1-2021

Discovering Topics from the Titles of the Indian LIS Theses

Sourav Mazumder
University of North Bengal, smazumderlis91@gmail.com

Tapan Barui
University of North Bengal, tapanbarui@nbu.ac.in

Discovering Topics from the Titles of the Indian LIS Theses

Sourav Mazumder[1] and **Dr. Tapan Barui**[2]

[1]Former student, Department of Library and Information Science, University of North Bengal, West Bengal, India, email: smazumderlis91@gmail.com, ORCID ID: https://orcid.org/0000-0003-0956-661X

[2]Assistant Professor, Department of Library and Information Science, University of North Bengal, West Bengal, India, email: tapanbarui@nbu.ac.in , ORCID ID: https://orcid.org/0000-0002-8023-4987

Abstract:

A lot of text data is being generated on the web in the form of scholarly articles, doctoral thesis, social media, library databases, and data archives. They are easy to use but complicated to process for research works. That is exactly why text mining is required and topic modeling is one of the most important techniques involved in text mining. In this paper, an attempt has been made to discover topics from the thesis titles (uploaded theses) in the field of Library and Information Science (LIS). For this work, the text data (n=2132) has been obtained from the Shodhganga. Then, topic modeling through Latent Dirichlet Allocation (LDA) has been applied. After employing preliminary investigation, the findings show: State universities of India have the highest contribution of the thesis (78.06%); most theses (106) belong to Karnatak University, and 60.83% of thesis falls under the period 2011-2020. The main results of this paper are (a) The keyword "library" (0.204) has the highest score regarding 10 topics and "Library use" can be inferred as the major topic; (b) the keywords "information", "technology", "communication", "survey", "comparative", "plant", "scientist", "city", "support", and "small" were discussed over 266 titles; and (c) "study", "university libraries", and "information-seeking behaviour' are the most frequent n-grams appeared in the titles. This work can be taken towards future research for more improvement and new applications.

Keywords: Topic modeling**,** Natural Language Processing, Text Mining, Data Mining

1. Introduction

Information Communication Technologies (ICTs) are massively used in almost every activity of human life. As the days go by, information is being generated rapidly. More specifically, they are being stored online through ICTs. Such information relates to scholarly communication, news, blogs, and wikis. In the social context, we can recognize the value and the application of the information. The necessity of information is entrusted by the increasing relationship between the economy of information and nations (Ifijeh, 2010). Therefore, we are witnessing information explosions too. Information explosion is simply defined as the rapid growth of various published literature (e.g Huth, 1989; Major & Savin-Baden, 2012). Hence, it opens the door for us to discover more useful information from the existing literature which is in large form. Villars *et al.* (2011) state the growth in the number, size, and importance of information assets is not limited to just large government agencies, large enterprises, and internet websites. This is where the concept of big data originates. Its rise can be traced to a research work of Gandomi and Haider (2015). A couple of issues can be aroused regarding adequate analysis of large data to discover meaningful information and finding appropriate methods to do it. There are some techniques called Data Mining and Text Mining to deal with large or big data. Nowadays, Text Mining has become a *"*growing interest*"* across the social sciences and digital humanities, medical sciences, and computational sociology and many more (Bernard & Ryan, 1998; Miller, 2018). Especially, text mining allows end-users to extract data from databases, books, and social media to establish significant information. Text mining is a machine learning technique and Hoth *et al.* (2005, p. 5) tried to differentiate three specific perspectives of text mining: information extraction from texts, text data mining, and knowledge discovery in database (KDD) processes. A bag of word identification; extraction and visualization of the most frequent word using a word cloud; topic modeling; text-clustering; and analysing statistical values of the text are instances of text mining applications (Goswami *et al*., 2021). This paper deals with topic modeling of titles of theses of library and information science (LIS) submitted to the various universities of India. In general, it is not an easy task to analyze text data adequately without any earmarked technique. In the context of text mining, topic modeling is one of the superior techniques to find abstract topics from a given text. According to Buenaño-Fernandez *et al.* (2020, p. 35319) "topic modelling focuses on the grouping of text documents, assuming that each document is a function of latent variables entitled topics."

The remainder of this paper is organized as follows. Section 2 combines some related works on topic modeling. Section 3 shows the objectives of this work. Section 4 describes the methods and materials used in this work. In section 4.2, a workflow has been illustrated. Section 5 discusses the results from each segment of the present work. Section 6 describes some limitations and further research direction. Finally, Section 7 presents the conclusion of this paper.

