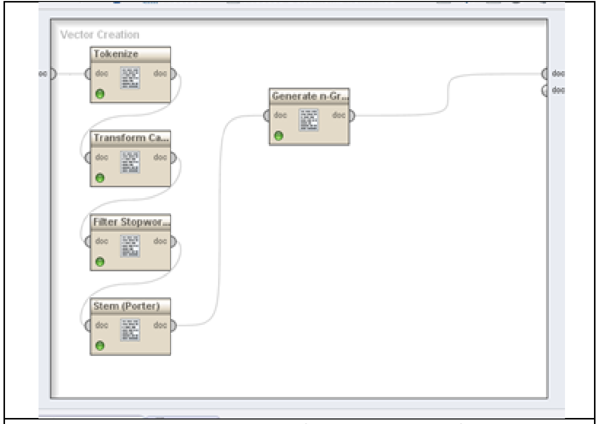
```
<DOC>
<DOCNO>1070101_sports_story_7205760.utf8
</DOCNO>
<TEXT>
The Telegraph - Calcutta : Sports Hussey backs
Moody  Michael Hussey Sydney: Batsman
Michael Hussey wants Tom Moody to succeed
John Buchanan as Australias next coach when
the job becomes vacant after the World Cup in
April, reports said on Sunday. The former
Australian allrounder has been coaching Sri
Lanka for 18 months and been widely credited
with the sides recent good form, including 11
wins from their past 13 one-day Internationals.
I know Tom well, Hussey told a newspaper.
(AGENCIES)
```

IV. EXPERIMENT

Our experiment have been carried out on Rapid Miner 5.3 version .This tools incorporates all features for the analysis on FIRE dataset. Here we used English sports documents text file. The FIRE 2007 sports file comprise of 7615 data item files of which 100 sample data is collected for experiment and analysis. In the given sample dataset the pre-processing technique is performed in the following steps :



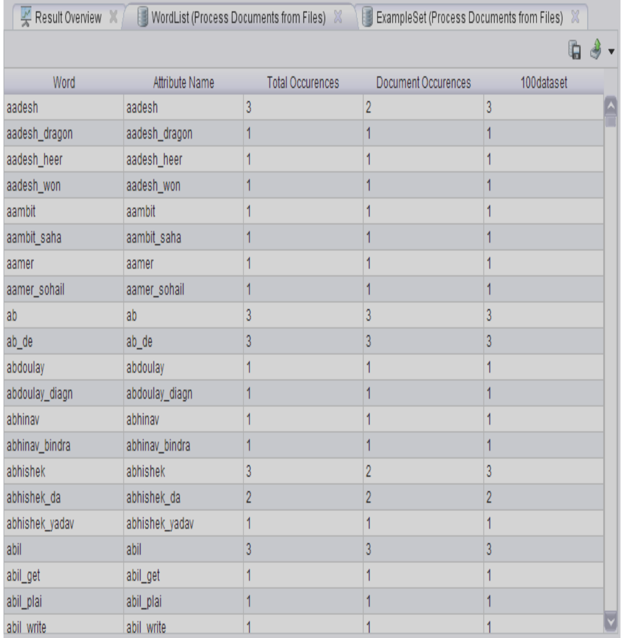
Step 1 : shows the main process called the process document which takes the data stored in the hard disk



Step 2: shows the technique of preprocessing of the sample data. Here the preprocessing comprise of tokenization, transformation, filtering and stemming.

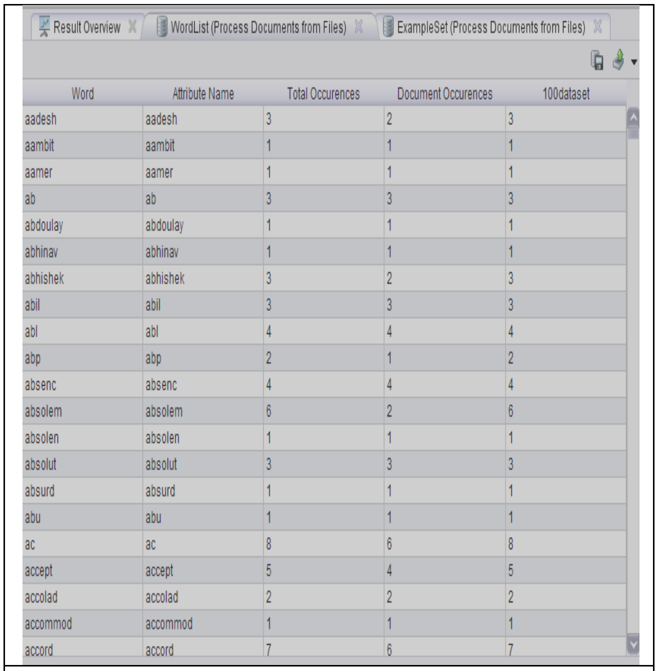
In the preprocessing technique first step involves the creation of main process which is taken in terms of sample data as an input from the hard disk, then tokenization of a document is performed which splits the text of a document into a sequence of tokens. Next step is transformation which transforms all characters in a document to either lower or upper case. Further filtering filters the tokens based on their length (i.e. an, and, or, is etc.). Finally the stemming operator stems English words using the various stemming algorithm applying an iterative, rule based replacement of word suffixes intending to reduce the length of the words until the minimum length is reached (i.e. go for going, name for named etc.).

