8

N-grams Analysis of Digital Humanities Research during 2017-2021: A Study based on Scopus Database

Sourav Mazumder
Research Scholar, Department of Library and Information Science,
University of North Bengal, West Bengal
E-mail: smazumderlis91@gmail.com

Tapan Barui
Asst. Professor, Department of Library and Information Science,
University of North Bengal, West Bengal
E-mail: tapanbarui@nbu.ac.in

Abstract

In the social sciences, Digital Humanities (DH) is gaining traction. To determine the contexts or topics of DH research, researchers used science mapping and text mining approaches. In the present study, we have applied n-grams analysis to understand the context of the DH research from the abstract of 1348 articles (2017-2021). The data was collected from the Scopus database. We used Orage for n-grams extraction. Further, we visualised the n-grams using the word cloud. We identified top-10 unigrams, bigrams, and trigrams and constructed the research contexts with human judgement using the frequencies of the n-grams. From the results, we have observed some major research contexts like DH research, the use of digital technologies, ICT, social networks, cultural heritage, DH projects, and natural language processing. Bigrams were identified as more significant. This study can be helpful for scholars to understand the current research context and usage of terms.

Keywords: Digital humanities; Text mining; Social computing; *N*-grams model; Bibliographic data, Scopus.

1. Introduction

Nothing in the twenty-first century can be imagined without the use of technology. We refer to it as "digital technology" or "Information Communication Technology" (ICT). Our society has evolved into a kind of information ecosystem related to the digital environment as information continues to proliferate (Floridi, 2007). ICT plays a pivotal role in enhancing society, economy, culture, and education. Moreover, being digital is essential for extensive scientific and industrial progress. The relevance of technical assistance has been seen in research and development. Certainly, advanced digital technologies are used to facilitate scientific activities (Berry, 2012). The application of technologies is applied in every discipline, such as Science, Technology, Arts, and Humanities. However, the application of ICT in the humanities is not new. It has emerged as an interdisciplinary prospect and one of the scholarly domains' rising fields (Svensson, 2010). What exactly does "digital humanities (DH)" imply? "Digital humanities" or "humanities computing," according to Berry (2019), is computer-based technology in the humanities. There has been a surge in interest in the DH in recent years. Researchers explore knowledge from various disciplines such as language, literature, history, media science, computer science, and information science (Berry, 2019). We can find many resources (Berry, 2012; Gold, 2012; Warwick et al., 2012) for understanding DH. On the other hand, a considerable amount of literature is published on DH for analysing its evolution, nature, intellectual structure, topics, and research productivity by using science mapping (Münster, 2019; Su & Zhang, 2021; Wang et al., 2020).

There are substantial DH projects in English, History, Performing Arts, and Crowdsourcing that deal with archives, databases, text mining, visualisation, and crowdsourcing (Lehigh University, 2022). Researchers in the social sciences are using cutting-edge methodologies like data science to analyse and extract insights from large amounts of data (B. Wright, 2019). Artificial intelligence, big data, data mining, text mining, data visualisation, data management and curation, modelling, and data science are all included in data science. In DH, text mining is a popular technique. Analytics, clustering, topic modelling, sentiment analysis, and *n*-grams analysis are all related to it (J. Han et al., 2012; Mazumder & Barui, 2021). For text analysis in DH, we can use a variety of tools from the Digital Research Tools (DiRT) Directory (e.g., Voyant, Google Ngram Viewer, WordHoard) (Lehigh University, 2022).

