DESIDOC Journal of Library & Information Technology, Vol. 39, No. 6, Nov 2019, pp. 315-321, DOI : 10.14429/djlit.39.6.15145
© 2019, DESIDOC

A Framework to Process Text Data of Web Discussion Forums: A Study of LIS Links

Mohit Garg[#,*] and Uma Kanjilal[$]

[#]*Indira Gandhi National Tribal University, Amarkantak – 484 887, India*
[$]*Indira Gandhi National Open University, New Delhi – 110 068, India*
[*]*E-mail: mohitji@igntu.ac.in*

ABSTRACT

Nowadays, people use the internet for both seeking and disseminating information in a collaborative way on various social media platforms like Quora, Yahoo Answers, LisLinks Forum, etc. This social interaction on different topics makes these platforms as a knowledge repository. Evaluation of these repositories can help to understand various trends. However, this evaluation is a challenging task because of unstructured data and the unavailability of application programming interfaces for the harvesting of a dataset. This study presented a framework to harvest and pre-processing of data available on LisLinks Forum. The proposed framework is implemented using statistical programming language R. The fourteen metadata elements were defined for the discussion forums. The framework automatically harvest and pre-process relevant data of posts.

Keywords: **T**ext mining; Discussion forums; LIS Links; Data pre-processing.

1. INTRODUCTION

The internet has become an important communication medium not only as a source of information but to disseminate experiences, knowledge, etc. With the rapid growth of technology especially mobile, web 2.0 and social media facilitated the freedom to people to express their feelings, views, and opinions in natural language about products or services they have used. These feelings, opinions, discussions shared by netizens on different platforms are termed as user-generated content or user-generated data. This plethora of data available on various web and social media platforms like Products reviews (e.g., Amazon), Hotel & Destination Place reviews(e.g., TripAdvisor), Human behavior on social networking sites (e.g., Facebook), Microblogging sites (e.g., Twitter), video sharing site (e.g., YouTube) or discussion forums(e.g., Yahoo Answers, LisLinks Forums) is often referred to as big data. As per the reports, every day 2.5 quintillion bytes of data is user-generated data.[1] The availability of an abundance of data is opening new opportunities in many fields of application to portrait the behavior of people. Mining this data provides an instantaneous understanding of what people are interested in and what is their opinion about a certain topic. This user-generated data contain digital traces of people that are suitable for academic research to understand the latest trends, behavior, etc.

Analyzing this massive volume of text data in natural language will help to uncover hidden knowledge which has endless opportunities mainly in social science research, where there was always scarcity of correct or more meaningful observations.

Online discussion forums are asynchronous communication platforms. These are popular among learners of different domains to seek help and discuss topics related to courseware, project etc. and widely applied tool in web-based education system.[18] The benefit of online community forums is that these have a social networking structure where communication is not only between friends, relatives but to a larger audience who all are related to the community irrespective of whether the person who posts the query and who reply are linked or not. The data like question answer dialogues on discussion forums can be used in solving a wide range of tasks like expert finding, thread summarisation, assessment of post quality, topic detection, etc.[2] What are the popular topics of discussion in the community? How these popular topics have evolved? , What type of post is replied?, are some of the research questions that can be answered with the analysis of the discussion forum's data. Large number of discussion forums are active on the internet, but very few have been able to attract audience for different reasons. Researchers have studied the motivation of the users to ask question on these forums. Choi, Shah (2016) found that Users seek opinions and advice on Yahoo Answers, whereas WikiAnswers was mostly used for finding the facts.[17] In the library and information science domain, one such popular discussion forum is www.lislinks.com. The activeness of the forum can be understood by the fact that it has more than 25000 registered members and more than 10,000 posts.

Received : 25 September 2019, Revised : 1 November 2019
Accepted : 26 November 2019, Online published : 11 December 2019

This large number of posts created on LisLinks Forums, the interaction between peer to peer make it as a big knowledge repository of discussion of Library professionals. This knowledgebase can act as a reference guide for students. Evaluation of these repositories can give valuable insight to know the latest trends. Analyzing whole text will help in to discover the topics of discussion for the organisation of posts and mining the popular topics. Analyzing this large knowledge will help to understand interesting patterns like trends of the question asked, key problems or issue discussed, etc.