Now we will generate the characters with n-gram and without n-gram of each token in the document.


Result Overview
WordList (Process Documents from Files)
ExampleSet (Process Documents from Files)

Word	Attribute Name	Total Occurences	Document Occurences	100dataset
aadesh	aadesh	3	2	3
aadesh_dragon	aadesh_dragon	1	1	1
aadesh_heer	aadesh_heer	1	1	1
aadesh_won	aadesh_won	1	1	1
aambit	aambit	1	1	1
aambit_saha	aambit_saha	1	1	1
aamer	aamer	1	1	1
aamer_sohail	aamer_sohail	1	1	1
ab	ab	3	3	3
ab_de	ab_de	3	3	3
abdoulay	abdoulay	1	1	1
abdoulay_diagn	abdoulay_diagn	1	1	1
abhinav	abhinav	1	1	1
abhinav_bindra	abhinav_bindra	1	1	1
abhishek	abhishek	3	2	3
abhishek_da	abhishek_da	2	2	2
abhishek_yadav	abhishek_yadav	1	1	1
abil	abil	3	3	3
abil_get	abil_get	1	1	1
abil_plai	abil_plai	1	1	1
abil_write	abil_write	1	1	1

Step 3: Word list process document with n-gram (where n=2)


Result Overview
WordList (Process Documents from Files)
ExampleSet (Process Documents from Files)

Word	Attribute Name	Total Occurences	Document Occurences	100dataset
aadesh	aadesh	3	2	3
aambit	aambit	1	1	1
aamer	aamer	1	1	1
ab	ab	3	3	3
abdoulay	abdoulay	1	1	1
abhinav	abhinav	1	1	1
abhishek	abhishek	3	2	3
abil	abil	3	3	3
abl	abl	4	4	4
abp	abp	2	1	2
absenc	absenc	4	4	4
absolem	absolem	6	2	6
absolen	absolen	1	1	1
absolut	absolut	3	3	3
absurd	absurd	1	1	1
abu	abu	1	1	1
ac	ac	8	6	8
accept	accept	5	4	5
accolad	accolad	2	2	2
accommod	accommod	1	1	1
accord	accord	7	6	7

Step 4: Word list process document without n-gram

In this preprocessing technique four types of stemming algorithms are involved which supports the English text. These four algorithms are :

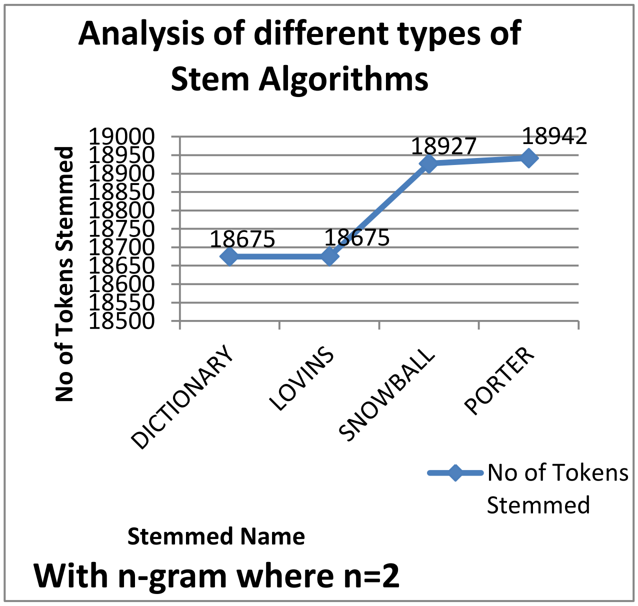
a. Porter Stemming algorithm
b. Lovins Stemming algorithm
c. Snowball Stemming algorithm
d. Dictionary Stemming algorithm

By using these stemming algorithms we find out word count from our sample data set .Of all these the Porter stemming algorithm is the best because it produce the maximum word count with n-gram and without n-gram.

The table given below depicts the number of tokens stemmed in all the four algorithms with n-grams

Stemmed Name	No of Tokens Stemmed
PORTER	18942
LOVINS	18675
SNOWBALL	18927
DICTIONARY	18675

The graph given below depicts the analysis between the four types of stemming algorithm out of which Porter algorithm is the most efficient as it gives the maximum word count.