Science mapping has already been identified as one of the most prominent methods for identifying essential concepts and contexts in published literature. Furthermore, using topic modelling, researchers have uncovered important research topics from scholarly communication text data (e.g., X. Han, 2020). We can see how topic modelling and scientific mapping are important in understanding the research context. *N*-grams are considered powerful aspects for representing text in sequential order (Welbers et al., 2017). Wang et al., (2007) identified four topics along with words and phrases from the dataset of Neural Information Processing Systems (NIPS) Conferences (1987-1999) using the Topical *N*-grams (TNG) model. Another study was based on an *n*-grams analysis of 3,367 papers from the journal Communications of the ACM. It was revealed that *n*-grams are useful for identifying information systems (Soper & Turel, 2012). Bouras and Tsogkas (2013) used *n*-grams to cluster news articles collected from several news portals. Bharadwaj and Shao (2019) employed *n*-grams to measure the correlation between TF (Term Frequency) and IDF (Inverse Term Frequency) in locating fake news. Wyskwarski (2020) mined job offerings data from five websites to identify the responsibilities of a business analyst using *n*-grams. These previous works provided insights into how *n*-grams can be utilised to extract the frequently used *n* numbers of terms in text data. However, no previous approach has used *n*-grams to determine context from abstracts of scholarly articles on DH.

The present study focuses on *n*-grams to identify the major terms or phrases that occurred in the abstracts of DH research during the period 2017-2021. The main objectives of the study are :

(i) to find out the top-10 *n*-grams that appeared in the text of abstracts;

(ii) to construct the context from the *n*-grams; and

(iii) to compare top-5 *n*-grams based on five years(2017-2021).

2. Materials and methodology

2.1 Data collection

In this study, there were several phases of data collection. First, a survey was conducted to determine the trends in published DH literature between 2010 and 2021. For this, the Scopus database was used to search the literature (search terms: “digital humanities,” “social comput*,” and “social science comput*”). A total of 12798 (Figure-1) publications were found. The publications comprise articles, chapters, conference proceedings, reviews etc. We can see a clear spike in 2020. However, the present study is delimited to a few criteria (Table 1). Second, the raw bibliographic data (includes title, authors, year, source title, and abstract) consisting

of 1393 articles (2017-2021) were extracted from the same database. Figure-2 shows the top-5 source titles (journals). Most of the publications can be found in "Digital Humanities Quarterly" (47).

2.2 Data preprocessing

We used Google Sheets for organising raw text data and Orange data mining software (Demšar et al., 2013) for text preprocessing of abstracts and generating *n*-grams. Some preprocessing activities were executed, including converting text to lowercase, tokenizing, removing English stopwords, and selecting the *n*-grams range. To make the dataset more efficient, a custom 706 stopwords (years and irrelevant words) was created using Notepad. We also eliminated 45 documents (having no abstracts) from the collection. We analysed the dataset containing the abstracts of 1348 documents. Moreover, we prepared subsets of the main datasets (year-wise) for further analysis.

2.3 *N*-grams

An n-gram is a contiguous sequence of *n* words or tokens in a text document in computational linguistics and probability. It is a probabilistic language model based on the Markov model (Jurafsky & Martin, 2021; Wikipedia contributors, 2022). *N*-grams can be classified into three major categories depending on the unit that incorporates them. Unigrams (1-gram) depict a single word (for example, "library" or "science" or "research" or "history"); Bigrams (2-grams) depict a sequence of two words (for example, "library science" or "science research"); and Trigrams (3-grams) depict three words (for example, "library science research" or "science research history"). The use of *n*-grams can be diverse in text analysis. For instance, it could be used to "detect spelling errors", "query expansion," "match strings," and "cluster text" (Robertson & Willett, 1998). In this study, we applied the *n*-grams (unigrams, bigrams, and trigrams) features using the data mining software. The frequency of the *n*-grams has been shown. By that, we tried to determine the context of the DH research.