Social media platforms like Facebook, Twitter, GoodReads, and YouTube provide the facility to access their data for academic and research purposes through different Application Programming Interfaces (APIs), but accessing the data of discussion forums whose APIs are not available is a challenging task. Apart from accessing this data, another big challenge is how to transform this inchoate data streams into meaningful formats. The user-generated text data are not immediately amenable to analysis as it contains different types of noises like spelling errors or typos, unknown acronyms, grammatical mistakes, etc.

The manual extraction of data and the removal of noises from it will be a very difficult task. Many studies have shown that the pre-processing of data is a critical and most time-consuming process.[19] This brought the need for an automated system that harvest and pre-processes user-generated data of LisLinks discussion forums. In this present study, we will be focusing of designing the framework that serves both the challenge of harvesting data from forums and pre-processing of it. The organisation of the article as follows, The first section introduce and set the premise about research on discussion forums. The next section reviews the literature focusing on the assessment of different discussion forums. The third section discusses objectives and scope of the research work. The framework and the results are discussed in fourth section. And the last section, conclude and talk about future work.

2. LITERATURE REVIEW

To the best of our knowledge, no study has been traced out in the literature which discusses the harvesting and pre-processing of the data from the discussion forums. However, many researchers have analysed user-generated data of different forums like yahoo answers, trip advisors, and stack overflow to categorise the type of information discussed on these platforms. Liu, 2014 evaluated knowledge on heart disease discussion forums. They showed that these discussion forums provide less biased information on adverse drug events than the information provided by the drug regulatory agency.[3] Calefato, Lanubile, & Novielli, 2018 empirically studied important factors to get answers on Stack Overflow. The study recommended that presentation-quality, a question to be précised includes code snippet to get answered.[4] Sahu, Nagwani, & Verma, 2016 proposed a model for automation of identification of authoritative users on two stack exchange websites, i.e. Stack OverFlow and AskUbuntu.[5] Aikawa, Sakai, & Yamana, 2011 studied the behavior of community users to understand whether they are looking for subjective answers or not?[6] Dearman & Truong, 2010 examined what the reason is behind to not answer a question on Yahoo Answers.[7] Chen, Zhang, & Mark, 2012 investigated the intention of the user to seek answer for a given question.[8] Wen, Yang, & Rosé, 2014) qualitatively analysed the sentiments of Massive Open Online Course (MOOC) discussion forums.[9] Zhou et al., 2008 proposed a framework to suggest questions, and found that the framework is effective than other frameworks.[10] Chai, Hayati, Potdar, Wu, & Talevski, 2010 proposed a method to measure content quality in forums.[11] Baldwin, Martinez, & Penman, 2007 analysed the text data of Linuz web user forums to visualise the problems and solutions described in the forum's threads. They classified the thread in to three different categories like whether the thread discuss specific or general problem, whether the initial problem is solved or not and whether the initial post is complete or not.[12] Choi, 2013 examined the motivations and expectations of the people asking questions in the online Q& A sites.[13] Carenini, Ng, & Zhou, 2007 summarised the email conversations based on clue words.[14] Wang & Zhang, 2016 examined characteristics and activity patterns of zhihu.com, expert social question answering site in china. The study showed that more questions are asked by starters. The person who answers the questions add more content to the sites.[15] Fu and Fan (2016), investigated the dataset of question-answer Music StackExchange site and found that the majoritarian asked questions related to music performance.[16]

3. OBJECTIVES

The present study aims to develop a framework to harvest unstructured data available on web discussion forums and pre-process it to prepare for the analysis.

The scope of the study is limited to the discussion forum of the Library and Information Science domain,i.e. LisLinks Forums. However, the framework can also be used for other discussion forums with some configuration based on their structure.