2. Related literature

Many studies have been conducted to analyze large data by topic modeling technique (Barde & Bainwad, 2017; Buenaño-Fernandez *et al*., 2020; Nikolenko *et al*., 2017; Tong & Zhang, 2016). Most specifically, some similar studies are systematically reviewed in this section. Jelodar *et al.*(2019) surveyed scholarly articles on topic modeling to investigate various "proposed models" based on Latent Dirichlet Allocation (LDA), "current trends", and "intellectual structure of topic modeling." Also, they introduced various tools and datasets in topic modeling. In another study, Han (2020) used LDA to examine research topics in LIS. The author collected 14035 documents from 1996 to 2019. An enhanced data-selection strategy was devised by the author to provide a dynamic journal list. According to the findings of the study, there were less top topics related to "library issues" during 2000-2005, rather "bibliometrics", "information retrieval", and "organizational activities" emerged heavily as research topics. The study also discovered that conventional technological context began to change due to the appearance of social media and mobile technologies. Qualitative research on sociological aspects, opinion analysis, and media can be benefited by using topic modeling techniques. Based on these considerations, Nikolenko *et al.* (2017) identified two problems regarding human judgement in grasping the topics and revealing the specific subtopics. For that, the authors firstly proposed "TF-IDF coherence metrics" for overcoming human judgement adequately. Secondly, they brought the interval semi-supervised approach (ISLDA) for topic extraction. Tong and Zhang (2016) conducted an experimental study to explore topics from Wikipedia and Twitter data. They researched and proposed a model based on LDA for text mining solutions. Furthermore, they explored topic probabilities, terms, top-10 shortest distances of a Wikipedia article. Newspapers are one of the important sources of information for scholarly communication. Since newspapers have a high monetary worth, Yang *et al.* (2011) tried to find the most prominent topics using automatic clustering from newspapers published in Texas from 1829 to 2008. They found the top-10 topics from the 1865-1901 dataset.

Additionally, they discovered topics that associated more strongly with the word "cotton". Wang and Blei (2011, pp. 448–456) stated some issues finding relevant research publications on online databases. They developed an algorithm that "combines the merits of traditional collaborative filtering and probabilistic topic modeling". They investigated the CiteULike (social bookmarking service) dataset and showed the algorithm's effectiveness for collaborative filtering with latent structure.

Goswami *et al.* (2021) worked on text mining of biomedical literature where they tried to find co-occurring keywords. They discovered 10 different latent topics with positive and negative weights. In another study, Sun and Yin (2017) applied LDA to discover 50 key topics from the abstracts of 17,163 articles on transportation from 22 journals. Their study explored the context of each topic and the variation of topics across the country, journals, time, and networked word co-presence. The results of their study revealed that most journals cover different topics. In addition, their research can provide a classification scheme for transportation research. Social media like Twitter is one of the vital information sources for human beings. Hong and Davison (2010) asserted that Twitter messages can be understood using standard text mining techniques. Since Twitter messages are restricted, they cannot be employed fully for exploring anything further. Then, the authors addressed issues using different topic models. As a result, they devised a few schemes to train models and compared the quality of the model based on qualitative and quantitative aspects.

All of the work was done with different perspectives in different subject domains. In the field of LIS, there is least work on topic modeling for doctoral thesis. This paper tries to present topics from the Indian LIS theses using the LDA. The main objectives are specified in the following section.

3. Research objectives

The main objectives of this paper are:

i. to explore latent topics from the text of thesis titles;
ii. to visualize the topics using multidimensional scaling;
iii. to infer topics with the most representative titles;
iv. to discover the frequency of the distribution of the topic keywords across all titles; and
v. to identify the frequency of the n-grams.

4. Methodology and materials

4.1 Data collection and dataset preparation

For this study, we surveyed and collected data of 2132 theses (dated 15/01/2021) using two searching parameters: subject search (https://sgsubjects.inflibnet.ac.in/) and department search on the Shodhganga (https://shodhganga.inflibnet.ac.in/), a digital repository of Indian theses and dissertations maintained by the Information and Library Network Centre (INFLIBNET Centre) and put them into a spreadsheet (Google Sheets) to create a dataset (Fig.1). The dataset includes information about title of the ID, thesis title, scholar, guide, department, university/library, and year. We analyzed the text of the thesis titles to discover topics.

4.2 Data pre-processing, modeling and visualization

We used Google Colab IDE (Integrated Development Environment), a web-based notebook for running Python codes. We applied the LDA topic modeling method to uncover the latent topics using the `Gensim` library package. Gensim is used for topic modeling, document indexing and similarity retrieval (Rehurek & Sojka, 2010). The dataset contains words in uppercase or lower-case, stopwords, punctuations, and unnecessary characters. So, we defined functions to clean, preprocess using `NLTK` (Perkins, 2011) and tokenize text data. Furthermore, we lemmatized the text data using the `spaCy` library package (Srinivasa-Desikan, 2018) for keeping the words that belong only to nouns, adjectives, verbs, adverbs, person, language, quantity, and geographical location. After stimulating the vocabulary size, we trained the LDA model. For topic visualization, `pyLDAvis` (a python library) was used to show topics and the frequency of the keywords belonging to each topic. This library is a port of the "R" package where Sievert and Shirley (2014) introduced LDAvis, a web-based interactive visualization of topics estimated using Latent Dirichlet Allocation. It was developed with the combination of R and D3.The following figure (Fig.1) displays two different workflows in a single frame.

