TEXT MINING: APPLICATION OF STEMMING ALGORITHMS FOR INFORMATION RETRIEVAL IN HEALTHCARE WITH SPECIAL REFERENCE TO VIRAL INFECTIVE DISEASES

Article · March 2021

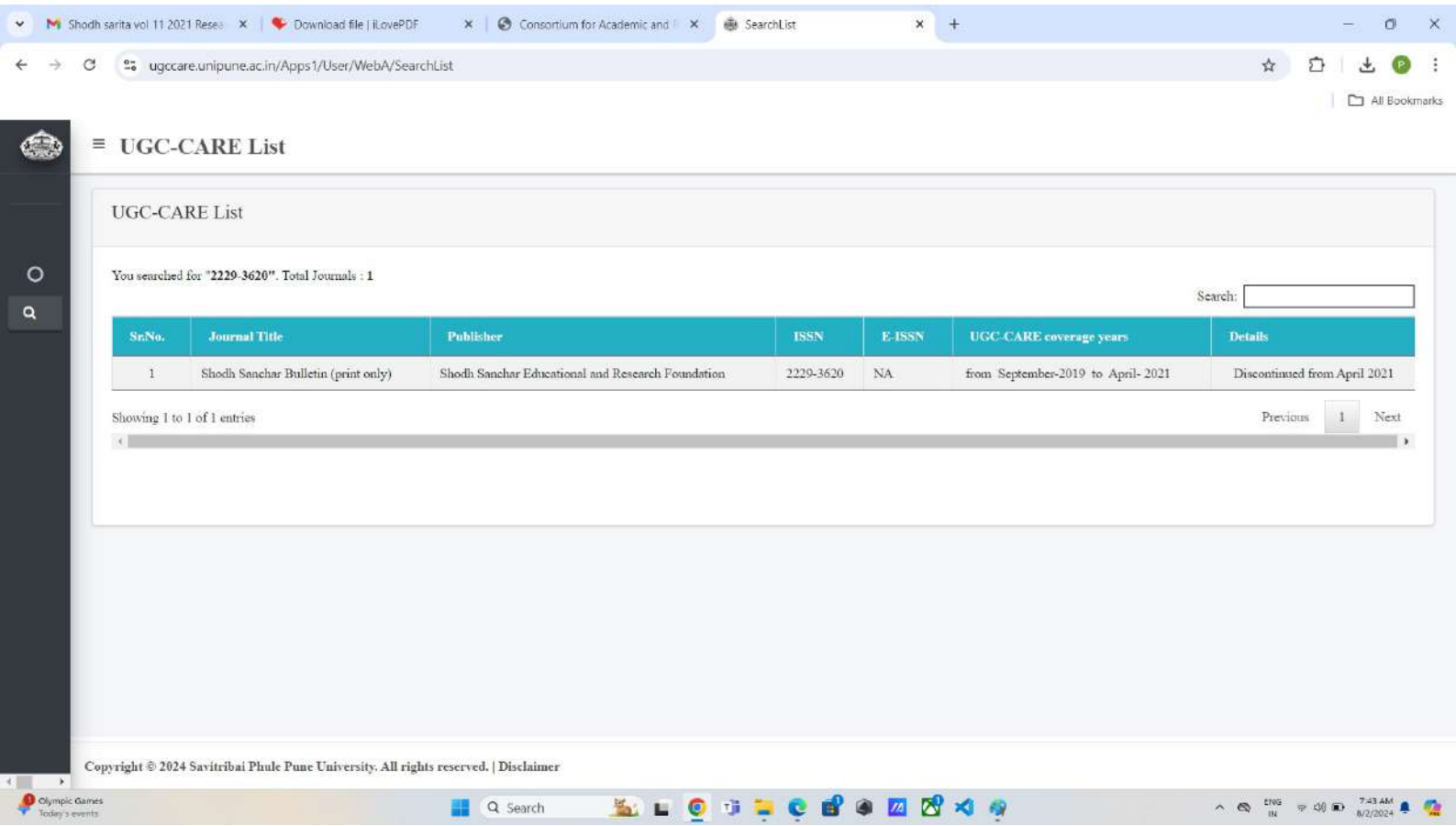
CITATIONS
0

READS
2

2 authors:

Shivaji Dattu Mundhe
37 PUBLICATIONS **127** CITATIONS
SEE PROFILE

Ashwini Manish Brahme
International Institute of Management Science , Pune
8 PUBLICATIONS **0** CITATIONS
SEE PROFILE


Shodh sarita vol 11 2021 Rese
Download file | iLovePDF
Consortium for Academic and
SearchList
ugccare.unipune.ac.in/Apps1/User/WebA/SearchList
All Bookmarks
UGC-CARE List
UGC-CARE List
You searched for "2229-3620". Total Journals : 1
Search:
Sr.No. | Journal Title | Publisher | ISSN | E-ISSN | UGC-CARE coverage years | Details
1 | Shodh Sanchar Bulletin (print only) | Shodh Sanchar Educational and Research Foundation | 2229-3620 | NA | from September-2019 to April- 2021 | Discontinued from April 2021
Showing 1 to 1 of 1 entries
Previous 1 Next
Copyright © 2024 Savitribai Phule Pune University. All rights reserved. | Disclaimer
Olympic Games
Today's events
Search
ENG
IN
7:43 AM
8/2/2024

GOVT. OF INDIA RNI NO.: UPBIL/2015/62096

UGC Approved Care Listed Journal

ISSN 2229-3620

SIC

SHODH SANCHAR Bulletin

An International
Multidisciplinary
Quarterly Bilingual
Peer Reviewed
Refereed
Research Journal

Vol. 11

Issue 41

January to March 2021

Editor in Chief
Dr. Vinay Kumar Sharma
D. Litt. - Gold Medalist

sanchar
Educational & Research Foundation

CONTENTS

ISSN - 2229-3620
UGC CARE LISTED JOURNAL
SIC
SHODH SANCHAR Bulletin
January-March, 2021
Vol. 11, Issue 41
Page Nos. 105-110
AN INTERNATIONAL BILINGUAL PEER REVIEWED REFEREED RESEARCH JOURNAL

TEXT MINING: APPLICATION OF STEMMING ALGORITHMS FOR INFORMATION RETRIEVAL IN HEALTHCARE WITH SPECIAL REFERENCE TO VIRAL INFECTIVE DISEASES

Dr. Ashwini Manish Brahme*
Dr. Shivaji D. Mundhe**

ABSTRACT

The information all over the globe is in diverse format and it is specified in three aspects namely structured, instructed and semi-structured. Data mining is only feasible for structured data and not for unstructured and semi-structured. Mining of such heterogeneous, complex and huge amount of data is creating challenges day by day in the field of data mining. The paper entitles towards the text mining phases such as text transformation, text pre- processing, filtration and stemming. The paper also aimed towards the high frequency viral infective diseases textual online news from various newspapers and processing it for better information retrieval. The research is aligned on various stemming techniques and their comparison.

Keywords: *Text mining, Stemming, Viral Diseases, Data Mining, Text Pre-Processing.*

I. INTRODUCTION

The information over the globe and Internet ear is in various forms namely graphs, email, scripts, blogs, audio-video, reports, etc. To take such heterogeneous and large information and generating patterns through the same is complex is one of the significant task in the information mining era. There are various applications of information, data, text mining some of them are IT, social media, online video/audio streaming channels, education, medical, banking and insurance, etc.

Healthcare text mining and information retrieval is one of the vital challenge to extract the knowledge from the medical domain; it includes natural language processing (NLP), the diverse forms of data and information, text refining, summarization, identifying the patterns form information, knowledge discovery, complexity, data dimensions, identifying the relationships between the healthcare data, longitudinal data , and many more. Therefore the present study is focused towards the better text mining and knowledge discovery for effective and efficient decision making system for healthcare sector.

