

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/382796772>

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

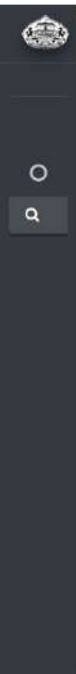


Ashwini Manish Brahme

International Institute of Management Science , Pune

8 PUBLICATIONS 0 CITATIONS

SEE PROFILE



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

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

S. No.	TITLE	NAME OF AUTHORS	PAGE No.
1.	AN ANALYTICAL STUDY OF FLEXI LEAVE POLICY AND ITS IMPACT ON PBO'S EMPLOYEES	Dr. Vijay Dhole	1
2.	STUDY OF IMPACT OF CORONA PANDEMIC AND THE GLOBAL RECESSION ON JOB OPPORTUNITIES FOR INDIAN YOUTH IN KPO SECTOR	Dr. Pallavi Sajanapwar	7
3.	INNOVATION IN OUT OF HOME (OOH) MEDIA ADVERTISING RESEARCH: A LITERATURE REVIEW	Prof. Rupa Rawal Dr. Amod Markale	13
4.	ROLE OF COMPUTER TECHNOLOGY ON PHARMACY IN INDIA	Dr. D. R. Vidhate Ms. D. D. Vidhate	19
5.	UP SKILLING THE SKILLS FOR FUTURE EMPLOYMENT CHALLENGE-2030	Prof. Manoj Sathe Dr. Sudarshan Pawar	24
6.	INVESTORS PREFERENCES TOWARDS GOLD AS AN INVESTMENT OPTION WITH REFERENCE TO PUNE DISTRICT.	Mr. Swapnil Patil Dr. Eknath B. Khedkar	29
7.	A STUDY OF THE IMPACT OF PANDEMIC COVID 19 ON THE HEALTH CARE SECTOR	Dr. Gauri Prabhu	33
8.	LIFELONG MACHINE LEARNING IN SENTIMENT CLASSIFICATION: CONCEPTS AND IMPLEMENTATIONS	Krishna Priya S. Kavita S. Oza	37
9.	A COMPARATIVE STUDY OF URBAN AND RURAL HOUSEHOLDS INVESTMENT AVENUES IN MAHARASHTRA.	Ghodake Shamrao P. Dr. E. B. Khedkar	43
10.	LITERATURE REVIEW ON EMPLOYEE ENGAGEMENT PRACTICES	Dr. Pushpraj Wagh Ms. Pooja Salvekar	52
11.	NEW WORLDVIEW IN SECURITY FOR BANKS: DECEPTION TECHNOLOGY	Ms. Ashwini R. Chavan Dr. R. D. Kumbhar	58
12.	COMPARATIVE STUDY OF ONLINE V/S OFFLINE MODE OF EDUCATION	Mrs. Rupali Kalekar Ms. Aanchal Priya	64
13.	A STUDY OF INBOUND MARKETING AND ITS IMPORTANCE IN INNOVATION MANAGEMENT	Prof. Asmita Abhijeet Gaikwad Dr. Amod Markale	68
14.	POST COVID-19: A TECHNICAL APPROACH FOR HIGHER EDUCATIONAL EXCELLENCE	Ms. Sarika Choudhari Mrs. Arati Patil	73
15.	REDEFINING HR WITH ARTIFICIAL INTELLIGENCE: CHATBOTS THE NEW ASSISTANT!	Sarah Dsouza Saylee Anil Karande Angshupriya Datta	80