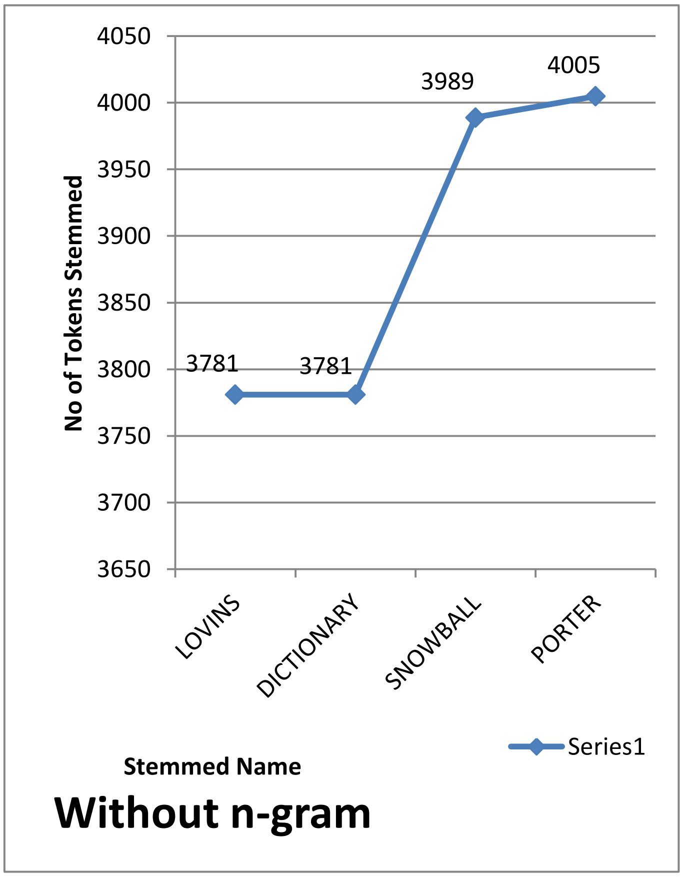

Analysis of different types of Stem Algorithms

With n-gram where n=2

The table given below depicts the number of tokens stemmed in all the four algorithms without n-grams

Stemmed Name	without n-gram
PORTER	4005
LOVINS	3781
SNOWBALL	3989
DICTIONARY	3781

The graph given below depicts the analysis between the four types of stemming algorithm out of which Porter algorithm is the most efficient as it gives the maximum word count.



Without n-gram

V. CONCLUSION

In the given context of study the different stemming algorithms have been applied and it was found that Porter algorithm works better with the FIRE Data set. We have also experimented with n-gram and without n-gram method with the same dataset. Based on the comparative study of this in the various stemming algorithms we can conclude that out of the four stemming algorithm Porter algorithm is the most efficient for English text.

REFERENCES

[1] Hearst, M. A. (1997) Text data mining: Issues, techniques, and the relationship to information access. Presentation notes for UW/MS workshop on data mining, July 1997.

[2] Zdenek Ceska and Chris Fox, “The Influence of Text Pre-processing on Plagiarism Detection”, International Conference. RANLP 2009 - Borovets, Bulgaria, pages 55–59

[3] www.techopedia.com

[4] Nikita P.Katariya *et al*, International Journal of Computer Science and Mobile Applications,Vol.3 Issue. 1, January- 2015, pg. 01-05 ISSN: 2321-8363]

[5] Ms. Anjali Ganesh Jivani et.al., “A Comparative Study of Stemming Algorithms”, Int. J. Comp. Tech. Appl., Vol 2 (6), 1930-1938 IJCTA NOV-DEC 2011.

[6] D’Amore, R. and Mah, C. One time complete indexing of text: Theory and practise. Eighth Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pages 155-164

[7] Inderjit Dhillon,” Introduction to Data Mining”, Spring 2009.

[8] Rangra Kalpana and Bansal K. L, “Comparative Study of Data Mining Tools”, International Journal of Advanced Research in Computer Science and Software Engineering Research Paper, Volume 4, Issue 6, June 2014 ISSN: 2277 128X.

[9] 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015) Monolingual Information Retrieval using Terrier: FIRE 2010 Experiments based on n-gram indexing Santosh K. Vishwakarmaa , Kamaljit I Lakhtariab , Divya Bhatnagarc , Akhilesh K Sharmad a Gyan Ganga Institute Of Technology & Sciences, Jabalpur, Madhya Pradesh, India

[10] Ljiljana Dolamic & Jacques Savoy UniNE at FIRE 2010: Hindi, Bengali, and Marathi IR

[11] Paul McNamee and James Mayfield, Character N0gram Tokenization for European Language Text Retrieval. Information Retrieval, 7:73-97,2004.

[12] Mierswart al, “ Rapid prototyping for complex data mining tasks”, In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 935–940. ACM, 2006.

[13] Land Sebastian and Fisher Simon ,”RapidMiner in academic use”, 27th August 2012 Rapid-I www.rapid-i.com.

[14] Mierswa, I. et al “YALE: Rapid Prototyping for Complex Data Mining tasks”, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-06), pp. 935-940, 2006.

[15] Ralf Mikut and Markus Reischl Wiley ,” Data Mining and Knowledge Discovery”, Volume 1, Issue 5, pages 431–443, September/October 2011.US UNIVERS

[16] Introduce to Data Mining with RapidMiner, 2008, Syracuse University, EECSRAPID MINER

[17] Paolo Palmerini, "On performance of data mining: from algorithms to anagement systems for data exploration", Technical Report, Universit`a Ca’ Foscari di Venezia, 2004.