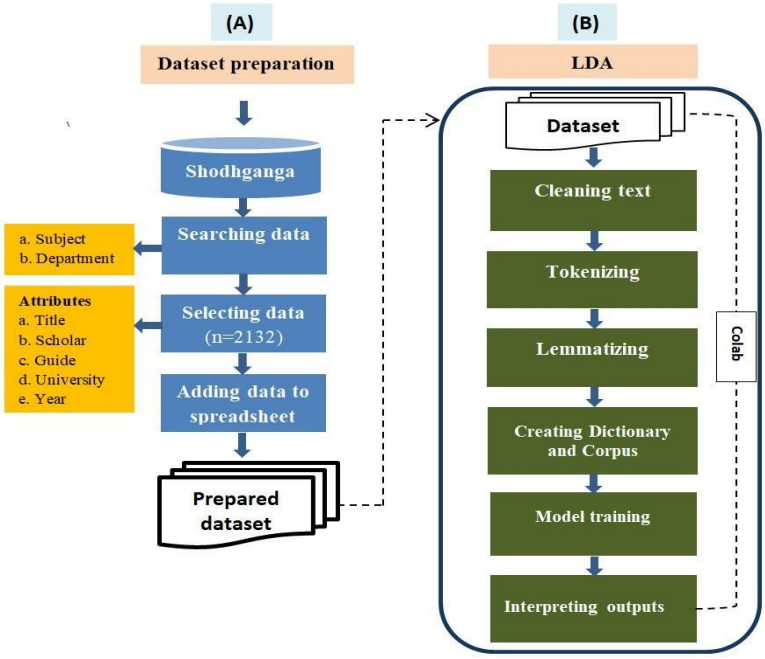

Fig.1: Workflow: (A) Dataset preparation and (B) LDA based topic modeling.

5. Results and discussion

5.1 Preliminary investigation of the uploaded theses on the Shodhganga

The preliminary investigation conveys different aspects of the thesis contributing universities. According to the official website of the repository, the actual number of contributing universities is approximately 490, with more than 300000 theses spanning various subject fields. If we access Shodhganga, we can find over 3500 unique departments that contribute full-text theses. It has also been found that several subjects, such as Chemistry, Engineering, Education, and Physics have the highest number of theses. The University of Madras is among the top universities that contributed most theses (approx 12515). This paper only brings up the number of universities that have contributed LIS theses. Table 1a and 1b provide the information related to the types of contributing universities and the number of available theses on Shodhganga.

Table 1a: Contributing universities

Types	Count	%
Central University	14	12.39
State University	68	60.18

Deemed University	12	10.62
Private University	19	16.81
Total	**113**	

According to this study, four different types of universities' theses have been uploaded to Shodhganga. Those universities are central university (12.39%), state university (60.18%), deemed university (10.62%) and private university (16.81%).

Table 1b: Thesis contribution by the universities

Types	**Count**	**%**
Central University	286	13.41
State University	1664	78.05
Deemed University	127	5.96
Private University	55	2.58
Total	**2132**	

This table shows that most of the theses (more than three quarter of all) belong to the State universities (78.05%) and this is followed by Central universities (13.41%), Deemed universities (5.96%) and Private universities (2.58%).

Table 2: Frequency count of Top 10 universities having the most number of theses available on Shodhganga

s/n	**Name of the university**	**Count**
1	Karnatak University	106
2	Manonmaniam Sundaranar University	106
3	Bharathiar University	82
4	Savitribai Phule Pune University (University of Pune)	80
5	Sri Venkateswara University	78
6	University of Mysore	69
7	Bharathidasan University	66
8	Aligarh Muslim University	62

9	Dr. Babasaheb Ambedkar Marathwada University	59
10	Alagappa University	57

As per our online survey on Shodhganga, it has been found that there are 113 contributing universities' theses archived on Shodhganga (up to January 15, 2021). Statistically, we compute the mean value of the frequency count of the thesis contributions from the top-10 universities is 76 (while mean= 18.7 for all 113 universities) and the standard deviation (SD) of top-10 universities is 18 (while SD= 22.5 for all 113 universities). The highest number of theses (jointly count:212) is found in Karnatak University and Manonmaniam Sundaranar University followed by Bharathiar University (82), Savitribai Phule Pune University (80), Sri Venkateswara University (78), University of Mysore (69), Bharathidasan University (66), Aligarh Muslim University (62), Dr. Babasaheb Ambedkar Marathwada University (59), and Alagappa University (57). In addition, Karnatak University has contributed a total of 4572 theses and the number of LIS thesis is 106 (2.31%). On the other hand, the percentage of the uploaded theses of the University of Madras is 0.25 which is proportionately less.

Table 3: Frequency count of awarded theses over the years from 1969 to 2020 (based on the archived metadata)

s/n	**Years**	**Count (n=2132)**	**%**
1	1969-1980	9	0.42
2	1981-1990	40	1.88
3	1991-2000	150	7.04
4	2001-2010	470	22.05
5	2011-2020	1297	60.83
6	No Date (n.d)	166	7.79

It shows the count of 2132 theses awarded-year based on upload by Shodhganga. First, a very few theses have been found during the period 1969-1980. But we can see the change after 1981. The highest numbers of theses (1297) found in the years between 2011-2020 followed by 2001-

2010 (470), no date (166), 1991-2000 (150), and 1981-1990 (40). One of the most notable points is that the thesis awarded dates are not mentioned in the metadata descriptions of 166 theses (7.79%) on the repository.

