Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2023.00089

Research Article

Can we identify the similarity of courses in computer science?

Tugay KARADAĞ[1,*], Coşkun PARIM[1], Ali Hakan BÜYÜKLÜ[1]

[1]*Department of Statistics, Yildiz Technical University, Istanbul, 34349, Türkiye*

ARTICLE INFO

Article history
Received: 13 July 2021
Revised: 28 September 2021
Accepted: 18 October 2021

Keywords:
Computer Science; Curriculum; Data Processing and Interpretation; Higher Education; Science, Knowledge; Technology

ABSTRACT

Especially on the Internet, popular topics in computer sciences which are artificial intelligence, big data, business analytics, data mining, data science, deep learning, and machine learning have been compared or classified using confusing Venn diagrams without any scientific proof. Relationships among the topics have been visualized in this study with the help of Venn diagrams to add scientificity to visualizations. Therefore, this study aims to determine the interactions among the seven popular topics in computer sciences. Five books for each topic (35 books) were included in the analysis. To illustrate the interactions among these topics, the Latent Dirichlet Allocation (LDA) analysis, a topic modeling analysis method, was applied. Further, the pairwise correlation was applied to determine the relationships among the chosen topics. The LDA analysis produced expected results in differentiating the topics, and pairwise correlation results revealed that all the topics are related to each other and that it is challenging to differentiate between them.

Cite this article as: Karadağ T, Parım C, Büyüklü AH. Can we identify the similarity of courses in computer science? Sigma J Eng Nat Sci 2023;41(4):812–823.

INTRODUCTION

Vast and complex information abounds in computer science topics including artificial intelligence, big data, business analytics, data mining, data science, deep learning, and machine learning. Especially on the Internet, these topics are compared or classified by many people without any scientific basis. One general approach is the creation of a Venn diagram about the topics to show which categories are subsets or supersets. How do these topics interact with each other? Is there a topic that encompasses all others? Are there any unrelated topics?

According to Kim [1], for example, deep learning is a machine learning technique that employs deep neural networks; this is a general opinion that is held by data scientists and researchers. However, we could not assert that deep learning is fully linked to machine learning; there may be slight differences between the two approaches.

During the last two decades, data in different fields have developed substantially. The total volume of generated and repeated data in the world is 175 zettabytes (175,000,000,000,000 gigabytes), as stated in a paper by the International Data Corporation (IDC) in 2018. The term "Big Data" was first used in the mid-1990s [2] and currently, the term is used to specify large data sets owing to the vast expansion of global data. In proportion to classic data sets, big data includes unstructured data stacks that

*Corresponding author.
*E-mail address: karadagt@yildiz.edu.tr

This paper was recommended for publication in revised form by Adem Kilicman

mostly call for more real-time analysis. Likewise, big data helps us to deeply understand hidden values; creates new opportunities to explore new values; and presents new challenges, such as how to organize and handle large data sets.

Size is the fundamental characteristic that comes to mind when asked, "What is big data?" However, other characteristics of big data have recently appeared. For example, the study of Laney [3] suggests that the challenges of data management have three dimensions (the Three Vs): Volume, Variety, and Velocity. The Three Vs have emerged as a universal structure for identifying big data [4, 5]. According to Gandomi and Haider [6], the Three Vs are defined as follows:

Volume is the magnitude of the data;

Variety is the structural heterogeneity in the dataset;

Velocity refers to the rate at which data are generated and then the rate at which they are analyzed.

Some industries that use big data include the medical industry, transportation industry, information technology industry, education, government, and banking and security.

Artificial intelligence is a branch of computer science that aims to create smart machines that behave as though they were intelligent [7]. It is the intelligence emerged by machines or programming [8]. Basically, it is the ability of any computer or computer-controlled robot to perform certain situations and behaviors in a similar way to that of intelligent living beings. Although they are far from real, the term artificial intelligence has also been applied to computer systems and programs that can do more complicated tasks than straightforward programming [9]. According to [10], artificial intelligence is the part of computer science that is related to the design of intelligent computer systems that show characteristics associated with intelligent human behavior such as reasoning, language learning, understanding, and solving problems [11, 12]. The primary application fields of artificial intelligence are automated customer support, personalized shopping experience, healthcare, finance, smart cars and drones, travel and navigation, social media, and smart home devices.

Another branch of computer science is business analytics, which refers to all methods and techniques used by an organization to evaluate performance. It is the combination of processes, technologies, and systems that are used by organizations to manage their work and business planning. It helps to develop new insights and understand business performance based on data and statistical methods. Business analytics solutions generally manage an organization's planning and measure its past performance. It may change the way of managing and making decisions in various organizations. It allows managers to better manage and make decisions using data and facts, rather than just making decisions based on experience, instinct, or intuition [13]. Some examples for the use of business analytics are improving productivity and collaboration, enhancing customer support, forecasting orders, and creating recipes at companies.

The process of meaningfully extracting potentially useful information from data in data stacks can be referred to as data mining. It is also called knowledge discovery in databases [14]. Some articles or documents use different names for this method, such as knowledge extraction and data archaeology, which contain minimal differences [15]. Data mining is the phase of storing large amounts of information, scanning for useful information, analyzing the information in some way, interpreting the results from an expert's point of view, and making predictions based on past data. Data are worthless unless processed. Batch data are processed, valued, and then converted into information. Data are divided into predefined classes based on their attributes, and data mining handles the recorded data to make sense of and convert it to information. It is currently a prominent study topic in the information and database technology field [16]. It is used by actors such as service providers, supermarkets and retail stores, and crime prevention agencies; and in fields such as science, engineering, education, and e-commerce.