The present study focused towards text mining of healthcare digital news available on the news archival of various websites. The researcher has aimed to text mine the common viral infectious diseases information.

II. TEXTMINING

The text mining is used to get the necessary data from large amount of data and generate the various patterns out of them; the phases included in the text mining are as represented in the following figure no.1.

*Assistant Professor - Sinhgad Institute of Management and Computer Application, Pune
**Director - International Institute of Management Science, Pune

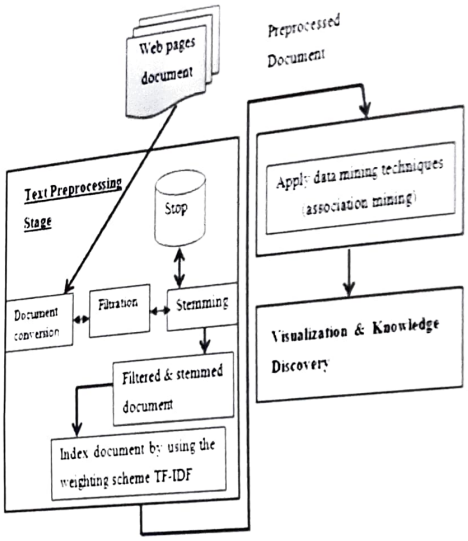


Figure No.1. Health Care Text Mining Theoretical Framework (Source: designed by Researcher)

The above figure no.1.Describes the healthcare text mining conceptual framework designed by the researcher for knowledge discovery in healthcaredomain the steps included in this process are Text Transformation, Text Pre-Processing, Data Mining on Processed Data/Information and Knowledge Discovery consequently.

A. Text Transformationis used to deal with the transformationof unstructured data to structured data to achieve efficient information retrieval from unstructured data. Then the second phase text pre-processing is carried out where the textual document is transformed to XML document. This step includes information filtration, text tokenization, and removal of stop words as well as performing stemming on the same. Into this step the entire step all special characters are removed from the text as well as the document is concerted to lower case which is necessary step for pre-processing.

B. The present study pre-processes the text of viral infective diseases news information which is available in unstructured format. The news is split into various parameters of natural language processing namely

character, symbol and words. This is known as tokenization which significant role for association rule and effective information retrieval represented in following figure no.2

id	type	filename	transaction	token
1	SwineFlu	...yr-old-dies-swine-flu-taking-toll-pune-81-3821	1	pune
2	SwineFlu	...yr-old-dies-swine-flu-taking-toll-pune-81-3821	1	a
.....	.....	.....	.....	.....
15418	Dengue	61098492.cms	44	xyz
15419	Dengue	61098492.cms	44	homes
15420	Dengue	61098492.cms	44	additional
.....	.....	.....	.....	.....
60777	HIV	61478954.cms	167	in
60778	HIV	61478954.cms	167	the
60779	HIV	61478954.cms	167	medical
60780	HIV	61478954.cms	167	college

Figure 2: Tokenization of Diseases New

C. Filtration of textual data plays signifi role to eliminate the less significant as as irrelevant words/tokens from the d this technique is also called as stop w removal. (The various stop works are as the, true, false, why,when,wh etc.).This step is important which re that the study can focus only on signifi and necessary words to achieve stemming, association mining, knowledge generation.

D. The result generated through stop w removal and tokenization is represented the following table no.1 :

Table 1: Tokenization and Stop word removal

Sr. No.	Particulars	Count
1	Number of tokens Generated	60,800
2	No. of Stop Words Removed	28,595
3	No. of Single Character Words	1685
4	Total Tokens remained for further Processing	30,520