5.2 Topic modeling using LDA

LDA is a "generative probabilistic model" which represents discrete texts. It is based on the Bayesian model. The collection of documents is a mixture of latent topics distinguished upon words or items (Blei *et al*., 2003). In other words, LDA finds different types of topics within a given text data. In this paper, the LDA model has been built to discover 10 topics. However, it can not determine a topic name of a certain document. Instead of a topic name, it provides us with a string of keywords associated with a certain document, such as the thesis title. In other words, a topic model cannot detect any topic, only humans can comprehend and identify a specific topic by inferring the keywords. Hence, we inferred 10 topics based on their keywords in Table 4. Each topic has ten different keywords associated with it, which are presented in order of their highest weightage.

Table 4: List of 10 topics along with their keywords

Topic	Topic Name	Keywords
0	Library use	library, study, information, use, resource, service, libraries, research, electronic, access
1	Open-source	student, digital, open, source, industry, medium, web, internet, indian, database
2	Management	user, management, approach, government, medical, search, evaluative, software, skill, new
3	University library	university, work, model, case, develop, state, analytical, change, library, assessment
4	Information seeking behaviour	information, behaviour, seek, district, need, literacy, pattern, base, people, habit
5	Collection development	special, reference, development, system, woman, agricultural, academic, scientist, design, collection
6	Performance evaluation	public, evaluation, awareness, performance, region, journal, economic, classification, knowledge, growth
7	ICT	impact, select, technology, analysis, communication, school, english, survey, narrative, novel
8	Higher education	college, engineering, education, application, affiliate, tamil,

		institution, ict, nadu, high
9	Librarianship	science, faculty, environment, professional, member, universities, social, satisfaction, librarian, job

Topic 0 (python generated the list with 0 as the first number) can be exemplified for interpretation. For instance, it represents the keywords with their weights:

Box 1: Topic 0 and its weights

(**0**,
'0.204*"library" + 0.202*"study" + 0.108*"information" + 0.084*"use" + '
'0.041*"resource" + 0.041*"service+ 0.034*"libraries" + 0.030*"research" +
"0.023*"electronic" + 0.019*"access"')

Among the 10 keywords in Topic 0, the keyword "library" has the highest weight (0.204). It shows that it is the most representative keyword. On the other hand, "access" has less weight and can be considered less representative than the other nine keywords. Topic 0 might be inferred to "Library Use" or "Information resource." Though, it is completely based on human judgement. The keywords are the most often used for a certain subject, such as "library use".

For each topic, each keyword has a distinct weight. For example, when the keyword "library" has a weight of 0.204 (20.4%) in Topic 0, the same keyword has a weight of 0.020 (2%) in Topic 3. So, it is relatively less representative for topic 3. We can discover "University library" as Topic 3. Another example is that in Topic 4, the keyword "information" might represent a different topic than Topic 0. The keywords suggest that the topic can be "Information-seeking behaviour." "Management," "Collection development," "Performance evaluation," "ICT," "Higher education," and "Librarianship" are the other topics. Furthermore, topic coherence (Gan et al., 2018) is also computed to measure the coherence score of the topics. The coherence value is 0.54.

5.3 Topics visualization

Fig.2 shows the visualization of topics. The first topic (bubble 1) is not numbered by 0. Topics with precise numbers are shown by `pyLDAvis` (from 1 to 10). It can be defined with two different approaches: term saliency (1) and terms relevance (2) (Chuang *et al*., 2012; Sievert & Shirley, 2014). The mathematical expressions behind the two approaches are given below.

Term saliency is described as:

$$saliency(w) = P(w) \times \sum_{T} P\,(T|w)\ log\frac{P(T|w)}{P(T)} \quad (1)$$

Where w= given word; T= topic; $P(T|w)$ = conditional probability; $P(T)$= marginal probability*;* *distinctiveness* is between $P(T|w)$ and $P(T)$.

Relevance is described as:

$$r(w,k|\lambda) = \lambda \log(\phi_{kw}) + (1-\lambda)\, log\left(\frac{\phi_{kw}}{P_w}\right) \quad (2)$$

Where λ= weight term w under topic k; ϕ_{kw}= probability of term w for topic k empirical corpus distribution.