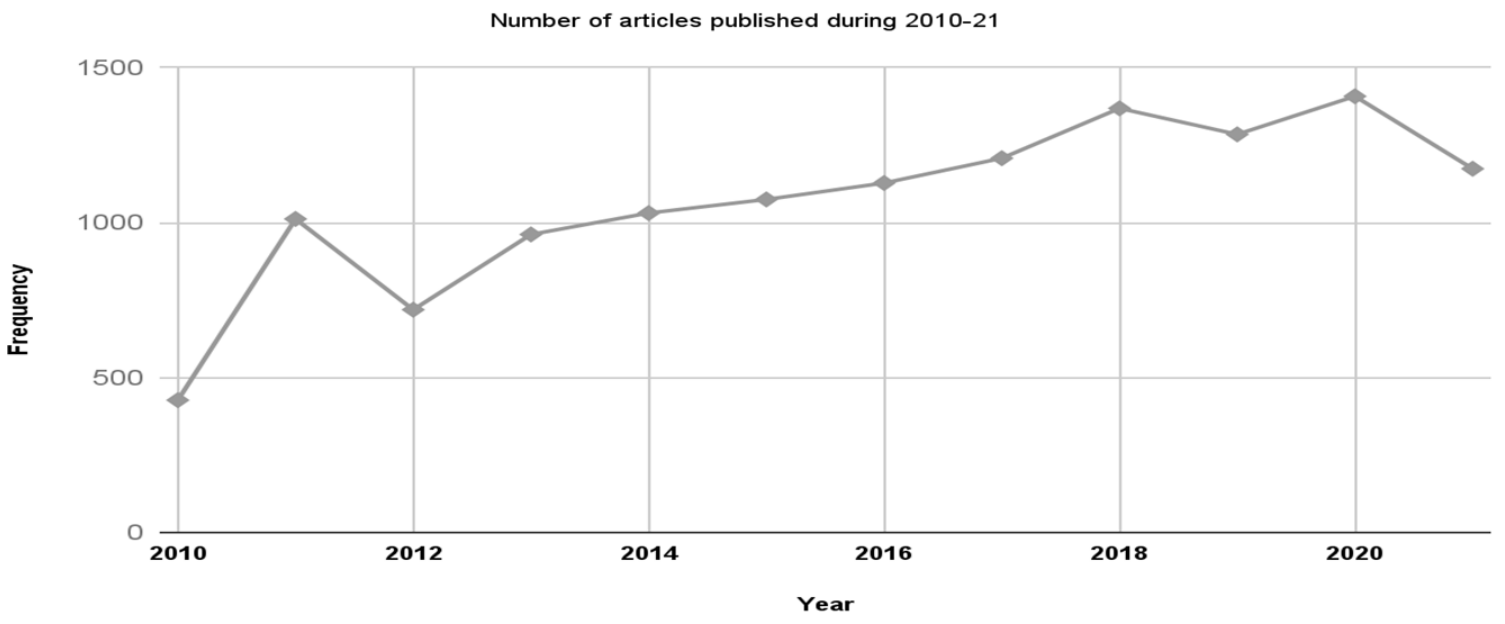

Figure-1: Trends of digital humanities research between the years 2010 and 2021

Table 1: Description of the search query on the Scopus for the present study

Characteristics	**Description**
Publication year	2017-2021
Document type	Articles
Subject area	Social science, Arts and Humanities