Data science aims to answer questions asked about data by using names and numbers, or, in other words, labels and categories. It examines every subject related to data. It provides assistance in processing, managing, analyzing, and assimilating large amounts of fragmented, structured, or unstructured data [17]. It has become popular today with the adaptation of concepts such as machine learning, deep learning, and artificial intelligence in industry and technology. Owing to its high commercial utility, data science has become popular as a formal discipline [18]. Some well-established statistical and computational methods have been used by big companies to mine high volumes of financial and social data [19]. Commonly used applications of data science include image recognition, speech recognition, fraud, and risk detection, and airline route planning. According to Kim [1], deep learning is a machine learning procedure that uses deep neural networks. Learning can be supervised, semi-supervised, or unsupervised [20-22]. Deep learning is a versatile tool that emerged with the aim of assimilating a large heterogeneous data. For complex and uncertain situations, deep learning provides reliable prediction results [23]. It is a type of machine learning algorithm that uses various layers to gradually determine advanced-level characteristics from raw data. For instance, in image processing, a lower layer may describe edges and a higher layer may describe human-meaningful components like digits, letters, or faces [24]. According to [25, 26], deep learning structures such as recurrent neural networks, deep belief networks, and deep neural networks have been applied to fields, including audio recognition, bioinformatics, medical image analysis, and natural language processing, where they have produced results comparable to and, in some cases, superior to human experts.

Machine learning is a technique that makes inferences from existing data with the help of mathematical and statistical methods to determine the "unknown." It is the

scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions and relying on patterns and inferences instead [27]. A mathematical model based on sample data, which is called "training data," is created to make decisions or predictions without being explicitly programmed to perform a specific task by machine learning algorithms. Examples of machine learning include image recognition, speech recognition, medical diagnosis, document classification, and spam detection.

In recent years, these seven topics of computer science have been taught to students in universities. While some of these topics may have distinct contents, their contents and methods overlap. While designing a course or curriculum, university departments offer some of these topics as courses without considering their similarities. Therefore, when designing a course or curriculum, this study could serve as a guide to avoid content duplication in education. This study aims to determine the interactions and relationships among seven of the recently popular topics in the field of Statistics and Computer Science. There has been no in-depth study investigating the interactions of these courses in the literature, to the knowledge of the authors. Therefore, this research topic will contribute to the literature and will be a trigger for future studies.

Related Works

An intelligent approach was proposed by Bastani, et al. [28] based on Latent Dirichlet Allocation (LDA) analysis to analyze Consumer Financial Protection Bureau consumer complaints. Further, Roque, et al. [29] described how topic modeling can be effectively used to identify co-occurrence patterns of attributes related to run-off-road crashes. In their study, LDA was applied to analyze the topics mentioned in their study and divided into two groups: discovered problems and proposed solutions. Natural language processing was applied in De Clercq, et al. [30] study to identify innovative technology trends related to food-waste treatment, biogas, and anaerobic digestion. In their study, LDA and the perplexity methods were used to assign the main topics that comprised the patent corpora and to determine which technological concepts were most associated with each topic. Griffiths and Steyvers [31] defined the LDA and used a Markov chain Monte Carlo (MCMC) algorithm for inference in their models, which showed that their algorithm worked with a small dataset. They then applied the algorithm to a corpus consisting of abstracts from the National Academy of Sciences Proceedings from 1991 to 2001, identifying the number of issues needed to account for the information contained in this corpus and extracting a series of topics. They used these topics to describe the relationships among several scientific disciplines and to assess trends and "hot topics" by analyzing topic dynamics and used the assignments of words to topics to highlight the semantic content of papers. Chen and Ren [32] proposed Forum-LDA for modeling the generative process of root posts and relevant and irrelevant response posts jointly. Silge and Robinson [33] analyzed the chapters of four different books with the LDA topic modeling method. They found that the books on four different topics were separated according to the LDA method and that the words on each topic (together with the chapters) were assigned mostly to the same topic.

Data Collection

Due to the fact that they are the most popular blending topics of computer science and statistics, our study analyzes seven topics: data science, big data, business analytics, machine learning, artificial intelligence, deep learning, and data mining. We searched for books by topic on link.springer.com. We then collected related books according to the corresponding topic. We used link.springer.com, as our institution is affiliated with this platform, to obtain downloadable books (e.g., to get a book about deep learning, we wrote "deep learning" in the search text field and downloaded the most relevant and downloadable book). Since only one book per topic would not be satisfactory, and for some topics, it is not possible to get more than five downloadable books, we decided to download five books per topic. Furthermore, we decided that the number of books for each topic should be equal. We collected a total of 35 books for data collection. The topics, titles, authors, publication years, and page counts of the books can be seen in Table 1.

MATERIALS AND METHODS

This study applied a variety of text mining techniques. According to Gharehchopogh and Khalifelu [67], text mining is the method of analyzing text to extract information that is useful for particular goals. Text is mostly unstructured and complicated; nonetheless, it is the most common tool for the formal exchange of information. Nowadays, classical information extraction techniques are inadequate due to the increasing amount of text data. Generally, only a small portion of available data is suitable for a user. Without an understanding of the contents of documents, it is challenging to create productive queries to analyze and extract useful information from data. Users need tools to compare several documents, sort them by importance, and determine their suitability or find patterns and trends in multiple documents. Accordingly, text mining has become progressively popular and an essential topic in data mining [67-69]. This section explains the processes used in this study and provides the necessary information about pre-processing, the LDA process, visualization, and pairwise correlation methods.