16.	CONTRIBUTIONS OF LOGISTICS AND SUPPLY CHAIN IN ENTREPRENEURIAL ECOSYSTEM IN DEVELOPING COUNTRIES WITH SPECIAL REFERENCE TO TANZANIA	Dr. Maige Mwakasege Mwasimba Mr. Alfred Sallwa	87
17.	CRITICAL ANALYSIS OF THE CORONA DISEASE PANDEMIC AND ITS IMPACT ON EDUCATION SYSTEM IN INDIA	Prof. Manoj Ashok Sathe Prof. Yugandhara R. Patil	93
18.	PERCEPTIVE CROSS - CURRENTS IN VOCATIONAL TRAINING	Dr. Vandana Mohanty Colonel (Dr.) J. Satpathy Ms. Dhanashree Shinde	98
19.	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	105
20.	STUDY OF THE FINANCIAL IMPLICATIONS OF PSB MERGERS IN INDIA	Mr. Yogesh K. Nakhale Dr. Gauri Prabhu	111
21.	A POST IMPLEMENTATION STUDY OF SUGARCANE IRRIGATION MANAGEMENT SYSTEM	Dr. Nilam Jadhav Dr. Shivaji Mundhe	119
22.	A COMPARATIVE ANALYSIS OF CAPITAL STRUCTURE ADJUSTMENT WITH RESPECT TO PARTIAL CAPITAL STRUCTURE ADJUSTMENT IN THE SELECTED BSE LISTED INDIAN MANUFACTURING COMPANIES	Vikas R. Adhegaonkar Dr. E. B. Khedkar	124
23.	ENVIRONMENTAL ANALYSIS: TO ANALYZE VARIOUS ENVIRONMENTAL FACTORS AFFECTING ESSITY AND IDENTIFY THE KEY DRIVERS AMONGST THEM.	Tejas Gujar Dr. Bharati Jadhav	128
24.	CUSTOMER REVIEWS SENTIMENTS ANALYSIS USING NATURAL LANGUAGE PROCESSING (NLP) AND DEEP LEARNING.	Dr. Sachin Misal Dr. Shivaji Mundhe	134
25.	BEHAVIOUR OF INVESTOR ON INVESTMENT DECISION WITH SPECIAL REFERENCE TO MUTUAL FUND VS. TRADITIONAL INVESTMENT AVENUES- A REVIEW	Prof. Mahesh Mahankal Dr. Prabha Singh	139
26.	FACE IDENTIFICATION IN A GROUP BASED ON LBP ALGORITHM AND NEURAL NETWORK CLASSIFIER FOR CLASS ATTENDANCE	Narayan Kulkarni Dr. H. S. Fadewar	145

27.	FACTORS AFFECTING CONSUMER'S BUYING BEHAVIOR TOWARDS ORGANIZED RETAILING WITH REFERENCE TO STAR BAZAAR IN PUNE CITY	Dr. Pushpraj Wagh	151
28.	VEHICLE REGISTRATION NUMBER RECOGNITION USING MACHINE LEARNING	Dr. Sachin Misal Mr. Sunil Joshi	158
29.	INNOVATION- AN INSTRUMENT FOR SUSTAINABLE ECONOMY	Singha Roy Ekta	164
30.	AUTOMATIC IRIS RECOGNITION SYSTEM IN HCI	Ganesh K. Awasthi Dr. H. S. Fadewar	168
31.	VOCATIONAL EDUCATION AND TRAINING: IS IT REALLY EFFECTIVE?	Dr. Vandana Mohanty Dr. Ashutosh Zunjur	175
32.	AN EMPIRICAL STUDY ON TECHNIQUES TO CREATE A POSITIVE LEARNING ENVIRONMENT	Dr. Sonali Dharmadhikari Dr. Shweta Joglekar	182
33.	EFFECT OF BONUS ISSUE ON PRICES OF COMPANIES AT STOCK EXCHANGE	Prof. Mahesh Mahankal Mr. Rahul Bari	187
34.	TREND ANALYSIS OF FINTECH USAGE IN PUNE CITY DURING COVID-19	Mrs. Poorva Pachpore Dr. Gauri Prabhu	191
35.	A STUDY OF AWARENESS REGARDING DIABETES MELLITUS AMONG RURAL COMMUNITY OF AHMEDNAGAR DISTRICT	Dr. Gorakshanath T. Gund Dr. Prashant Radhakrishna Tambe	196
36.	VALIDATING CONSUMER ETHNOCENTRISM SCALE (CETSCALE)	Dr. Ashutosh Zunjur Dr. Joe Lopez	202
37.	NLP FOR SOCIAL MEDIA DATA PROCESSING	Miss. Ankita D. Garud Prof. Prashant N. Wadkar	208
38.	A COMPARATIVE STUDY OF WIDELY USED VIRTUAL CLASSROOM TECHNIQUES	Dr. Nilam Jadhav Dr. Kedar Marulkar	214
39.	LINKING BETWEEN KNOWLEDGE AND TALENT MANAGEMENT IN AN IT FIRMS – LITERATURE REVIEW	Ms. Deepti M. Yadav Dr. Bholu Sarang S.	219
40.	STUDY OF E-MARKETING PRACTICES OF SELECTED SMARTPHONE BRANDS FOR PCMC REGION	Prof. Amarnath Gupta Ganesh Kalshetty	226