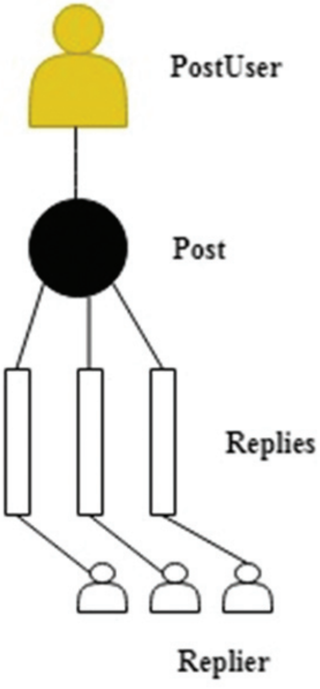


Figure 1. Structure of Lsilinks forums.

4. FRAMEWORK

The framework takes the benefit of the structure of website design. Web page documents are written in HTML and designed with CSS. A discussion forum's website can be defined as a collection of the webpage designed in a tree structure, where the root of the website contains several branches and each branch contain many sub-branches. LisLinks Forums is one such branch of root website www.Lislinks.com where discussion can be started on different topics. The registered users can start a genuine discussion after signing on the portal as each post is approved by the administrators. After the approval, each post has been assigned a URL that has prefix http://www.lislinks.com/forum/topics/. The specific information of the posts is stored in one particular web element. For Example, the "h1" element contains information about the title of the post. So all posts contain the title of the post in "h1" only. This consistency of the structure of the posts helps in automating the harvesting processing of a large number of posts (Fig. 1).

Table 1. Metadata elements

Metadata Element	Description
PostTitle	Title of the Post
PostBody	A text description of the post
PostLang	A language of the Post
PostCategory	The type of Post (as per predefined 18 categories)
PostURL	Unique Identifier of the Post
PostDate	A date associated with the creation of the post
PostTime	A time associated with the creation of the post
PostUser	User who created the post
PostViews	Number of views of the Post
PostReplyCount	Number of replies to the Post
PostRepliedBody	A text description of the reply of the post
PostRepliedDate	A date associated with the creation of the post reply
PostRepliedTime	A time associated with the creation of the post reply
PostRepliedUser	A user who replied to the post

The structure of the post can be understood by seeing the figure. The figure shows, how social interactions take place on the LisLinks forum. The PostUser can start the discussion on the forum. Each post has replies given by the other users based on their interest in the topic. However, not all posts get replied by the users of the community.

Metadata elements were defined for the discussion forums. The framework harvest data based on these metadata. Table 1 shows the list of 14 metadata elements related to discussion forums.

The framework is divided into two phases. The first phase of the framework harvest the data and store in the database whereas the pre-processing of data is done in the second phase. Figure 2 shows the data harvesting phase of the proposed framework.

The first phase of the framework consists of two data harvester written in a statistical programming language, R. The first data harvester (DH1) will harvest the PostUrl of all the posts of a given page range. The forum's page display ten posts per page in a default setting. The input for DH1 is the base URL of the page, the information containing element (ICe) and the page range. The base URL of the page is "http://www.lislinks.com/forum?page", ICe is the element "a" along with its attribute "href", and page range is a sequence of numbers from 1 to n, where n is the total number of pages. After passing the inputs, the DH1 request the discussion forums and harvest all the URLs of posts and stores it in a database. Table 2 shows the PostUrl of the first ten posts harvested on 17/06/2019.

PostUrl is the input for DH2 which harvests other data of the posts. ICe for DH2 will be different for different