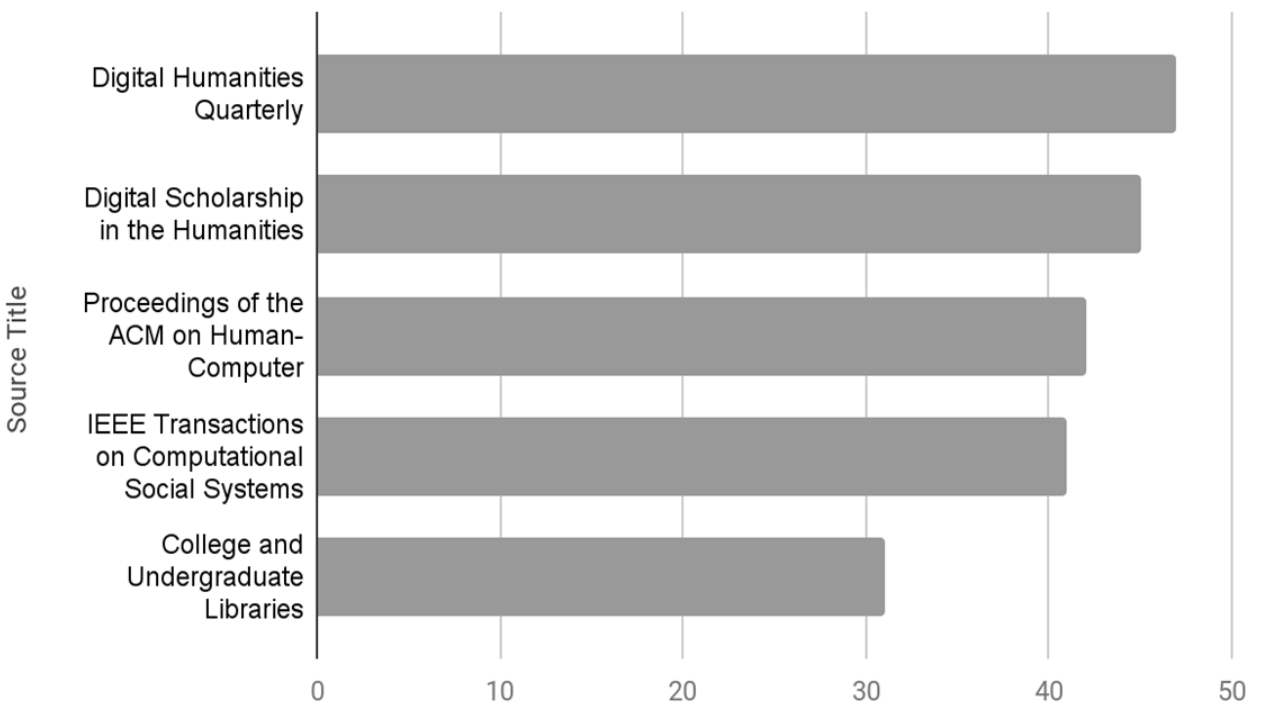

Figure-2: Top-5 journals for DH research during 2017-2021

3. Results and Discussion

After setting all the parameters for analysis (mentioned in section 2), we performed multiple actions in the software to generate *n*-grams. Additionally, we trained the subsets of the main dataset to present the list of unigrams, bigrams, and trigrams year-wise. The results have been presented in Table 2-3, and Figure 3-5.

3.1 Top-10 n-grams

Table 2 shows the top-10 *n*-grams found in the abstracts of DH scholarly articles published between 2010 and 2021. It is evident that unigrams dealt with only one word, bigrams with a sequence of two words, and trigrams with a string of three words at one time. The corpus had 134435 tokens and 14655 distinct words (single words or unigrams), 110773 bigrams, and 127930 trigrams after the stopwords were removed. The term "digital" appeared 2387 times in the abstracts, followed by "humanities" (1424), "research" (1366), "social" (1284), "information" (644), "history" (439), "network" (429), "technology" (423), "work" (421), and "knowledge" (403). Second, 110773 bigrams were retrieved in total. "Digital humanities" (1015) was the most popular bigram, followed by "social media" (191), "social networking" (182), "cultural heritage" (122), "social networks" (121), "humanities research" (93), "social sciences" (75), "social computing" (69), "digital technologies" (59), and "digital scholarship" (59). Third, we also took out a list of 127930 trigrams. The term "Digital humanities research" (69) was discovered as the highly occurring trigram. In addition, other trigrams were "digital humanities

projects: (42), “field digital humanities” (34), “humanities social sciences” (25), “social media literature” (20), “natural language processing” (20), “within digital humanities” (19), “research digital humanities” (18), “social media platforms” (17) and “online social networks” (16).

Table 2 can now be used to deduce the DH’s context. The search phrases, on the other hand, were not removed because they could cause complications when contextualising the text. As expected, the most often occurring unigrams are “digital” and “humanities.” The majority of the research, we might assume, focused on digital technology, social networks, and social factors. It does not, however, provide a clear depiction. We can much more discern the context if we look at the bigrams list. The word “digital humanities” came up frequently, as one might expect. The table reveals several themes, including social media, cultural heritage, digital scholarship, and digital technologies. Researchers approached scholarly work on social media or social networks to understand its usage and effects. Research (e.g., Brumann, 2015) related to cultural heritage consists of history, cultural sites, and social practices on conservation. Therefore, digital scholarship emerged along with other research contexts. Generally, it is based on a digital lifecycle that embraces some activities like the use of digital technologies, data management and curation, and data preservation (Zhou, 2021). Galleries, libraries, archives, and museums (GLAM) also contribute to the development of scholarly resources, which benefits researchers (Hilburger et al., 2021). So, in this context, we can presume that digital scholarship related studies are well associated with library and information science.