Pre-Processing

In general, data collected from any source contains noise. Therefore, it needs to be processed in several steps before text mining is implemented [70]. Our pre-processing

Table 1. The title, authors, publishing years, and page counts of the books

Topic	Book No.	Page	Title of Book	Author(s) and Years
AI	1	484	Abstraction in Artificial Intelligence and Complex Systems	[34]
	2	283	Aspects of the Theory of Artificial Intelligence: The Proceedings of the First International Symposium on Biosimulation Locarno, June 29–July 5, 1960	[35]
	3	216	Intelligence and Artificial Intelligence: An Interdisciplinary Debate	[36]
	4	316	Introduction to Artificial Intelligence	[7]
	5	402	Philosophy and Theory of Artificial Intelligence	[37]
BD	6	183	Big Data and Analytics: Strategic and Organizational Impacts	[38]
	7	263	Big Data and Visual Analytics	[39]
	8	216	Big Data Bootcamp: What Managers Need to Know to Profit From the Big Data Revolution	[40]
	9	267	Big Data Management	[41]
	10	400	Big Data Technologies and Applications	[42]
BA	11	243	Advanced Business Analytics	[43]
	12	156	Advanced Business Analytics	[44]
	13	189	Business Analytics for Managers	[45]
	14	162	Business Analytics: A Practitioner's Guide	[46]
	15	310	R for Business Analytics	[47]
DM	16	734	Data Mining: The textbook	[48]
	17	374	Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery	[49]
	18	241	Journeys to Data Mining: Experiences from 15 Renowned Researchers	[50]
	19	440	Principles of data Mining	[51]
	20	397	Scientific Data Mining and Knowledge Discovery	[52]
DS	21	272	Intelligent Techniques for Data Science	[53]
	22	769	Data Science: Third International Conference of Pioneering Computer Scientists, Engineers and Educators	[54]
	23	251	Data Science: Create Teams That Ask the Right Questions and Deliver Real Value	[55]
	24	342	Data Science: Innovative Developments in Data Analysis and Clustering	[56]
	25	213	Mathematical Problems In Data Science: Theoretical and Practical Methods	[57]
DL	26	151	Matlab Deep Learning, in With Machine Learning, Neural Networks, and Artificial Intelligence	[1]
	27	329	Deep Learning in Natural Language Processing	[58]
	28	191	Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence	[59]
	29	497	Neural Networks and Deep Learning : A Textbook	[60]
	30	508	Deep Learning: A Practitioner's Approach	[61]
ML	31	172	Machine Learning: Modeling Data Locally and Globally	[62]
	32	240	Advances in Machine Learning and Data Analysis	[63]
	33	291	An Introduction to Machine Learning	[64]
	34	210	Machine Learning with R	[65]
	35	358	Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python	[66]

AI: Artificial Intelligence | BD: Big Data | BA: Business Analytics | DM: Data Mining | DS: Data Science | DL: Deep Learning | ML: Machine Learning

Table 2. Pre-processing example

Table 2a. The process of uploading PDFs to R Studio, first process

Historically, the first explicit theory of abstraction started at the axiomatic level. Plaisted [419] provided a foundation of theorem proving with abstraction, which he sees as a mapping from a set of clauses to another one that satisfies some properties related to the deduction mechanism. Plaisted introduced more than one abstraction, including a mapping between literals and a semantic mapping. A more detailed description of his work will be given in Chap. 4. Later, Tenenberg [526] pointed out some limitations in Plaisted's work, and defined abstraction at a syntactic level as a mapping between predicates, which preserves logical consistency.

Table 2b. Tokenization process

Historically the first explicit theory of abstraction started at the axiomatic level Plaisted [419] provided a foundation of theorem proving with abstraction which he sees as a mapping from a set of clauses to another one that satisfies some properties related to the deduction mechanism Plaisted introduced more than one abstraction including a mapping between literals and a semantic mapping A more detailed description of his work will be given in Chap 4 Later on Tenenberg [526] pointed out some limitations in Plaisted's work and defined abstraction at a syntactic level as a mapping between predicates which preserves logical consistency

Table 2c. Transforming the data in a one-word-per-row format

Line	Word	Line	Word	Line	Word	Line	Word	Line	Word	Line	Word
1	Historically	18	of	35	another	52	including	69	be	86	defined
2	the	19	theorem	36	one	53	a	70	given	87	abstraction
3	first	20	proving	37	that	54	mapping	71	in	88	at
4	explicit	21	with	38	satisfies	55	between	72	Chap	89	a
5	theory	22	abstraction	39	some	56	literals	73	4	90	syntactic
6	of	23	which	40	properties	57	and	74	Later	91	level
7	abstraction	24	he	41	related	58	a	75	on	92	as
8	started	25	sees	42	to	59	semantic	76	Tenenberg	93	a
9	at	26	as	43	the	60	mapping	77	[526]	94	mapping
10	the	27	a	44	deduction	61	A	78	pointed	95	between
11	axiomatic	28	mapping	45	mechanism	62	more	79	out	96	predicates
12	level	29	from	46	Plaisted	63	detailed	80	some	97	which
13	Plaisted	30	a	47	introduced	64	description	81	limitations	98	preserves
14	[419]	31	set	48	more	65	of	82	in	99	logical
15	provided	32	of	49	than	66	his	83	Plaisted's	100	consistency
16	a	33	clauses	50	one	67	work	84	work		
17	foundation	34	to	51	abstraction	68	will	85	and		

Table 2d. Removal of stop words, numbers, and punctuation

Line	Word	Line	Word	Line	Word	Line	Word	Line	Word	Line	Word
1	historically	9	provide	17	satisfy	24	abstraction	31	description	38	syntactic
2	explicit	10	foundation	18	property	25	including	32	chap	39	level
3	theory	11	theorem	19	related	26	mapping	33	tenenberg	40	mapping
4	abstraction	12	proving	20	deduction	27	literals	34	limitations	41	predicates
5	started	13	abstraction	21	mechanism	28	semantic	35	plaisted	42	preserve
6	axiomatic	14	mapping	22	plaisted	29	mapping	36	defined	43	logical
7	level	15	set	23	introduced	30	detailed	37	abstraction	44	consistency
8	plaisted	16	clauses								

Table 2e. Word frequencies of the example paragraph

Word	Freq*	Word	Freq*	Word	Freq*	Word	Freq*	Word	Freq*	Word	Freq*
4	abstraction	1	clauses	1	explicit	1	literals	1	historically	1	1
4	mapping	1	consistency	1	foundation	1	logical	1	including	1	1
2	level	1	deduction	1	historically	1	description	1	introduced	1	1
2	plaisted	1	defined	1	including	1	detailed	1	limitations	1	1
1	axiomatic	1	description	1	introduced	1	explicit	1	literals	1	1
1	chap	1	detailed	1	limitations	1	foundation	1	logical	1	1

* *Frequencies*

method for obtaining word frequencies of each book is described below.