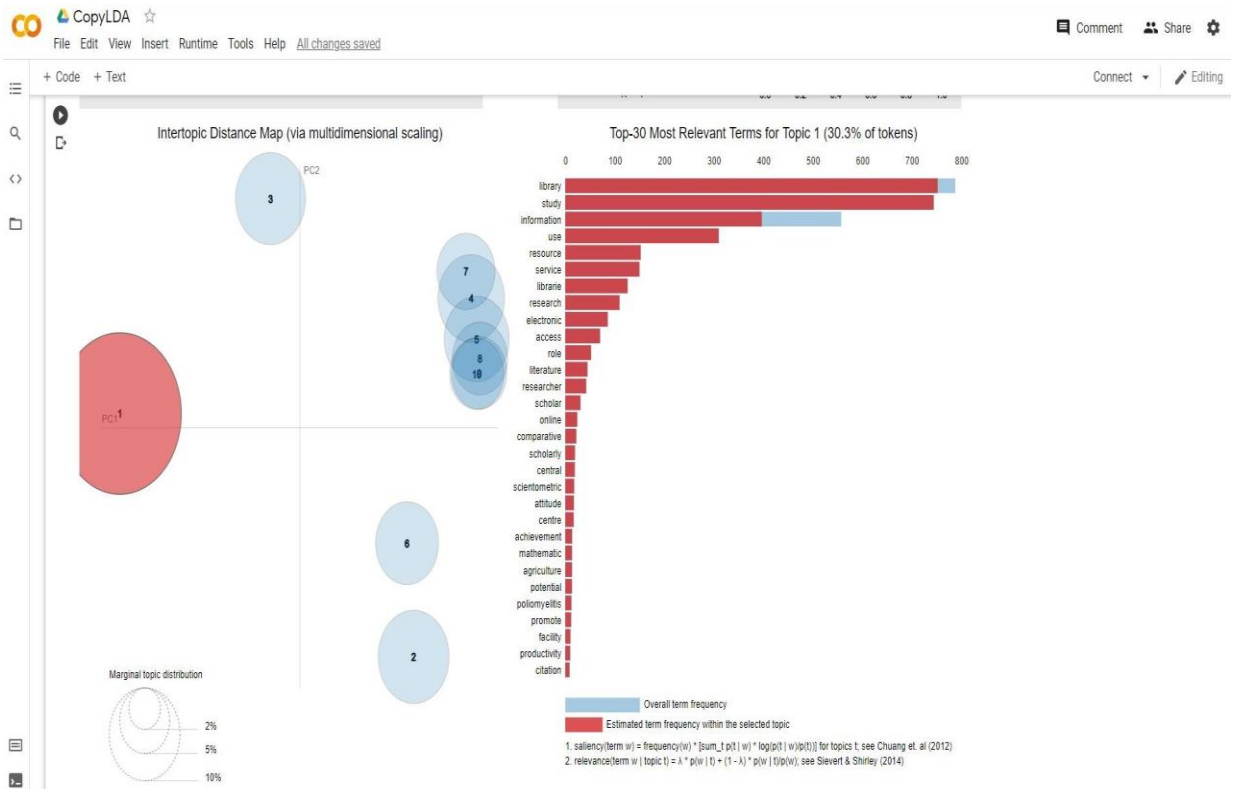

Fig.2: Visualization of topics and keywords (The numbers inside the bubbles are the topic numbers)

The information in this Fig.2 pertains to two panels. The left panel comprises two-dimensional planes (bubbles) of 10 topics. The right panel (bars) deals with top-30 most frequent terms to the

10 topics. There are two panels that are connected to each other. On the right panel, there are also two different coloured bars. In the left panel, blue bars (overall term frequencies) determine the list of all keywords under 10 topics, while red bars (estimated word frequency inside the selected topic) display the specific words of that topic. Topic 1 is used in this figure to look at the frequency of the keywords. It is the largest bubble on the plane, and it is responsible for the majority of the tokens. It also reveals that "library" (≥ 700) has the highest frequency, followed by "study" (≥ 700), "information" (≥ 400), "use" (≥ 300), "resource" (≥ 100), and so on. The majority of the titles are associated with Topic 1. Other 9 topics may be carried out in the same way by picking the bubbles on the global plane. In comparison to the Fig.2, Fig.3 illustrates the frequency of comprehensive keywords. It is the default "relevance" metrics of the 30 keywords and it will not show any estimated term frequency until a topic is selected in the left panel.

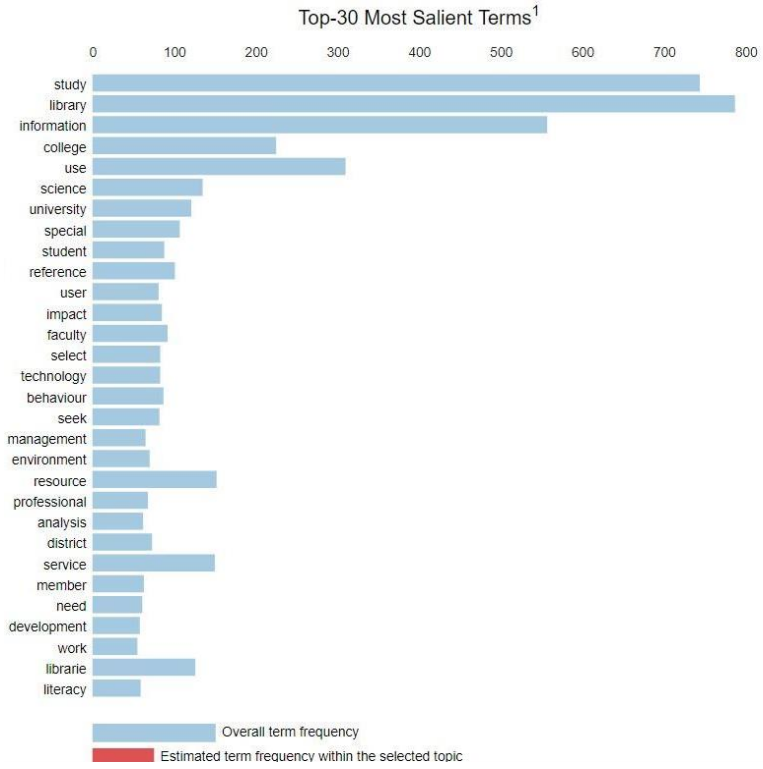

Fig.3: Overall frequency of the top-30 terms.