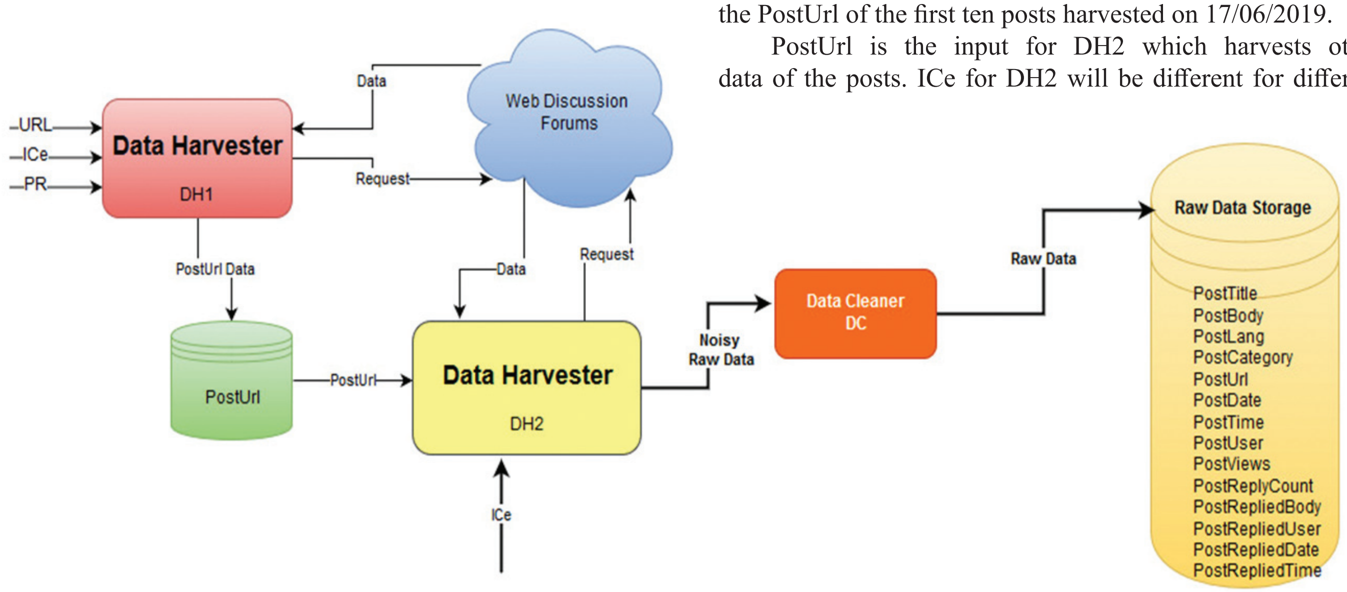


Figure 2. Data harvesting phase.

Table 2. Harvested URLs of the post

Post Url
http://www.lislinks.com/forum/topics/please-suggest-us-to-read-prepare-for-ugc-net
http://www.lislinks.com/forum/topics/university-grants-commission-notification-new-delhi-the-18th-july
http://www.lislinks.com/forum/topics/publisher-of-colon-classification-7th-edition
http://www.lislinks.com/forum/topics/oclc-products-and-services-with-previous-questions-from-net-exam
http://www.lislinks.com/forum/topics/how-to-face-interview-at-azim-premji-university-for-library-train
http://www.lislinks.com/forum/topics/rssmb-librarian-exam-date-2019
http://www.lislinks.com/forum/topics/ceptam-of-drdo-document-verification-will-m-lib-be-considered-ins
http://www.lislinks.com/forum/topics/sticking-print-out-of-library-bill-in-stock-register
http://www.lislinks.com/forum/topics/source-of-funding-for-professional-development
http://www.lislinks.com/forum/topics/dr-deepak-kumar-shrivastava-honored-by-african-award-of

Table 3. Harvested data of ten posts

Post title	Post date	Post time	Post user	Post category	Post view	Post reply count
Please Suggest Us to Read & Prepare for UGC - NET	17-Jun-19	10:28	Chitra	UGC NET / SET Examination	527	6
University Grants Commission Notification New Delhi, the 18th July 2018: UGC Regulations on Minimum Qualifications for Appointment of Teachers and Other Academic Staff in Universities and Colleges...	20-Jul-18	10:09	Mr. Bharat Sondarva	Other Discussions	4575	3
Publisher of Colon classification 7th Edition	15-Jun-19	8:43	Tara Singh	Other Discussions	257	1
OCLC Products and Services with Previous Questions from NTA / UGC NET Examination	13-Jun-19	15:44	Abilash Achuthan	UGC NET / SET Examination	390	1
How to Face Interview at Azim Premji University for Library Trainee	9-Jun-19	17:47	Supreet sakat	Interview	323	2
RSSMB Librarian Exam Date: 6th July, 2019	11-Jun-19	22:53	VIJAY PAL	Other Discussions	1092	1
CEPTAM of DRDO Document Verification: Will M.Lib. be Considered Instead of Diploma in LIS	7-Jun-19	14:08	PK PATRA	Other Discussions	454	7
Sticking Print-out of Library Bill in Stock Register	15-Jun-19	13:37	Bhargav Thaker	Other Discussions	188	1
Source of Funding for Professional Development	14-Jun-19	20:06	UGANDHA BAJAJ	Other Discussions	99	1
Dr. Deepak Kumar Shrivastava honored by African Award of Excellence -2019	12-Jun-19	12:09	Dr. D.K. Shrivastava	Award / Fellowship Notification	184	6