1. After collecting downloadable books from link.springer.com in PDF format, books are uploaded to R (Please see an example of a paragraph in Table 2 [34]). When uploading a PDF to R, the program recognizes the file as shown in Table 2a.
2. After uploading a PDF to the program, the next step is the tokenization step (Table 2b).
3. In this step, data is transformed into a one-word-per-row format (Table 2c).
4. The stop words, numbers, punctuation, and unwanted letters are then removed (Table 2d).
5. The word frequencies of the paragraph (Table 2a) are retrieved. In this table, word frequencies can be seen in a large-to-small order (Table 2e).

Latent Dirichlet Allocation (LDA)

LDA was introduced by Blei, et al. [71] as a generative probabilistic model of a corpus for a text document. It is often used for topic modeling. According to this method, each document consists of a mixture of latent topics. In addition, each topic is characterized by its distribution over unique words and the relative importance of the topics that differ from one document to another. This can be used to find complex words related to each topic and to determine the topic of the document by using the words found in each document [28]. The generative process underlying LDA is shown in Figure 1. The boxes in this figure represent plate notations used to show the replications, and plates K, N, and D indicate the number of topics, the total number of unique words within documents, and the number of documents, respectively. The topic of interest in the LDA analysis is parameterized with distribution per document (θd), distribution over words (β (z)), and per-word topic assignment (zi). θd and β(z) can only be calculated using topic-index assignments zi [72]. The arrows indicate conditional dependencies between the variables. Descriptions of the notations mentioned in this process can be found in Table 3.

A document d in Labeled LDA, the sampling probability for a topic for position i in is given as [73]:

Table 3. Notation descriptions for the LDA process

Notation	Description
θ	A vector of topic proportions for the document
β	Conditional probability table of the words to topics
α	Hyperparameters for prior distribution of θ
η	Hyperparameters for prior distribution of β
z	A vector of topics corresponding to each word in the document (assignments)
w	A vector of words per document
N	Total number of unique words within a document
D	Number of documents
K	Number of topics

$$P(z_i = j \mid z_{-i}) \propto \frac{n_{-i,j}^{w_i} + \eta_{w_i}}{n_{-i,j}^{(\cdot)} + \eta^T 1} \times \frac{n_{-i,j}^{(d)} + \alpha_j}{n_{-i,\cdot}^{(d)} + \alpha^T 1} \quad (1)$$

The β multinomial topics are learned from the training set. Afterward, the sampling in formula 1, limited to tags, is used to determine per-word tag assignments z. This allows inferences to be made on any new labeled test document. Therefore, the posterior distribution θ over the topics can be calculated by appropriately normalizing the topic assignments z [73].

Visualization

After putting the words from every book into a one-word-per-row format, the data are ready for visualization. There are five books for each topic, for a total of 35. To determine the most common words used in the books, word frequencies are obtained for every topic. Word clouds are then drawn for each topic. The logic of the word clouds is quite straightforward; the more frequently a word is repeated in a text, the larger and thicker the word appears in the word cloud. For pre-processing and visualization, R and the "tidytext" library were used for analyses (https://www.tidytextmining.com).

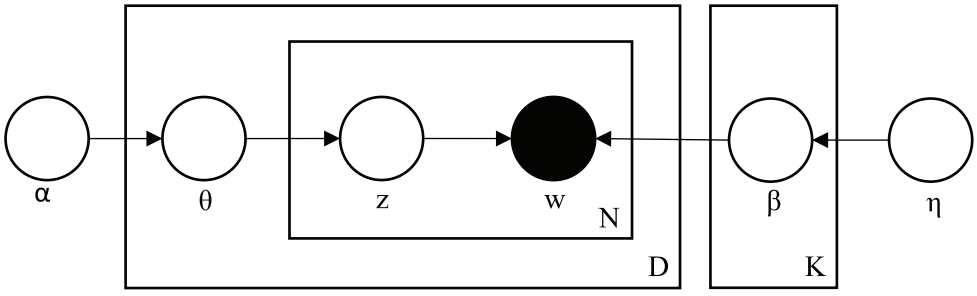


Figure 1. The Latent Dirichlet Allocation (LDA) process for topic modeling.

Pairwise Correlation

How often words appear separately or together can be a subject of research. In that case, it may be desirable to investigate their correlations. The phi coefficient is commonly used for binary correlation to find the correlation between words. The phi coefficient deals with whether the words A and B are together (Table 4). For example, n_{11} shows the number of documents in which both A and B appear; n_{00} represents the number of documents in which they do not appear; and n_{10} and n_{01} represent the absence of either A or B. The phi (ϕ) coefficient is:

$$\varphi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1.}n_{0.}n_{.0}n_{.1}}} \quad (2)$$

Table 4. Values used to calculate the phi coefficient

	Has word B	No word B	Total
Has word A	η_{11}	η_{10}	$\eta_{1.}$
No word A	η_{01}	η_{00}	$\eta_{0.}$
Total	$\eta_{.1}$	$\eta_{.0}$	η

The phi coefficient is equivalent to the Pearson correlation when applied to binary data [74].

In that case, correlation coefficients were found among the topics. A Venn diagram was drawn for the seven topics according to the correlation analysis. To create a Venn diagram, we used a graphing/charting and general data visualization application (www.meta-chart.com). When someone enters the values of each relation, the program returns a Venn diagram about the given topics.