5.4 Inferring topics with the titles

Table 5: List of selected original titles (most representatives) with topic keywords

s/n	Keywords	Original titles (text)	Topic name	TP*
1	research, literacy, study, universitie, output, scholar, mapping, skill, ict, productivity	*Scientific research in Central universities of northern India: a bibliometric analysis of research output during the period 2000-2008*	Science mapping	11.62
2	library, access, open, source, environment, software, journal, marketing, learning, product	*Open-source software open access resources enhancing library services: An exploratory study*	Open-source software	12.60
3	library, professional, work, satisfaction, job, problem, prospect, automation, personnel, perspective	*Work motivation job satisfaction organisational commitment among LIS professionals first grade college libraries North Karnataka: A study*	Job satisfaction	15.91
4	information, student, habit, school, internet, district, teacher, search, medium, read	*Reading Habit Among The Secondary School Students In Mysuru District: A Study*	Reading habit	14.27
5	system, public, information, community, thesis, rural, develop, submit, online, citation	*A study public library system community information centers Nagaland realities challenges*	Public library	12.69

*TP: Total percentage; sources of the 5 thesis titles are shown in the Appendix 1.

This section provides a comprehensive overview of the titles that are represented as inferring aspects of topics. In section 5.2, we discovered a distribution of ten topics and their keywords. It may, however, be unable to convey explicit sights in order to understand a certain topic. Titles (as text) might be considered as the representatives of topics to prevail over such obstacles. The distribution of Table 5 is the same as the original output format. Column five indicates the weights of each topic as "Topic Percentage". It is contingent on the contributions of all of the titles' topics. For instance, if we take the first title (title in full: "*Scientific research in Central universities of northern India: a bibliometric analysis of research output during the period 2000-*

2008"), we can assimilate multiple topics such as "science mapping", "bibliometrics", and "scientometrics". It is extensively an ideal amid all the titles to represent the two topics and the TP of contribution is 11.62%. The need for this demonstration is very obvious. It simply sets an example for a thematic aspect of the text. If we look at the green-colored keywords (s/n 2) in the table (library, access, open, source, software, product), we can see that the text is about "open-source software." The second title may be regarded to be the most representative. Undoubtedly, this makes it very easy to understand how to perceive a topic with the help of keywords. Similarly, other 4 titles could be labelled as "Job satisfaction" (s/n 3), "Reading habit" (s/n 4), and "Public library" (s/n 5).

5.5 Distribution of the most important topic keywords

Table 6: Frequency count of top-10 topic keywords with the highest number of documents

s/n	Keywords	Count	OP*
1	information, technology, communication, survey, comparative, plant, scientist, city, support, small	266	12.48%
2	library, evaluation, web, plan, collection, performance, status, website, region, initiative	140	6.57%
3	information, seek, behaviour, centre, medical, scientist, behavior, coimbatore, requirement, package	133	6.24%
4	information, student, habit, school, internet, district, teacher, search, medium, read	128	6%
5	college, faculty, engineering, member, affiliate, aid, graduate, affiliated, pune, stress	125	5.86%
6	reference, special, user, attitude, libraries, scientific, activity, measure, strategy, profession	120	5.63%
7	science, study, education, indian, service, art, critical, legal, sector, implementation	113	5.30%
8	system, public, information, community, thesis, rural, develop, submit, online, citation	112	5.25%
9	study, libraries, application, quality, survey, economic, tool, industry, total, level	102	4.78%
10	library, professional, work, satisfaction, job, problem, prospect, automation, personnel, perspective	97	4.55%

*OP: Overall percentage (%)

This table shows the distribution of the top-10 frequently discussed topic keywords across 1336 (63%) titles. The actual list of topic keywords is 20, while the remaining 10 keywords cover a total of 796 (37%) titles. The first string of topic keywords: “information,” “technology,” “communication,” “survey,” “comparative,” “plant,” “scientist,” “city,” “support,” and “small” were discussed 266 (12.48%) times in 2132 titles. As a result, the keywords may be used to find a wide range of topics. If three terms (technology, information, and communication) are concatenated, we may clearly understand that the topic is “Information communication technology.” The frequency from the second to the fourth topic keyword strings is nearly identical ranging from 6 to 6.5%; fifth to eighth strings are under one cluster which incorporates the enumeration of frequency from 112 to 125 (5.25% to 5.86%); and the last two strings have the frequency from 97 to 102 (4.55% and 4.78%)