metadata. ICe is ".xg_headline-2l" for data related to main post like PostDate, PostTime, PostCategory, PostUser and ICe is ".xg_lightborder" for data related to reply of the post like PostRepliedUser, PostRepliedDate, PostRepliedTime.

The DH2 requires only two inputs, i.e. PostUrl and ICe. After passing the input, DH2 request the discussion forums and harvest the data. But DH2 harvest irrelevant information along with the data. To avoid the extraction of irrelevant information, a Data Cleaner is used which captures only relevant information and stores it in a database as raw data.

Table 4. Harvested reply data of one posts

Post replied by	Post reply date	Post reply time	Post reply body
PK PATRA	June 8, 2019	0:14	Any DRDO professional present here please reply to this ....
Jawed Akhtar	9-Jun-19	1:39	DRDO consider MLib degree for STA B post. Don't worry
PK PATRA	9-Jun-19	8:43	Conform,, no problems.. I have doubt because of higher qualifications not allowed.
Moin Raza	9-Jun-19	15:03	I Don't know the right answer but it was mentioned in advertise that all applications with higher qualifications will be rejected. So I think you are clear. Where you found Document verification date ?
Moin Raza	June 9, 2019	23:45	I mean there should be no problem. Cause one of my friend applied for mechanical post with B.tech degree and according to him he got enough marks to clear tier-1 but he was not called for tier-2 because of higher education. For more info you can watch videos of you tube channel " Mechanical Adda " about drdo.
KASINATH MISHRA	14-Jun-19	9:37	I have clear it from drdo ...they accepted mlib ...so dont worry
Moin Raza	16-Jun-19		Hii Kashinath mishra..Have you also provisionally selected for DRDO?