RESULTS AND DISCUSSION

We performed the data cleansing process and the LDA analysis using the R software. For the LDA analysis, we used "topicmodels" integrated into the R package tm for text mining applications. This package provides an interface for estimating the LDA topic model with Gibbs sampling and variational expectation-maximization algorithm [75]. Since η and α parameters are generally defined as 0.1 in the literature, they were 0.1 by default in our study, and k = 7, because there are seven different types of books. The aim was to determine whether the books are differentiated by topic. As mentioned in the related works section, many studies conducted for topic modeling attempt to determine

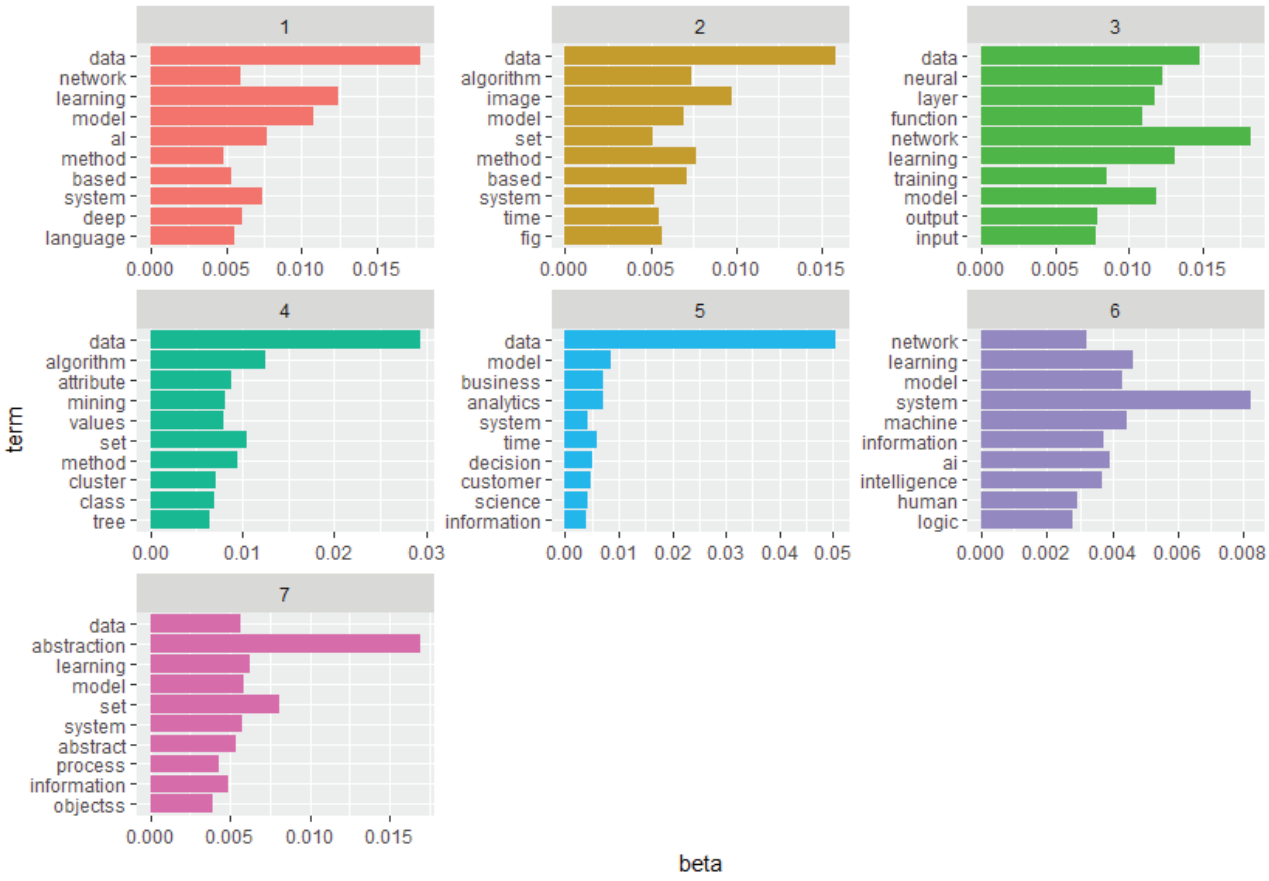


Figure 2. Top ten words with the most frequency for each book using LDA.

the number of topics. However, in our study, since the number of topics is seven, we used the classification feature of the LDA topic modeling to prove that the topics could not be decomposed into each other. The top ten words calculated from β values and contributing to all the seven topics are given in Figure 2. For example, the word "data" is among the top 10 words for all topics except the sixth topic. Similarly, other words occur in more than one topic. However, their extent of contribution is different.

It would be possible to estimate to which book each graphic calculated from β values corresponds if the topics and their contents were distinct. However, since words such as "data," "method," "algorithm," and "function" could be mentioned in all the topics, it is difficult to find out to which topic each graphic relates. This is also consistent with the aim of our study, because, from these graphics, it is easy to conclude that these seven topics share common characteristics. One can get an idea of which word represents which topic with some words (even if not with all the words). For example, with the word "business" in the fifth topic, as shown in Figure 2, one can consider that the word is under the topic "business analytics." Here, we have to say that each book is used as a document. Figure 3 reveals the topic to which the words in each topic are assigned with the help of LDA. Each line in this graph shows the right topics for the words, and each column shows to which topic the words are assigned. About 48% of the words in the "artificial intelligence" topic are assigned to the "artificial intelligence" topic. The words were most correctly assigned in the case of this topic. The degree of appropriate assignments of artificial intelligence, machine learning, deep learning, data science, business analytics, big data, and data mining was found to be approximately 48%, 41%, 40%, 35%, 21%, 20%, and 19%, respectively. These results reveal the high similarity among the topics.