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 unstructured, structured 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 era 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 T, 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

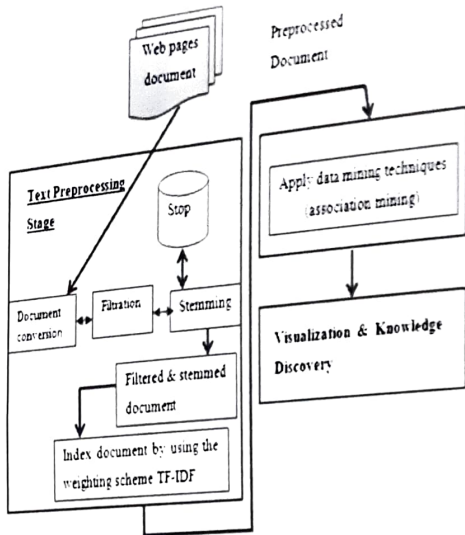


Figure No.1. Health Care Text Mining Theoretical Framework (Source: designed by Researcher)

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

- A. Text Transformation is used to deal with the transformation of 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 converted 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 process is known as tokenization which plays a significant role for association rule mining and effective information retrieval. This is represented in following figure no.2

id	type	file name	transaction	tokens
1	SwineFlu	7-yr-old-dies-swine-flu-taking-toll-pune-S1-3821	1	pune
2	SwineFlu	7-yr-old-dies-swine-flu-taking-toll-pune-S1-3821	1	a
15418	Dengue	61098492.cms	44	vip
15419	Dengue	61098492.cms	44	homes
15420	Dengue	61098492.cms	44	additional
60777	HEV	61478954.cms	167	in
60778	HEV	61478954.cms	167	the
60779	HEV	61478954.cms	167	medical
60780	HEV	61478954.cms	167	college

Figure 2: Tokenization of Diseases News

- C. Filtration of textual data plays significant role to eliminate the less significant words as irrelevant words/tokens from the data. This technique is also called as stop word removal. (The various stop words are as the, true, false, why, when, who, etc.). This step is important which results that the study can focus only on significant and necessary words to achieve stemming, association mining, knowledge generation.
- D. The result generated through stop word removal and tokenization is represented in 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