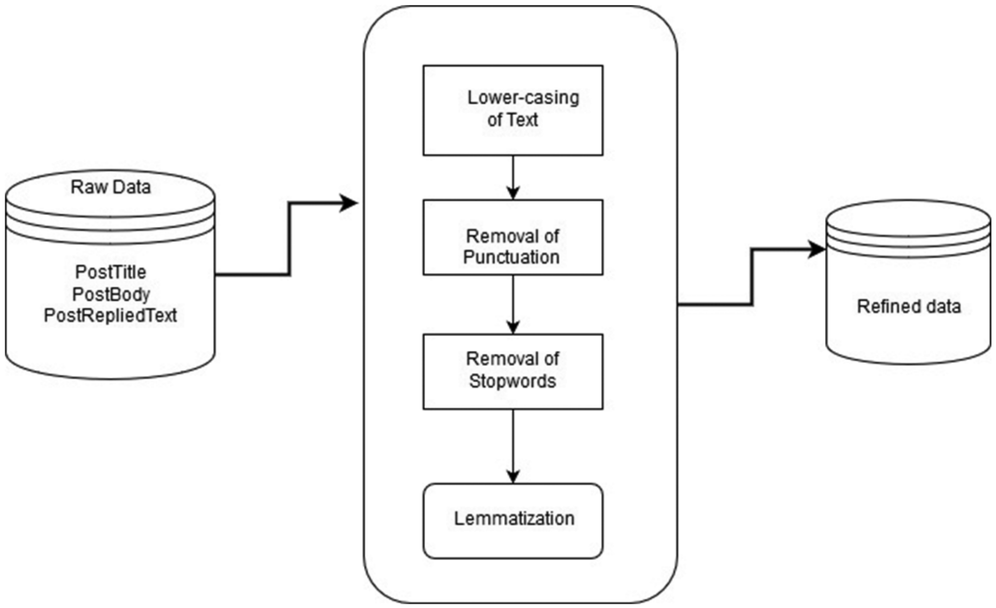
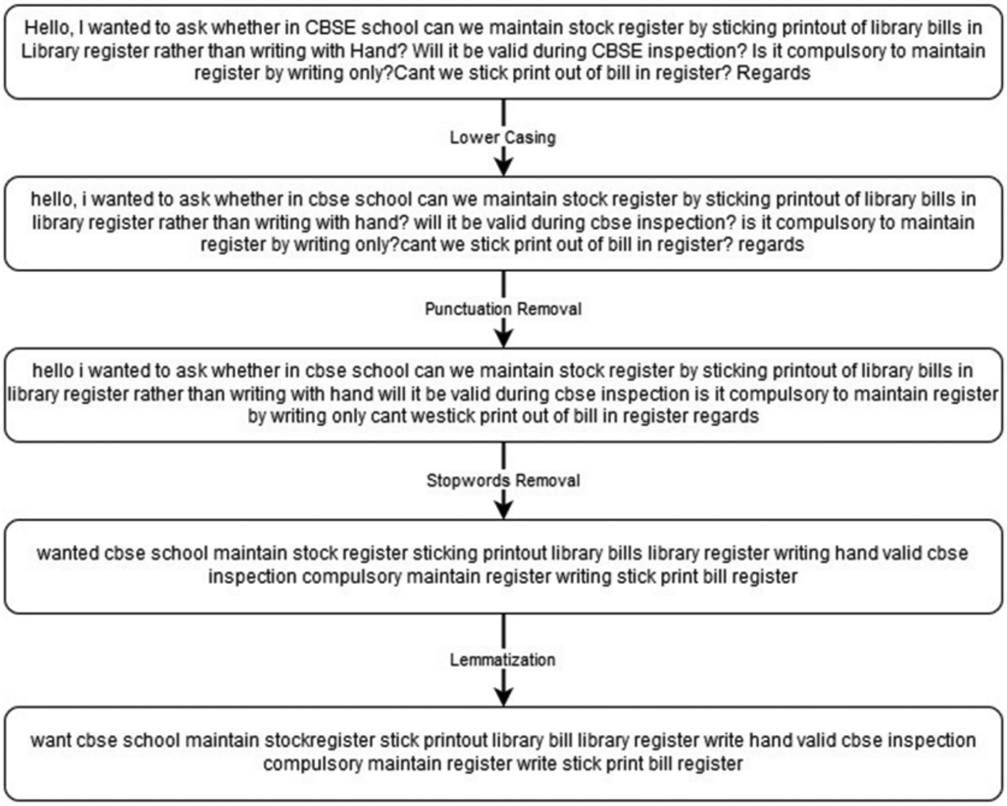


Figure 3. Data pre-processing phase.

Table 3 shows the harvested data of 10 posts on the first page on 17/06/2019. Table 4 present the reply thread of the seventh post of the first page.

The text data contain noises, which to be removed to make it analyzable. This step is known as pre-processing. The second phase of the framework is shown in Fig. 3. The data of PostTitle, PostBody, and PostReplyBody to be pre-processed as these contain text data. At first step, the whole text is converted in to lower case text, This is done to maintain uniformity of text of different styles. For example, "Library", "LibRary", "LIBRARY", or "library" are treated as four different words, but these words contain the same information. This writing style sometimes deviates the results, to avoid any deviation of a result, all text to be converted in to lower text. After lowering of text, the word will be "library"whose frequency of occurrence in the text is 04.

In English language punctuations are used for understanding and reading text perfectly. But this doesn't add extra information to the data. The removal of punctuation reduces the size of the data. The commonly used words like a, an, the, have, etc are known as stop words. These also don't add extra information about the text. The list of stop words of SMART (System for the Mechanical Analysis and Retrieval of Text) information retrieval system is the most popular in text analysis. This system was developed in the 1960s at Cornell University. The list contains 571 stop words. Apart from these, some stop words are domain-specific, which appear only in that domain. The words like dear, hello, please, regards, etc are some of the words which can be seen very often in discussion forums. These all stop words including domain-specific words were removed at the next step. At the last and final step, lemmatisation was done to map similar words with the root words. Lemmatisation is the process of morphological analysis to map the inflectional form of words to the root word