For instance, 57% of the words in the deep learning topic were assigned to the machine learning topic. This rate is higher than the words in the deep learning topic being assigned to the deep learning topic. These results demonstrate that deep learning and machine learning are very similar. Another detail that draws attention is the machine learning topic. The words in the five topics (artificial intelligence, business analytics, data mining, deep learning, machine learning) were assigned to the machine learning topic the most. Machine learning was found to be the second most assigned topic only in big data and data science topics. Data science was found to be the topic with the most assignment percentage for "big data" and "data science." Overviewing all the topics, the correct assignment rate was calculated as approximately 30%, a meager rate. It is difficult to separate the books because they contain similar topics. The correlation matrix for each topic, which was created by using the "corrplot" package in R [76], is given in Figure 4. The lowest correlation value among all topics is 0.4, between artificial intelligence and big data. In other words, even the least related topics are found to be 40% related. In addition, the correlation coefficient between data science and big data is approximately 0.86. The Venn diagram created using the correlation matrix can be examined in Figure 5. All other topics comprised the machine learning topic, which is consistent with the results of the previous stage. In other words, in the LDA analysis, the words in each topic were assigned to machine learning with the highest ratio, and all other topics were most similar to machine learning. As such, the results from the LDA analysis and the Venn diagram generated from the correlation matrix were consistent.

Word clouds are generally used in text mining for visualization. Therefore, word clouds for each topic and a word cloud for the total of 35 books were created; the eight-word clouds can be seen in the Appendix.

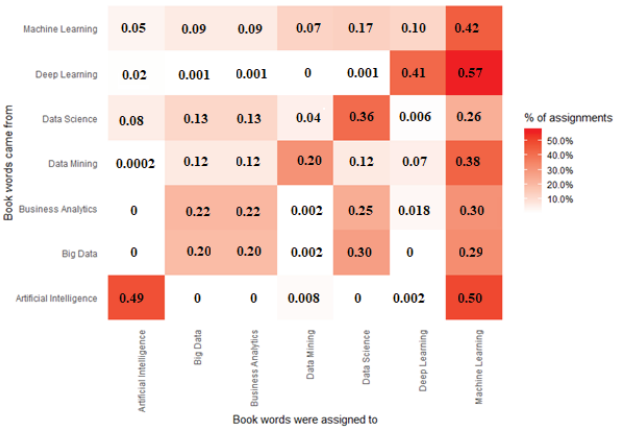


Figure 3. Accuracy rates of assignment for each topic using the LDA.

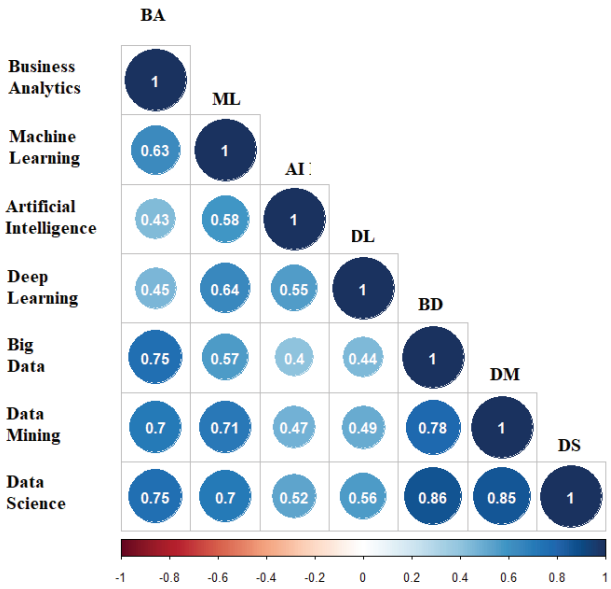


Figure 4. Correlation matrix results.

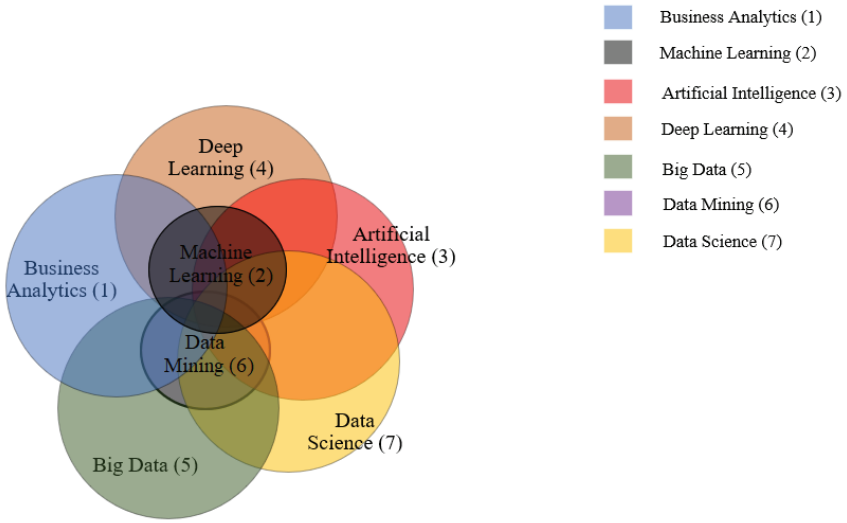


Figure 5. Venn diagram derived from the correlation matrix.

CONCLUSION

This study investigated the interactions among seven popular course topics in the field of computer science. Analyses results proved that the seven topics chosen are very similar to each other; it is challenging to differentiate them. Further, the machine learning topic comprised all other topics in the context of this study. Based on our results, these educational titles that are used in many scientific fields are intertwined; their correlative structures are similar, and they often contain the same words. Previous studies have shown the relationships among these topics through graphs without any statistical analysis or scientific proof. However, this study scientifically determines the common points of the seven popular scientific titles. Moreover, this study could serve as a guide in designing relevant courses to avoid duplication in education, especially in computer sciences, statistics, and other departments that offer courses related to these seven popular topics. University departments offer some of these topics as courses without considering the high level of similarity in their contents. Therefore, this study should be considered as the first step in minimizing such a duplication problem. By applying correlation analysis over the words in the content of the course books, the high correlations of the subjects with each other do not of course reveal the duplication in education problem by itself. However, the results obtained in this study suggest that there is a duplication in the education problem. In future studies, it is recommended to delve deeper into this subject and conduct a rigorous and comprehensive study.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

REFERENCES

[1] Kim P. Matlab Deep Learning: with Machine Learning, Neural Networks and Artificial Intelligence. New York: Apress; 2017.