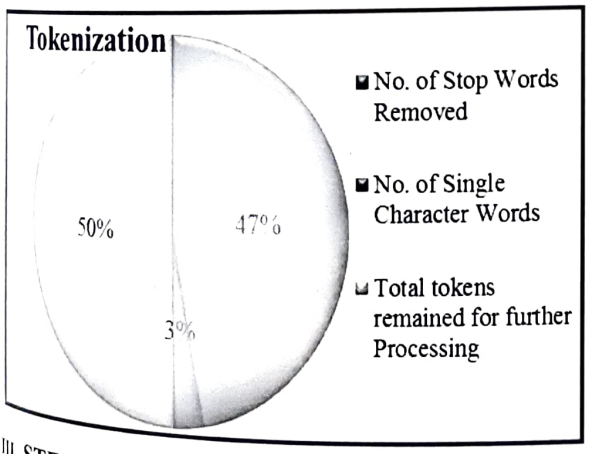
Source: Compiled by researcher

Therefore above table no.1.total tokens generated are 60,800, total 28,595 are stop words, 1685 are single characters; from total tokens the stop words and single character words are reduced and 30,520 are significant tokens/words considered for further processing.

Hence it has been resulted that, approximate 47.03 % words are irrelevant and 2.77 % are single character words which are not useful for the present study. Therefore it has depicted that approximate 49.80 % tokens/words are significant for the further study.

Therefore, it has been concluded that 50 % tokens are irrelevant and 50 % are relevant for association mining, stemming, and knowledge discovery. Also it outcomes the effective and efficient information retrieval from unstructured data/information represented as shown in the following graph no.1.

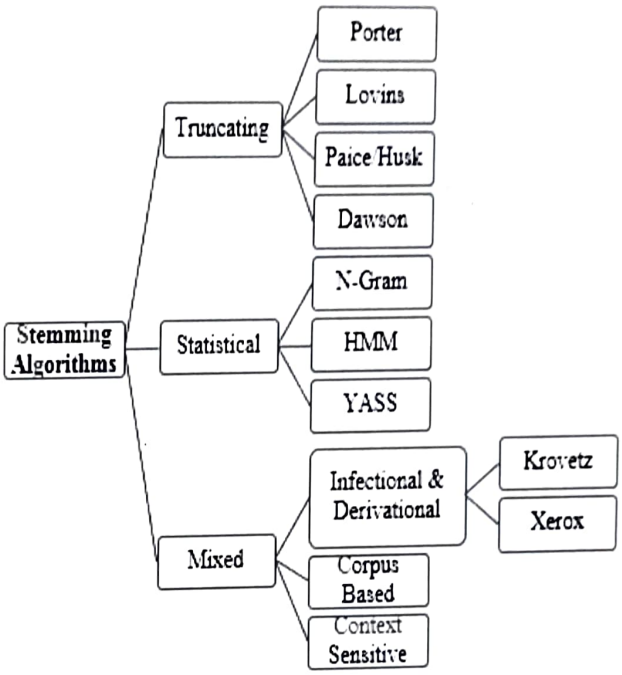
Graph no. 1: Filtered tokens and stop word removal



III. STEMMING

Stemming plays vital role in NLP, text mining, data mining, information retrieval for performing test pre-processing. Stemming is used to remove the suffix stripping of words and convert the word to itoriginal root word. For example if the word computer will stemmed as Compute, comput, Compu , etc. There are 10 stemming techniques fragmented into statistical stemming, truncating stemming and mixed stemming techniques. It is represented in the following figure no. 3.

Figure No.3. Stemming Techniques/Algorithms



IV. IMPLEMENATION OF STEMMING TECHNIQUES ONVIRAL DISEASES ONLINENEWS

The study focused towards the implementation of truncating and missed stemming algorithms for prefix and suffix striping. The Lovins, porters, Korvetz, and Paice/Husk stemming algorithms are considered to implement on most frequently occurring viral infective diseases specifically Influenza Flu, Swine Flu, Diarrhea and HIV, Dengue and Chikungunya. As shown in the following figure no.4.