5.6 Text analysis

Table 7: Frequency count of top-10 n-grams

s/n	1-gram	Count	2-grams	Count	3-grams	Count
1	study	1037	university libraries	194	information seeking behaviour	86
2	information	741	college libraries	156	library information science	60
3	libraries	684	special reference	151	engineering college libraries	44
4	university	484	information seeking	111	information communication technologies	43
5	library	479	library information	102	libraries special reference	24
6	india	281	seeking behaviour	90	college libraries affiliated	23
7	research	235	analytical study	80	electronic information resources	22
8	use	232	comparative study	79	libraries tamil nadu	22
9	college	209	case study	79	study special reference	22
10	services	208	faculty members	73	libraries andhra pradesh	21

This section shows some form of text analysis of the titles. Text analysis is the process of analysing text data in order to extract useful information. It also allows obtaining quantitative outlines in the corpus. Nevertheless, text analysis and text mining can be considered similar techniques. A simple text analysis includes measuring word frequency, concordance, text classification, etc. (Bird *et al*., 2009). Measuring statistical properties (frequency) of words or sequence of words based on n-gram models has been utilized in this paper. The main purpose of showing it is to highlight the most frequent terms in the titles. It also aids in summarizing topical contents. In addition to topic modeling, frequency computation of terms provides important information. The table presents the list of the top-10 n-grams. First, it depicts 1-gram (also known as unigram) and the result reveals that the scholars used the word "study" 1037 times. Conspicuously, it can be conjectured that this is used more as a particular concern. For instance, it has been found that the word "study" is used 400 times in the subtitles. Next, the keyword "information" appears 741 times in the titles, followed by "libraries" (684), "university" (484), "library" (479), "India" (281), "research" (235), "use" (232), "college" (209), and "services" (208). Second, the list of 2-grams (also known as bigrams) shows a different result from the previous list. It evolved to meaningful term such as "university libraries" which denote that scholars had worked on university libraries 194 times. The second term "college libraries" (156) is also prevalent in the context of research areas and followed by "special reference" (151), "information-seeking" (111), "library information" (102), "analytical study" (80), "comparative study" (79), "case study" (79), and "faculty members" (73). Third, the results of the 3-grams (also known as trigrams) model are similar to the bigrams. This is a more organised approach to learning a subject matter. This group of words "information seeking behaviour" (86) makes more intellect, as it is almost similar to "information seeking" and "seeking behaviour". The only difference is that this term is more meaningful than the bigrams. Similarly, "library information science" (60), "engineering college libraries" (44), "library information science" (60), and "study special reference" (22) come under the same aspect. All of the n-grams are illustrated in Fig.4. It can be observed that the terms "study", "university library", and "information-seeking behaviour" are in bigger form as they appeared more frequently.

Fig.4: Word Cloud of most frequent words

We also can match these n-grams with our topics or keywords listed in the Table 4. For example, if we take the first 3-gram "information seeking behaviour" then we can match it with Topic 1 and Topic 4 directly. As "library use" is a part of user studies. However, n-grams cannot be used to evaluate all topics, but we may use them to understand the frequency of the terms that appear often in the titles.

6. Limitations and future research direction

There are certain limitations of this study. Some of them include the size of the sample, availability of theses, and experimental aspects. In terms of the actual number of LIS theses, this study only finds those theses that are available on the repository and discovers topics. It can be assumed that the actual number of awarded Indian theses is higher than the collected data. Then there were several issues with executing programmes to obtain results. First, Google Colab is an internet-based IDE and so it is slow compared to desktop Jupyter Notebook. Though, it has been used to avoid installation or downloading heavy packages. Second, the periodic version require-

ment of some libraries like `Numpy`, `Pandas`, and `Funcy` etc. interrupts during execution. Third, Fig.2 shows some topic overlaps in some parts (e.g. Topic 9 and 10). The overlapping topics have a semantic relationship with each other. They share common keywords perhaps and as a result, they are clustered together.

More research can be carried out to discover esteemed topics from theses, dissertations, articles, and essays. Shodhganga is one of the most reliable sources of getting data of Indian thesis.

The data of Shodhganga can be utilized in the future to perform more research relating to this paper. Furthermore, in this paper, only topic coherence is shown in a general way but the coherence values of topics and perplexity scores are not deliberated. As a result, researchers may investigate these in order to have better outcomes using MALLET (McCallum, 2002). Next, we only have generated a word cloud to show the most frequent terms. But separate word clouds of each topic (see Table 4) could be more significant for data visualization.

7. Conclusion

In this paper, we have presented the topic modeling technique based on LDA to discover latent topics from the titles of the Indian LIS theses. We applied several approaches to understand the implicit text data and tried to show it in a meaningful form. We found keywords of 10 prominent topics, such as “library use”, “open-source”, “management”, “university library”, and “information-seeking behaviour”, etc. We also visualized Topic 1 with relevant keywords (30.3% of tokens) and overall frequency of the top-30 most salient terms. To to infer a certain topic, we tried to show the most representative thesis titles with keywords. The frequency of the top-10 most discussed topic keywors in a particular text was measured. Furthermore, the frequency of n-items was determined to comprehend their appearances in the titles. We also discussed some limitations regarding sample size, updation of python packages, and betterment of the model. This paper can be expanded for future research to include more improvements and applications.