Hello, I wanted to ask whether in CBSE school can we maintain stock register by sticking printout of library bills in Library register rather than writing with Hand? Will it be valid during CBSE inspection? Is it compulsory to maintain register by writing only?Cant we stick print out of bill in register? Regards

hello, i wanted to ask whether in cbse school can we maintain stock register by sticking printout of library bills in library register rather than writing with hand? will it be valid during cbse inspection? is it compulsory to maintain register by writing only?cant we stick print out of bill in register? regards

hello i wanted to ask whether in cbse school can we maintain stock register by sticking printout of library bills in library register rather than writing with hand will it be valid during cbse inspection is it compulsory to maintain register by writing only cant westick print out of bill in register regards

wanted cbse school maintain stock register sticking printout library bills library register writing hand valid cbse inspection compulsory maintain register writing stick print bill register

want cbse school maintain stockregister stick printout library bill library register write hand valid cbse inspection compulsory maintain register write stick print bill register

Figure 4. Processed data.

based on the dictionary form of a word, known as the lemma. Lemmatisation is similar to stemming with the difference that it does not simply remove popular prefix or suffix from the words. In stemming, “library” and “libraries” are mapped to the word “librari” whereas in lemmatisation “library” and “libraries” are mapped to “library”. The lemmatisation is an advanced stemming process with the function of the dictionary look-up. It maps the word to root word as per the dictionary.

After this whole process, the refined data is stored in the database for the analysis. Figure 4 shows the step by step of the pre-processing of data.

5. DISCUSSION

The rapid increase in the usage of discussion forums made these as a knowledge repository. The evaluation of this knowledge can help us to understand the behavioral pattern of the community about a given topic. But the biggest challenge in evaluating these repositories is how to extract and analyse this unstructured form of data. In this paper, we presented a framework to extract data from LisLinks Forums and pre-processing of it. The framework is divided into two phases, i.e. Harvesting Phase and Pre-processing phase. Fourteen Metadata elements were defined for forum’s dataset. An information container of thirteen metadata elements was located on the forum web page. Two data harvester,i.e. DH1 & DH2 and one data cleaner (DC) were developed in R. DH1 to take base URL, ICe and Page Range as input and extract the PostUrl of all post of a given page range. DH2 takes this PostUrl and ICe of the other twelve metadata as Input and harvest the desired data. DC removes irrelevant data harvested by DH2 and stores it in a database. The harvested data of PostTitle, PostBody, PostRepliedBody were pre-processed at the second phase. Text lowering, Removal of stop words and punctuation, and lemmatisation are used to remove the noise from the dataset. The framework was implemented in R language and its packages like Rweka, tm, rvest, stringr, stringi, etc. were used at harvesting and pre-processing phase.

6. CONCLUSIONS & FUTURE WORK

It is observed that the present framework provides a convenient interface for harvesting and processing user-generated data available on web discussion forums. It will automate the process of harvesting of text data. Further, the harvested data can be cleaned and pre-processed for analysis. This framework can directly be used for the discussion forums which have a structure similar to LisLinks. However, it can also be used for other discussion forums like Yahoo Answers, Trip Advisors, etc with few customisations based on the structure of the webpage. As of future work, the processed data will be analysed using different statistical algorithms like LDA, semantic analysis, etc.

REFERENCES

1. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#b21203c60ba9 (accessed on 30 June 2019).
2. Hoogeveen, D.; Wang, L.; Baldwin, T. & Verspoor, K.M. Web forum retrieval and text analytics: A survey. *Found. Trends® in Inf. Retr.*, 2018, **12**(1), 1-163.
3. Liu, X.; Liu, J. & Chen, H. Identifying adverse drug events from health social media: A case study on heart disease discussion forums. *In* International conference on smart health, July 2014,Springer, Cham, pp. 25-36.
4. Calefato, F.; Lanubile, F. & Novielli, N. How to ask for technical help? Evidence-based guidelines for writing questions on Stack Overflow. *Inf. Software Technol.*, 2018, **94**, 186-207.
   doi: 10.1016/j.infsof.2017.10.009.
5. Sahu, T.P.; Nagwani, N.K. & Verma, S. Multivariate beta mixture model for automatic identification of topical authoritative users in community question answering sites. *IEEE Access*, 2016, **4**, 5343–5355.
   doi: 10.1109/ACCESS.2016.2609279.