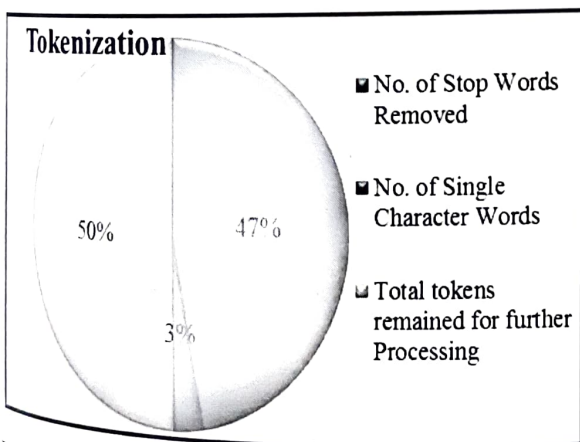
Source: Compiled by researcher

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

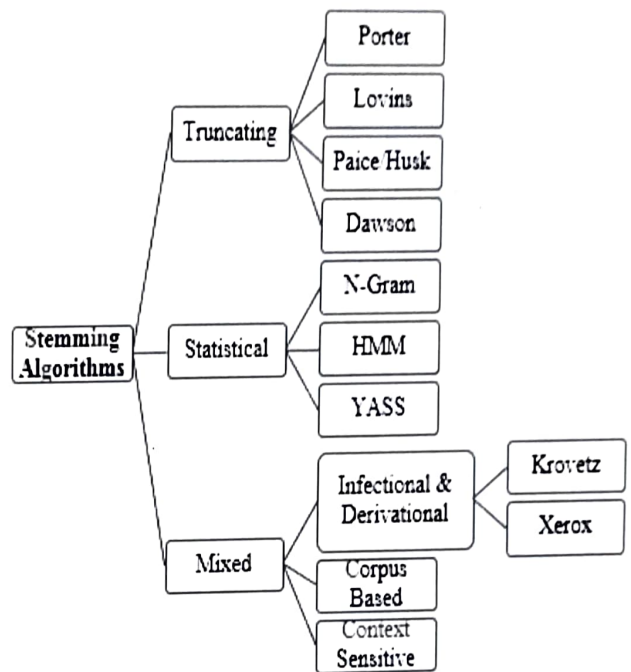


III. STEMMING

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. IMPLEMENTATION OF STEMMING TECHNIQUES ONVIRAL DISEASES ONLINENEWS

The study focused towards the implementation of truncating and missed stemming algorithms for prefix and suffix stripping. 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	Porters 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		
2	year	year	TRUE	year	TRUE	year	TRUE	year	TRUE		
3	old	old	TRUE	old	TRUE	old	TRUE	year	TRUE	pune	TRUE
4	boy	boy	TRUE	boi	FALSE	boy	TRUE	old	TRUE	year	TRUE
5	lost	lost	TRUE	lost	TRUE	lost	TRUE	boy	TRUE	old	TRUE
6	battle	battle	TRUE	battl	FALSE	battl	FALSE	lost	TRUE	boy	TRUE
202	died	die	FALSE	di	FALSE	di	FALSE	battl	FALSE	lost	TRUE
203	due	due	TRUE	due	TRUE	du	FALSE	died	FALSE	battl	TRUE
204	swine	swine	TRUE	swine	TRUE	sw	FALSE	due	FALSE	di	FALSE
205	flu	flu	TRUE	flu	TRUE	flu	FALSE	swin	TRUE	due	FALSE
417	virus	virus	FALSE	viru	FALSE	virus	TRUE	swin	FALSE	swine	TRUE
433	issued	issue	FALSE	issu	FALSE	issu	TRUE	flu	TRUE	flu	TRUE
463	prepared	prepare	FALSE	prepar	FALSE	issu	FALSE	vir	FALSE	flu	TRUE
514	vegetable	vegetable	FALSE	veget	FALSE	prepar	FALSE	issu	FALSE	virus	TRUE
527	bacterial	bacterial	FALSE	bacteri	FALSE	prepar	FALSE	prep	FALSE	issu	FALSE
545	admission	admission	FALSE	admiss	FALSE	bacter	FALSE	veget	FALSE	prepar	FALSE
746	saturday	saturday	TRUE	saturdai	FALSE	admiss	FALSE	bact	FALSE	veget	FALSE
747	morning	morning	FALSE	mom	FALSE	admis	FALSE	admit	FALSE	bacten	FALSE
1159	expressing	express	FALSE	express	TRUE	saturda	FALSE	saturday	FALSE	admission	TRUE
						mom	FALSE	morning	TRUE	saturday	TRUE
						expres	FALSE	express	TRUE	morning	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

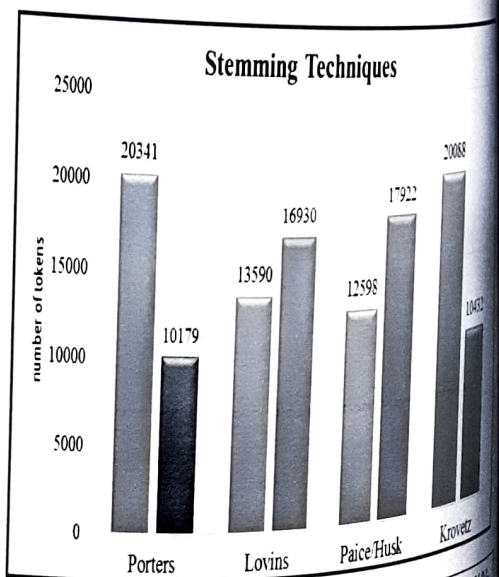
Source: Compiled by researcher

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 stems resulted in the Lovins stemming technique.

- Paice/Husk occasioned 12598 correct and 17922 incorrect stems and
- Krovetz has results 20088 correct and as 10432 incorrect stems

Graph no. 2: Porters, Lovins, Paice/Husk, Krovetz Stemming Techniques comparison 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 (IJ AIS)*, 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/stemmer.html>
 11. <http://www.punediacry.com/html/press.html>
 12. <https://www.sccollege.edu/Library/Pages/primarysources.aspx>
 13. <https://www.practo.com/pune>, (Accessed website on 12 /11/2016)
 14. <http://timesofindia.indiatimes.com/archives>
 15. <http://www.news-medical.net/>
 16. <http://www.punecorporation.org/en/health-department-3>
 17. <http://imapune.org/>, Indian Medical Association Pune Branch