Figure No.4: Stemming of Viral diseases tokens

No.	Original Tokens	Root words	Correct / Incorrect	Poters Stemming	Correct / Incorrect	Lovins Stemming	Correct/ Incorrect	Paice/ Husk Stemming	Correct / Incorrect	Krovetz Stemming	Correct/ Incorrect
1	pune	pune	TRUE	pune	TRUE	pun	FALSE	pun	FALSE	pune	TRUE
2	year	year	TRUE	year	TRUE	year	TRUE	year	TRUE	year	TRUE
3	old	old	TRUE	old	TRUE	old	TRUE	old	TRUE	old	TRUE
4	boy	boy	TRUE	boi	FALSE	boy	TRUE	boy	TRUE	boy	TRUE
5	lost	lost	TRUE	lost	TRUE	lost	TRUE	lost	TRUE	lost	TRUE
6	battle	battle	TRUE	battl	FALSE	battl	FALSE	battl	FALSE	battl	TRUE
202	died	die	FALSE	di	FALSE	di	FALSE	died	FALSE	di	FALSE
203	due	due	TRUE	due	TRUE	du	FALSE	due	TRUE	due	TRUE
204	swine	swine	TRUE	swine	TRUE	sw	FALSE	swin	FALSE	swine	TRUE
205	flu	flu	TRUE	flu	TRUE	flu	TRUE	flu	TRUE	flu	TRUE
417	virus	virus	FALSE	viru	FALSE	virus	TRUE	vir	FALSE	virus	TRUE
433	issued	issue	FALSE	issu	FALSE	issu	FALSE	issu	FALSE	issu	TRUE
463	prepared	prepare	FALSE	prepar	FALSE	prepar	FALSE	prep	FALSE	prepar	FALSE
514	vegetable	vegetable	FALSE	veget	FALSE	veges	FALSE	veget	FALSE	veget	FALSE
527	bacterial	bacterial	FALSE	bacteri	FALSE	bacter	FALSE	bact	FALSE	bacter	FALSE
545	admission	admission	FALSE	admiss	FALSE	admis	FALSE	admit	FALSE	admission	TRUE
746	saturday	saturday	TRUE	saturdai	FALSE	saturda	FALSE	saturday	TRUE	saturday	TRUE
747	morning	morning	FALSE	morn	FALSE	morn	FALSE	morn	FALSE	morning	TRUE
1159	expressing	express	FALSE	express	TRUE	expres	FALSE	express	TRUE	expressing	FALSE

The stemming of diseases news results in the form of correct and incorrect stemming as represented in the following table no.2.

Table 2: Correct and incorrect stemming

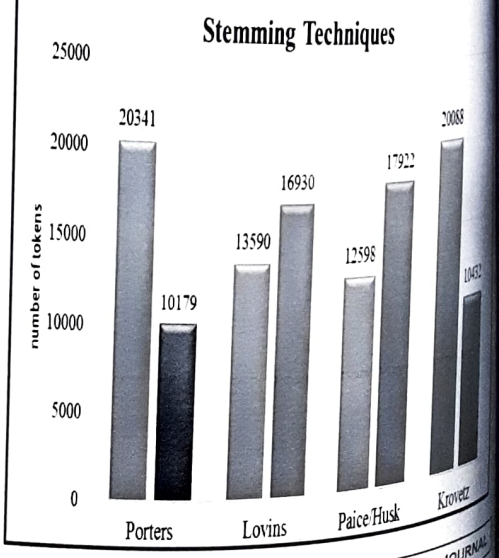
Stemming Technique	Porters	Lovins	Paice/Husk	Krovetz
Correct Stems	20341	13590	12598	20088
Incorrect Stems	10179	16930	17922	10432
% of Correct Stemming	**66.65**	**44.53**	**41.28**	**65.82**
% of incorrect Stemming	**33.35**	**55.47**	**58.72**	**34.18**

Source: Compiled by researcher

The table no. 2 results correct and incorrect stems of as

- Porters stem outcomes 24341 correct and 10179 are incorrect stem out of 30520 tokens.
- 13590 correct stems while as 16930 are incorrect stemsresulted in the Lovins stemming technique.
- Paice/Husk occasioned 12598 cor and 17922 incorrect stems and
- Korvetz has results 20088 correct wo as 10432 incorrect stems

Graph no. 2: Porters, Lovins, Paice/Husk a Korvetz Stemming Techniques comparat analysis



V. PORTERS STEMMING

The above table and graph no 2 depicts the comparative analysis of correct and incorrect stemming. It results that porters stemming performs better stemming as compares to others as 66.65 % correct stemming.