We also can acquire insights from trigrams as well. Firstly, we know about the DH research but some studies (e.g., Arana-Catania et al., 2021) were project-centric. For example, citizens’ engagement in the decision-making process. Second, it was discussed earlier that social media has been widely used for research purposes. So, contexts like social media literacy and social media platforms occurred many times in the abstracts. Third, natural language processing (NLP) was found as one of the most occurring trigrams. It is evidence that researchers were heavily interested in the application of NLP, machine learning, and artificial intelligence on DH.

However, some of the *n*-grams lacked sufficient context. For example, the *n*-grams model led to terms like “field digital humanities,” “humanities social sciences,” and “inside digital humanities.” Another aspect we found was that the two terms (social network or social networks) occurred differently. In the abstracts, the authors mentioned the two terms on different occasions. Furthermore, it was based on stemming algorithms (Lovins, 1968). We have also presented the world cloud of the *n*-grams (Figure 3-5)

Table 2: Frequency (*f*) distribution of top-10 unigrams, bigrams, and trigrams that assimilate 134435 tokens and 14655 unique words.

Sl.	Unigrams	*f*	Bigrams	*f*	Trigrams	*f*
1	digital	2387	digital humanities	1015	digital humanities research	69
2	humanities	1424	social media	191	digital humanities projects	42
3	research	1366	social network	182	field digital humanities	34
4	social	1284	cultural heritage	122	humanities social sciences	25
5	information	644	social networks	121	social media literature	20
6	history	439	humanities research	93	natural language processing	20
7	network	429	social sciences	75	within digital humanities	19
8	technology	423	social computing	69	research digital humanities	18
9	work	421	digital technologies	59	social media platforms	17
10	knowledge	403	digital scholarship	59	online social networks	16

Figure-3: World cloud of unigrams

Figure-4: Word cloud of bigrams

Figure-5: Word cloud of trigrams

3.2 Comparison of top-5 *n*-grams

In this section, we presented a year-wise (2017-2021) comparison of *n*-grams. Table 3 is quite revealing in several ways. First, unlike Table 2, it presents a timeline of mostly occurring words or phrases in the abstracts over the years. Second, it displays several significant contexts derived from bigrams and trigrams. Table 3 shows that the majority of the unigrams remained constant. Their frequency, however, varied

throughout the course of the five years. Bigrams, for the most part, remained the same, with a few exceptions. In most cases, digital humanities, social network, and social media ranked neck to neck. The bigrams' column showed no significant changes. Like unigrams, high and low frequencies of the bigrams also were noticed. What is striking about the data in the trigrams' frequency column is some uncommon trigrams. For example, in 2017, the second most popular trigram was "commodification rural space." We observed two different trigrams in 2018: "genealogy family history" and "geo social interaction." These phrases are associated with sociological (Pine, 2021) and geographical aspects in which humans are involved. In section 3.1, we already have discerned that NLP is one of the most popular themes for DH research. Here in this table, we also can see the remarks on NLP-related research. Authors sometimes used NLP and only "natural language." Nonetheless, it makes sense in the context. Another trigram "social value orientation" (e.g., Murphy et al., 2011) was identified in 2020. Generally, it dealt with the judgement and decision-making aptitude of human beings. These findings imply that there were no significant changes in the frequency of unigrams and bigrams during the period, except for their top-5 frequencies, which decreased over time but remained constant as top-5. Bigrams were no exception; they were similar to unigrams. However, few new contexts were detected as trigrams.