Appendix 1: Sources of the thesis titles (Table 5)

s/n	Researcher	University	Year	Unique identifier
1	Jahan, Tabassum	Aligarh Muslim University	2009	http://hdl.handle.net/10603/49852
2	Agrawel, Pawan R	Maharaja Sayajirao University of Baroda	2014	http://hdl.handle.net/10603/87285
3	Turamari, R. Nagappa	Madurai Kamraj University	2014	http://hdl.handle.net/10603/136666
4	Nagaraja, S.	Bharathidasan University	2017	http://hdl.handle.net/10603/220323
5	Imchen, A Takatemsu	Gauhati University	2004	http://hdl.handle.net/10603/68296

References

Barde, B. V., & Bainwad, A. M. (2017). An overview of topic modeling methods and tools. *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 745–750. https://doi.org/10.1109/ICCONS.2017.8250563

Bernard, H. R., & Ryan, G. W. (1998). *Text Analysis: Qualitative and Quantitative Methods*. https://www.rand.org/pubs/external_publications/EP19980030.html

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media. https://books.google.co.in/books?id=KGIbfiiP1i4C

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://ai.stanford.edu/~ang/papers/jair03-lda.pdf

Buenaño-Fernandez, D., González, M., Gil, D., & Luján-Mora, S. (2020). Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach. *IEEE Access*, *8*, 35318–35330. https://doi.org/10.1109/ACCESS.2020.2974983

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization Techniques for Assessing Textual Topic Models. *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 74–77. https://doi.org/10.1145/2254556.2254572

Gan, G., Li, B., Li, X., & Wang, S. (2018). *Advanced Data Mining and Applications: 14th International Conference, ADMA 2018, Nanjing, China, November 16–18, 2018, Proceedings*. Springer International Publishing. https://books.google.co.in/books?id=pI2wvQEACAAJ

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Goswami, S., Mazumder, S., & Chakrabarty, S. (2021). Text mining of biomedical literature: Discovering new knowledge. *Library Philosophy and Practice (e-Journal)*, 31.

Han, X. (2020). Evolution of research topics in LIS between 1996 and 2019: An analysis based on latent Dirichlet allocation topic model. *Scientometrics*, *125*(3), 2561–2595. https://doi.org/10.1007/s11192-020-03721-0

Hong, L., & Davison, B. D. (2010). Empirical Study of Topic Modeling in Twitter. *Proceedings of the First Workshop on Social Media Analytics*, 80–88. https://doi.org/10.1145/1964858.1964870

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*.

Huth, E. J. (1989). The information explosion. *Bulletin of the New York Academy of Medicine*, *65*(6), 647–672. PubMed. https://pubmed.ncbi.nlm.nih.gov/2590751

Ifijeh, G. (2010). Information Explosion and University Libraries: Current Trends and Strategies for Intervention. *Chinese Librarianship: An International Electronic Journal*. http://eprints.covenantuniversity.edu.ng/5824/#.X_rpaegzZEY

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Major, C. H., & Savin-Baden, M. (2012). *An Introduction to Qualitative Research Synthesis: Managing the Information Explosion in Social Science Research*. Taylor & Francis. https://books.google.co.in/books?id=hXO9ZdzuV30C

McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu/about.php

Miller, A. (2018). Text Mining Digital Humanities Projects: Assessing Content Analysis Capabilities of Voyant Tools. *Journal of Web Librarianship*, *12*(3), 169–197. https://doi.org/10.1080/19322909.2018.1479673

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88–102. https://doi.org/10.1177/0165551515617393

Perkins, J. (2011). *Python Text Processing with Nltk 2.0 Cookbook: Lite*. Packt Publishing. https://books.google.co.in/books?id=XjXXnWPkd-AC

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 40–50.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. https://doi.org/10.3115/v1/W14-3110

Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing. https://books.google.co.in/books?id=48RiDwAAQBAJ

Sun, L., & Yin, Y. (2017). Discovering themes and trends in transportation research using topic modeling. *Transportation Research Part C: Emerging Technologies*, *77*, 49–66. https://doi.org/10.1016/j.trc.2017.01.013

Tong, Z., & Zhang, H. (2016). A Text Mining Research Based on LDA Topic Modelling. In *Computer Science & Information Technology* (Vol. 6, p. 210). https://doi.org/10.5121/csit.2016.60616

Villars, R. L., Olofson, C. W., & Eastwood, M. (2011). Big data: What it is and why you should care. *White Paper*, *14*, 1–14.

Wang, C., & Blei, D. M. (2011). Collaborative Topic Modeling for Recommending Scientific Articles. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining, 448–456. https://doi.org/10.1145/2020408.2020480

Yang, T.-I., Torget, A., & Mihalcea, R. (2011). Topic Modeling on Historical Newspapers. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 96–104. https://www.aclweb.org/anthology/W11-1513