There are many researchers working on improving the stemming efficiency and correctness as the porters stemming has some pitfalls as:

- There are 60 rules and five steps for suffix stripping.
- It has over-stemming and under-stemming demerit.
- It does not generate the correct stems many times.
- Due to the same it gives irrelevant information retrieval and changes the proper or root meaning of the word.

VI. FUTUER SCOPE

The comparative study of selected stemming represents the pitfalls of porters stemming technique and creates new opportunity for modification or invention of new advanced stemming for effective correct stemming. Text mining a better stemming consequences into enhanced NLP, information retrieval and knowledge discovery and association mining and future forecasting.

Hence the present study envisioned the designing of new stemming algorithm to curb the pitfalls of exiting techniques of porters stemming.

VII. CONCLUSION

The ample amount of online healthcare information and data is increasing day by day all over the globe in heterogeneous form. The research was conducted on information/ text mining of viral infective diseases news from news archival form internet. The text mining and text processing was conducted and selected stemming techniques specifically Lovins, Paice/Husk, Porters, Korvetz are applied on the selected dataset. It results in correct and incorrect stemming and concludes that the porters stemming performs better stemming as compare to others. There is need to improve the efficiency and maximum correct stems it has given rises for further improvements and better natural language processing.

REFERENCES

1. Jivani A. (2011). A Comparative Study of Stemming Algorithms. Int. J. Comp. Tech. Appl. (IJCTA), 2 (6), 1930-1938, www.ijcta.com
2. S. Giridhar, V. Prema, Reddy S. (2011). Giridhar N S., Prema K.V., N. V Subba Reddy (2011). A Prospective Study of Stemming Algorithms for Web Text Mining. Ganpat University Journal of Engineering & Technology,1(1), 28-34
3. Sharma D. (2012). Stemming Algorithms: A Comparative Study and their Analysis. International Journal of Applied Information Systems (IJAIS), Foundation of Computer Science FCS, New York, USA, 4(3), 7-12, www.ijais.org
4. N. Sandhya, Y. Sri Lalitha, Sowmya, K. Anuradha, A. Govardhan (2011). Analysis of Stemming Algorithm for Text Clustering. IJCSI International Journal of Computer Science, 8(5), 352-359, www.IJCSI.org
5. Ramasubramanian C., Ramya R. (2013). Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 2(12), 4536-4538, www.ijarcce.com
6. S. Vijayarani., J. Ilamathi, Nithya. Preprocessing Techniques for Text Mining -

An Overview, International Journal of Computer Science&Communication Networks, 5(1), 7-16

7. Kulkarni M., Kulkarni S. (2016). Knowledge Discovery in Text Mining using Association Rule Extraction, International Journal of Computer Applications (0975 – 8887), 143(12), 30-35

8. M.F. Porter, 1980, An algorithm for suffix stripping, Program, 14(3) pp 130−137

9. Kulkarni A., Mundhe S. (2018). Implementation of Effective Stemming Algorithm of Text Mining for Knowledge Discovery in Healthcare. SavitribaiPhule Pune University

10. http://snowball.tartarus.org/algorithms/porter/ste mmer.html

11. http://www.punediary.com/html/press.html

12. https://www.sccollege.edu/Library/Pages/primary sources.aspx

13. https://www.practo.com/pune, (Accessed website on 12 /11/2016)

14. http://timesofindia.indiatimes.com/archive.cms

15. http://www.news-medical.net/

16. http://www.punecorporation.org/en/health-department-3

17. http://imapune.org/, Indian Medical Association Pune Branch