Table 3: Frequency distribution of Top-5 *n*-grams during the 2017-2021

Year	Unigrams	*f*	Bigrams	*f*	Trigrams	*f*
2017	digital	472	digital humanities	206	digital humanities research	8
	humanities	311	social network	29	commodification rural space	8
	research	202	social media	29	social network sites	8
	social	201	social networks	24	digital humanities projects	7
	information	130	digital scholarship	21	online social networks	6
2018	digital	334	digital humanities	144	field digital humanities	9
	research	229	social media	48	genealogy family history	7
	social	223	social networks	26	digital humanities research	7
	humanities	185	social network	22	digital humanities projects	6
	information	94	digital scholarship	16	geo social interaction	5

Year	Unigrams	*f*	Bigrams	*f*	Trigrams	*f*
2019	digital	465	digital humanities	213	digital humanities research	23
	humanities	297	social media	35	support digital humanities	10
	social	267	social network	35	natural language processing	8
	research	269	humanities research	26	research digital humanities	7
	information	147	social networks	25	field digital humanities	7
2020	digital	546	digital humanities	229	social media literature	19
	humanities	339	social media	65	digital humanities projects	16
	social	33	social network	46	social media platforms	11
	research	330	cultural heritage	29	digital humanities research	10
	information	152	social networks	20	social value orientation	7
2021	digital	570	digital humanities	223	digital humanities research	21
	research	336	cultural heritage	52	cultural heritage crowdsourcing	10
	humanities	292	social network	50	character social network	10
	social	257	humanities research	28	social network relationships	9
	information	121	social networks	26	natural language processing	8

4. Conclusion

To analyse *n*-grams, we used Scopus to obtain bibliographic data from DH research articles published between 2017 and 2021. A total of top-10 unigrams, bigrams, and trigrams were presented to comprehend the context of research works. The findings showed bigrams were more comprehensive than unigrams and trigrams. The *n*-grams revealed some important themes such as social media, cultural heritage, DH initiatives, digital scholarship, and NLP. A comparison of top-5 *n*-grams occurring during the period was also shown to understand the trend. It showed there were no

such big changes for unigrams and bigrams except for their frequencies. We observed a couple of new terms from the trigrams. This study may help researchers and DH practitioners to recognise the current approaches in DH research works. Though, this study is restricted to frequency analysis. In the near future, we will apply a predictive model to get a more informative context.

References

Arana-Catania, M., Lier, F.-A. V., Procter, R., Tkachenko, N., He, Y., Zubiaga, A., & Liakata, M. (2021). Citizen Participation and Machine Learning for a Better Democracy. *Digital Government: Research and Practice*, *2*(3), 1–22. https://doi.org/10.1145/3452118

Berry, D. M. (2012). Introduction: Understanding the Digital Humanities. In D. M. Berry (Ed.), *Understanding Digital Humanities* (pp. 1–20). Palgrave Macmillan UK. https://doi.org/10.1057/9780230371934_1

Berry, D. M. (2019). *What are the digital humanities?* [Blog]. The British Academy. https://www.thebritishacademy.ac.uk/blog/what-are-digital-humanities/

Bharadwaj, P., & Shao, Z. (2019). Fake News Detection with Semantic Features and Text Mining. *International Journal on Natural Language Computing (IJNLC)*, *08*(3), 17. https://doi.org/10.5121/ijnlc.2019.8302

Bouras, C., & Tsogkas, V. (2013). Enhancing News Articles Clustering using Word N-Grams. In *DATA* (pp. 53–60). http://telematics.upatras.gr/telematics/system/files/publications/1962DATA_2013_12_CR.pdf

Brumann, C. (2015). Cultural Heritage. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (pp. 414–419). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.12185-3

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data mining toolbox in python. *The Journal of Machine Learning Research*, *14*(1), 2349–2353.