[2] Elsahlamy E, Eshra A, Eshra N, El-Fishawy N. Empowering GIS with Big Data: A Review of Recent Advances. 2021 International Conference on Electronic Engineering (ICEEM), Menouf, Egypt, 2021, pp. 1–7. [CrossRef]

[3] Laney D. 3D data management: Controlling data volume, velocity and variety. META Group Research Note 2001;6:1.

[4] Chen H, Chiang RH, Storey VC. Business intelligence and analytics: From big data to big impact. MIS Quarterly 2012;36:1165–1188. [CrossRef]

[5] Kwon O, Lee N, Shin B. Data quality management, data usage experience and acquisition intention of big data analytics. Int J Inform Manag 2014;34:387–394. [CrossRef]

[6] Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. Int J Inform Manag 2015;35:137–144. [CrossRef]

[7] Ertel W. Introduction to Artificial Intelligence. New York: Springer; 2018.

[8] Swathi P. Real time applications of artificial intelligence-review. Int J Inf Res Rev 2021;2:99–101.

[9] Mellit A, Kalogirou SA. Artificial intelligence techniques for photovoltaic applications: A review. Prog Energy Combust Sci 2008;34:574–632. [CrossRef]

[10] Barr A, Feigenbaum EA, Cohen P. The Handbook of Artificial Intelligence, vols. 1-3, Los Altos, CA: William Kaufmann Inc; 1981.

[11] Kalogirou SA. Artificial intelligence for the modeling and control of combustion processes: a review, Prog Energy Combust Sci 2003;29:515–566. [CrossRef]

[12] Kalogirou S. Artificial intelligence in energy and renewable energy systems. New York: Nova Publishers; 2007.

[13] Piot-Lepetit I, Nzongang J. Business analytics for managing performance of microfinance Institutions: A flexible management of the implementation process. Sustainability 2021;13:4882. [CrossRef]

[14] Piateski G, Frawley W. Knowledge Discovery in Databases. Cambridge, Massachusetts: MIT Press; 1991.

[15] Chen M-S, Han J, Yu PS. Data mining: an overview from a database perspective. IEEE Trans Knowl Data Eng 1996;8:866–883. [CrossRef]

[16] Wu Y, Wang Z, Wang S. Human resource allocation based on fuzzy data mining algorithm. Complexity 2021;2021:9489114. [CrossRef]

[17] Subrahmanya SVG, Shetty DK, Patil V, Hameed BMZ, Paul R, Smriti K, et al. The role of data science in healthcare advancements: applications, benefits, and future prospects. Ir J Med Sci 2022;191:1473–1483. [CrossRef]

[18] Navarro FCP, Mohsen H, Yan C, Li S, Gu M, Meyerson W, et al. Genomics and data science: an application within an umbrella. Genome Biol 2019;20:109. [CrossRef]

[19] Davenport TH, Patil DJ. Data scientist: the sexiest job of the 21st century. Harv Bus Rev 2012;90:70–76, 128.

[20] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35:1798–828. [CrossRef]

[21] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–444. [CrossRef]

[22] Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015;61:85–117. [CrossRef]

[23] Yang B, Xu Y. Applications of deep-learning approaches in horticultural research: a review. Hortic Res 2021;8:123. [CrossRef]

[24] Deng L, Yu D. Deep learning: methods and applications. Found Trends Signal Process 2014;7:197–387. [CrossRef]

[25] Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification, in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3642–3649: IEEE. [CrossRef]

[26] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks, Adv Neural Inform Process Syst 2012;25:1097–1105.

[27] Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.

[28] Bastani K, Namavari H, Shaffer J. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints, Exp Syst Appl 2019;127:256–271. [CrossRef]

[29] Roque C, Cardoso JL, Connell T, Schermers G, Weber R. Topic analysis of road safety inspections using latent Dirichlet allocation: A case study of roadside safety in Irish main roads. Accid Anal Prev vol. 2019;131:336–349. [CrossRef]

[30] De Clercq D, Wen Z, Song Q. Innovation hotspots in food waste treatment, biogas, and anaerobic digestion technology: A natural language processing approach. Sci Total Environ 2019;673:402–413. [CrossRef]

[31] Griffiths TL, Steyvers M. Finding scientific topics. Proceed Nat Acad Sci 2004;101:(Suppl 1)5228–5235. [CrossRef]

[32] Chen C, Ren J, Forum latent Dirichlet allocation for user interest discovery. Knowledge Based Syst 2017;126:1–7. [CrossRef]

[33] Silge J, Robinson D. Topic modeling. Available at: https://www.tidytextmining.com/topicmodeling.html Last Accessed Date: 21.08.2023.

[34] Saitta L, Zucker J-D. Abstraction in Artificial Intelligence, in Abstraction in Artificial Intelligence and Complex Systems: Springer, 2013, pp. 49–63. [CrossRef]

[35] Musés CA. Aspects of the Theory of Artificial Intelligence: The Proceedings of the First International Symposium on Biosimulation Locarno, June 29-July 5, 1960. Springer, 2013.

[36] Ratsch U, Richter MM, Stamatescu I-O. Intelligence and artificial intelligence: An interdisciplinary debate. Berlin, Heidelberg, Dordrecht, and New York: Springer Science & Business Media; 2013.

[37] Müller VC. Philosophy and Theory of Artificial Intelligence. New York: Springer; 2012.

[38] Morabito V. Big Data And Analytics, Strategic and Organisational Impacts. New York: Springer; 2015. [CrossRef]

[39] Suh S, Anthony T. Big Data and Visual Analytics. New York: Spinger; 2017. [CrossRef]

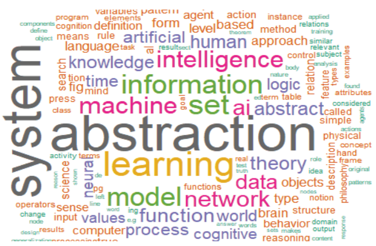
[40] Feinleib D. Big Data Bootcamp: What Managers Need to Know to Profit from the Big Data revolution. New York: Apress; 2014. [CrossRef]

[41] Márquez FPG, Lev B. Big Data Management. New York: Springer; 2017.