6. Aikawa, N.; Sakai, T. & Yamana, H. Community QA question classification: Is the asker looking for subjective answers or not? *IPSJ Online Trans.*, 2011, **4**, 160–168. doi: 10.2197/ipsjtrans.4.160.
7. Dearman, D. & Truong, K.N. Why users of yahoo!: Answers do not answer questions. *In* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 2010, ACM, pp. 329-332. doi: 10.1145/1753326.1753376.
8. Chen, L.; Zhang, D. & Mark, L. Understanding user intent in community question answering. *In* Proceedings of the 21st International Conference on World Wide Web, April 2012, ACM, pp. 823-828. doi: 10.1145/2187980.2188206.
9. Wen, M.; Yang, D. & Rose, C. Sentiment analysis in MOOC Discussion forums: What does it tell us?. *Educ. Data Min., 2014*.
10. Zhou, B.; Jia, Y.; Liu, C. & Zhang, X. A distributed text mining system for online web textual data analysis. *In* 2010 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, October 2010, IEEE, pp. 1-4.
11. Chai, K.; Hayati, P.; Potdar, V.; Wu, C. & Talevski, A. Assessing post usage for measuring the quality of forum posts. *In 4*th IEEE International Conference on Digital Ecosystems and Technologies - Conference Proceedings of IEEE-DEST 2010, DEST 2010, 233–238. doi: 10.1109/DEST.2010.5610640.
12. Baldwin, T.; Martinez, D.; & Penman, R.B. Automatic thread classification for linux user forum information access University of Melbourne. *In* Proceedings of the 12th Australasian Document Computing Symposium, Melbourne, Australia, 10 December 2007, 72–79.
13. Choi, E. Motivations and expectations for asking questions within online Q&A. 2013. http://citeseerx. ist. psu. edu/ viewdoc/download.
14. Carenini, G.; Ng, R.T. & Zhou, X. Summarizing email conversations with clue words. *In* Proceedings of the 16th international conference on World Wide Web, May 2007, ACM, pp. 91-100. doi: 10.1145/1242572.1242586.
15. Wang, Z. & Zhang, P. Examining user roles in social Q&A: The case of health topics in Zhihu.com. *In* Proceedings of the Association for Information Science and Technology, 2016, **53**(1), 1–6.
16. Fu, H. & Fan, Y. Music information seeking via social Q&A: An analysis of questions in music StackExchange community. *In* JCDL2016: Proceedings of the Joint Conference on Digital Libraries, 2016, ACM Press, New York, pp. 139-142.
17. Choi, E. & Shah, C. User motivations for asking questions in online Q&A services. *J. Assoc. Inf. Sci. Technol.*, 2016, **67**(5)**,** pp. 1182-1197.
18. Helic, D.; Maurer, H. & Scerbakov, N. Discussion forums as learning resources in web-based education. *Adv. Technol. Learn.*, 2004, **1**(1), 8-15.
19. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#8ca18276f637.(accessed on 30 June 2019).

CONTRIBUTORS

Mr. Mohit Garg is working as Assistant Librarian in Indira Gandhi National Tribal University (A Central University) Amarkantak, MP-484887, India. He is pursuing PhD from School of Social Science, Indira Gandhi National Open University under the guidance of Prof. Uma Kanjilal. His area of interest are Information Retrieval, Data Science, Quantitative analysis and Machine learning.

Mr. Mohit contributed in finding the knowledge gap by reviewing the related literature, developing the framework and completing the initial draft of the paper.

Prof. Uma Kanjilal is Professor and Head in the Faculty of Library and Information Science in the Indira Gandhi National Open University (IGNOU) is also holding charge of Director of Centre for Online Education (COE) at IGNOU. She has more than 29 years of experience in the Open and Distance Learning System. Her specialization includes e-learning, multimedia courseware development, ICT applications in Libraries and Digital Libraries.

Her contribution to the current study is the conceptualisation, improvement in contents and preparation of the final draft of the paper.