Floridi, L. (2007). A Look into the Future Impact of ICT on Our Lives. *The Information Society*, *23*(1), 59–64. https://doi.org/10.1080/01972240601059094

Gold, M. K. (2012). *Debates in the Digital Humanities*. University of Minnesota Press. https://books.google.co.in/books?id=_6mo2tApzQQC

Han, J., Kamber, M., & Pei, J. (2012). Data Mining Trends and Research Frontiers. In J. Han, M. Kamber, & J. Pei (Eds.), *Data Mining (Third Edition)* (pp. 585–631). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00013-7

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

Hilburger, C., Langille, D., Nelson, M., Bordini, A., Greenhill, J. A., Dowson, R., & Goddard, L. (2021). Collaborating with GLAM Institutions. *Digital Studies / Le Champ Numérique*, *11*(Special Collection: Student Researchers within the DPN), Article Special Collection: Student Researchers within the DPN. https://doi.org/10.16995/dscn.377

Jurafsky, D., & Martin, J. H. (2021). N-gram Language Models. In *Speech and Language Processing* (pp. 1–29). https://web.stanford.edu/~jurafsky/slp3/3.pdf

Lehigh University. (2022). *Library Guides: Digital Humanities: Projects*. Lehigh University Libraries-Library Guides. https://libraryguides.lehigh.edu/ digitalhumanities/projects

Lovins, J. B. (1968). *Development of a Stemming Algorithm.* MASSACHUSETTS INST OF TECH CAMBRIDGE ELECTRONIC SYSTEMS LAB. https://apps.dtic.mil/sti/citations/AD0735504

Mazumder, S., & Barui, T. (2021). Discovering Topics from the Titles of the Indian LIS Theses. *Library Philosophy and Practice (e-Journal)*. https://digitalcommons.unl.edu/libphilprac/5924

Münster, S. (2019). Digital Heritage as a Scholarly Field-Topics, Researchers, and Perspectives from a Bibliometric Point of View. *Journal on Computing and Cultural Heritage*, *12*(3), 22:1-22:27. https://doi.org/10.1145/3310012

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*(8), 771–781.

Pine, L. G. (2021). *Genealogy*. Encyclopedia Britannica. https://www.britannica.com/topic/genealogy

Robertson, A. M., & Willett, P. (1998). Applications of n-grams in textual information systems. *Journal of Documentation*, *54*(1), 48–67. https://doi.org/10.1108/EUM0000000007161

Soper, D. S., & Turel, O. (2012). Who Are We? Mining Institutional Identities Using n-grams. *2012 45th Hawaii International Conference on System Sciences*, 1107–1116. https://doi.org/10.1109/HICSS.2012.642

Su, F., & Zhang, Y. (2021). Research output, intellectual structures and contributors of digital humanities research: A longitudinal analysis 2005–2020. *Journal of Documentation*, *78*(3), 673–695. https://doi.org/10.1108/JD-11-2020-0199

Svensson, P. (2010). The Landscape of Digital Humanities. *Digital Humanities*,

4(1). http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-37513

Wang, X., McCallum, A., & Wei, X. (2007). Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 697–702. https://doi.org/10.1109/ICDM.2007.86

Wang, X., Tan, X., & Li, H. (2020). The Evolution of Digital Humanities in China. *Library Trends*, *69*(1), 7–29. https://doi.org/10.1353/lib.2020.0029

Warwick, C., Terras, M., & Nyhan, J. (2012). *Digital Humanities in Practice*. Facet Publishing. https://books.google.co.in/books?id=hPhnDQAAQBAJ

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text Analysis in R. *Communication Methods and Measures*, *11*(4), 245–265. https://doi.org/10.1080/19312458.2017.1387238

Wikipedia contributors. (2022). N-gram. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=N-gram&oldid=1073019765

Wright, B. (2019). *Integrating Social Science & Data Science* [Video]. https://doi.org/10.4135/9781526491503

Wyskwarski, M. (2020). An attempt to determine the scope of duties of the business analyst – application of text mining analysis. *Zeszyty Naukowe. Organizacja i Zarządzanie / Politechnika Śląska*, *z. 148*. https://doi.org/10.29119/1641-3466.2020.148.59

Zhou, P. X. (2021). Towards a Sustainable Infrastructure for the Preservation of Cultural Heritage and Digital Scholarship. *Data and Information Management*, *5*(2), 253–261. https://doi.org/10.2478/dim-2020-0052