[42] Furht B, Villanustre F. Big Data Technologies and Applications. New York: Springer, 2016. [CrossRef]
[43] Márquez FPG, Lev B. Advanced Business Analytics. New York: Springer; 2015.
[44] Bhaduri SN, Fogarty D. Advanced Business Analytics. New York: Springer; 2016. [CrossRef]
[45] Jank W. Business Analytics for Managers. New York: Springer; 2011. [CrossRef]
[46] Saxena R, Srinivasan A. Business Analytics: A Practitioner's Guide. Berlin, Heidelberg, Dordrecht, and New York: Springer Science & Business Media; 2012.
[47] Ohri A. R for Business Analytics. Berlin, Heidelberg, Dordrecht, and New York: Springer Science & Business Media; 2012.
[48] Aggarwal CC. Data Mining: the textbook. New York: Springer; 2015. [CrossRef]
[49] Williams G. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. Berlin, Heidelberg, Dordrecht, and New York: Springer Science & Business Media; 2011. [CrossRef]
[50] Gaber MM. Journeys to Data Mining: Experiences from 15 Renowned Researchers. Berlin, Heidelberg, Dordrecht, and New York: Springer Science & Business Media; 2012. [CrossRef]
[51] Bramer M. Principles of Data Mining. New York: Springer, 2007.
[52] Gaber MM. Scientific Data Mining and Knowledge Discovery. New York: Springer; 2009. [CrossRef]
[53] Akerkar R, Sajja PS. Intelligent Techniques for Data Science. New York: Springer; 2016. [CrossRef]
[54] Zou B, Han Q, Sun G, Jing W, Peng X, Lu Z. Data Science: Third International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2017, Changsha, China, September 22-24, 2017. [CrossRef]
[55] Rose D. Data Science: Create Teams that Ask the Right Questions and Deliver Real Value. New York: Apress; 2016.
[56] Palumbo F, Montanari A, Vichi M. Data Science: Innovative Developments in Data Analysis and Clustering. New York: Springer; 2017. [CrossRef]
[57] Li M. Mathematical Problems in Data Science: Theoretical and Practical Methods. New York: Springer; 2017.
[58] Deng L, Liu Y. Deep learning in Natural Language Processing. New York: Springer; 2018. [CrossRef]
[59] Skansi S. Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence. New York: Springer; 2018. [CrossRef]
[60] Charu C. Neural Networks and Deep Learning: A Textbook. New York: Springer; 2019.
[61] Patterson J, Gibson A. Deep Learning: A Practitioner's Approach. Sebastopol, California: O'Reilly Media, Inc.; 2017.
[62] Huang K-Z, Yang H, King I, Lyu MR. Machine Learning: Modeling Data Locally and Globally. Berlin, Heidelberg, Dordrecht, and New York: Springer Science & Business Media; 2008. [CrossRef]
[63] M. Amouzegar, Advances in Machine Learning and Data Analysis. Berlin, Heidelberg, Dordrecht, and New York: Springer Science & Business Media; 2009.
[64] Kubat M. An Introduction to Machine Learning. New York: Springer; 2017. [CrossRef]
[65] Ghatak A. Machine Learning With R. New York: Springer; 2017. [CrossRef]
[66] Swamynathan M. Mastering Machine Learning With Python In Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python. New York: Apress; 2019. [CrossRef]
[67] Gharehchopogh FS, Khalifelu ZA. Analysis and evaluation of unstructured data: text mining versus natural language processing. in 2011 5th International Conference on Application of Information and Communication Technologies (AICT), 2011, pp. 1–4: IEEE. [CrossRef]
[68] Yin S, Wang G, Qiu Y, Zhang W. Research and implement of classification algorithm on web text mining, in Third International Conference on Semantics, Knowledge and Grid (SKG 2007), 2007, pp. 446–449: IEEE. [CrossRef]
[69] Weiguo F, Linda W, Stephanie R, Zhongju Z. Tapping into the power of text mining. J ACM 2005;49:76–82. [CrossRef]
[70] Cheng CH, Chen HH. Sentimental text mining based on an additional features method for text classification. PloS One 2019;14:e0217591. [CrossRef]
[71] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3:993–1022.
[72] Jayapal A, Emms M. Topic Models-LDA-Experiments, 2014.
[73] Ramage D, Hall D, Nallapati R, Manning CD, Labeled LDA. A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009, pp. 248–256. [CrossRef]
[74] Silge J, Robinson D. Text mining with R: A tidy approach. Sebastopol, California: O'Reilly Media, Inc.; 2017.
[75] Hornik K, Grün B, topicmodels: An R package for fitting topic models. J Stat Softw 2011;40:1–30. [CrossRef]
[76] Wei T, Simko V, Levy M, Xie Y, Jin Y, Zemla J. Package 'corrplot'. Statistician 2017;56:e24.

Appendix Word clouds for each topic

Artificial Intelligence



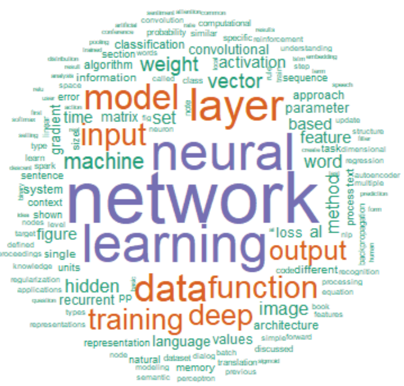
Big Data

Business Analytics

Data Mining

Data Science

Deep Learning



Machine Learning

Word Cloud of All (35) Books