CS-Insights: A System for Analyzing Computer Science Research

Terry Ruas[1], Jan Philip Wahle[1], Lennart Küll[2], Saif M. Mohammad[3], Bela Gipp[1]
[1]University of Göttingen Germany, [2]University of Wuppertal, [3]National Research Council Canada
[1]{ruas, wahle, gipp}@uni-goettingen.de
[2]leon.kuell@uni-wuppertal.de
[3]saif.mohammad@nrc-cnrc.gc.ca

Abstract

This paper presents CS-Insights, an interactive web application to analyze computer science publications from DBLP through multiple perspectives. The dedicated interfaces allow its users to identify trends in research activity, accessibility, author's productivity, venues, statistics, topics of interest, and the impact of computer science research on other fields. CS-Insights is publicly available, and its modular architecture can be easily adapted to domains other than computer science.[1]

1 Introduction

While the number of digital scientific publications keeps growing fast, our ability to analyze them does not follow the same speed, preventing us from uncovering implicit patterns among its main features (e.g., authors, venues) (Bornmann et al., 2021). The challenge in analyzing large amounts of articles, and possibly any type of data, comes largely from its storage and processing. Current solutions often focus on the storage of data (DBLP[2]), specific sub-areas (NLP Scholar (Mohammad, 2020)) or dataset augmentation (D3 (Wahle et al., 2022c)). Other robust alternatives such as Web of Science[3] and SCOPUS[4] offer more complete solutions, i.e., data storage, crawling, processing, and visualization, but unfortunately lie behind paywalls, which is prohibitive to those who would benefit the most from their resources (e.g., institutions in developing countries as they usually have a restrictive budget for such services).

With more than 389 million records, Google Scholar is probably the most comprehensive academic repository today (Gusenbauer, 2019). Even if not all records are publicly available, for research labs with limited funding, it is computationally unfeasible to build a tool that can store and process such massive amounts of data. Therefore, efforts in exploring scientific publications are focused on specific niches, such as NLP Scholar (Mohammad, 2020), a tool to analyze natural language processing publications, or PubMed[5], a system for medical sciences. Areas without dedicated solutions rely on scientometric studies on either general data repositories (e.g., arXiv), tools (e.g., VOSViewer (van Eck and Waltman, 2010)) or (semi-)manual approaches (Ruas and Pereira, 2014; Saheb et al., 2021).

As computer science publications have been growing exponentially in the last decades (Wahle et al., 2022d), and their presence in solving or facilitating other field-related problems is undeniable (e.g., plagiarism detection (Wahle et al., 2022a,b), media bias (Spinde et al., 2021, 2022)); we see computer science as a promising environment for developing a system to help understand its publications in an automated and transparent way.

We propose Computer Science Insights (CS-Insights), an open source[6] web-based application to retrieve and analyze computer science publications from DBLP through multiple perspectives interactively. CS-Insights is freely available through the project homepage[7]. The interactive tool enables its users to explore large amounts of data intuitively in the browser through several specialized dashboards: *papers*, *authors*, *venues*, *types of paper*, *fields of study*, *publishers*, *citations*, and *LDA topics*. Each dashboard offers a dedicated visualization panel and eight additional filters that can be combined to investigate authors, venues, publishers, and their publications. This paper details how we built our interactive tool, its main components, capabilities, and some exploratory experiments to show its main

[1]Demo: https://youtu.be/1ryjLK7lZXA
[2]https://dblp.org
[3]https://www.webofscience.com
[4]https://www.scopus.com/
[5]https://pubmed.ncbi.nlm.nih.gov/
[6]https://github.com/jpwahle/cs-insights
[7]https://jpwahle.com/cs-insights

functionalities.

In addition to the examples presented in Section 5, CS-Insights can also be used to explore particular topics of interest for individual researchers and find relevant subject experts, influential publications, or important venues to inform their own research. Conference organizers and research organizations such as ACL can use CS-Insights to identify the community's needs, inform policy decisions, and track how the implemented interventions impact broad publication trends over time. For example, in NLP, one can use CS-Insights to track how much research is performed in various fields; it can be used to track citation gaps across authors and venues; and in the future, it can be used to track the influence of big technology companies, highly-funded universities, and governments, and estimate the amount of research performed in various languages.

2 Related work

Even though tools such as Google Scholar[8], Semantic Scholar[9], NLP Scholar (Mohammad, 2020), and NLPExplorer (Parmar et al., 2020) provide information on scientific documents, their use is limited. Google Scholar and Semantic Scholar focus their platforms on authors and their metrics (e.g., h-index, number of papers, citations) but lack details on venues and publishers. Additionally, neither Google Scholar nor Semantic Scholar offer an interactive, customizable platform to browse their databases, preventing users from exploring explicitly available features on their website (e.g., the field of study). NLP Scholar and NLPExplorer offer a more personalized solution but only focus on natural language processing (NLP) publications. They use an interactive framework to visualize and correlate different characteristics simultaneously (e.g., venues, authors, and the field of study). NLP Scholar offers a dynamic interface as its reports are built on top of Tableau, preventing its use as an API. While NLP Scholar aligns the information between ACL Anthology and Google Scholar (45K articles), NLPExplorer uses only the ACL Anthology (77K articles) as a data source, thus, limiting their use for analyses of broader trends in computer science research.

CS-Insights offers four advantages over its competitors. First, CS-Insights uses DBLP, the largest collection of computer science publications, with over 6M, including ACL Anthology, arXiv, and sentient metadata (e.g., paper abstracts, author affiliations). Second, CS-Insights can be accessed as a REST API to retrieve the information we display in our frontend, enabling CS-Insights to be easily incorporated into other studies. Third, different from Web of Science (Clarivate) and SCOPUS (Elsevier), CS-Insights is a free and transparent tool to support anyone interested in investigating publications, independently of financial status or any other ethical barrier. Fourth, our architecture provides a scalable, customizable, and responsive service.

3 Main components

CS-Insights is composed of three main boards: A. *dashboards*, B. *filters*, and C. *visualizations*. *Dashboards* control the main views available in CS-Insights. *Filters* allow users to select what papers are visualized. *Visualizations* includes a series of interactive visualizations of the selected papers. Figure 1 shows the landing page of our system with a default filter setting and the papers dashboard.

3.1 Dashboards

The *dashboards* (A) are pre-set configurations focusing on specific elements of papers (e.g., number of papers, number of authors of papers, number of citations). There are eight dashboards in total.

Papers shows the absolute number of papers per year; **Authors** shows the full name and number of authors for each paper; **Venues** shows where each paper was published (e.g., ACL); **Types of Papers** lists types of papers published according to their BibTex entry classification (e.g., article); **Fields of Study** shows the number of papers in each research area (e.g., computer science); **Publishers** shows the responsible institution for publishing a given paper[10]; **Citations** shows the accumulated incoming (i.e., how often a single paper was cited) and outgoing citations (i.e., the number of bibliography entries in a single paper); **LDA Topics** performs a topic modeling analysis (Blei et al., 2003) considering papers' titles and abstracts.

3.2 Filters

Filters are adjustable feature-value pairs that can be configured to select a set of papers to be visualized. There are eight filters (B1) that can be applied for each available dashboard (A); six for textual values

[8] https://scholar.google.com/

[9] https://www.semanticscholar.org/

[10] Most publications in DBLP leave this field blank.

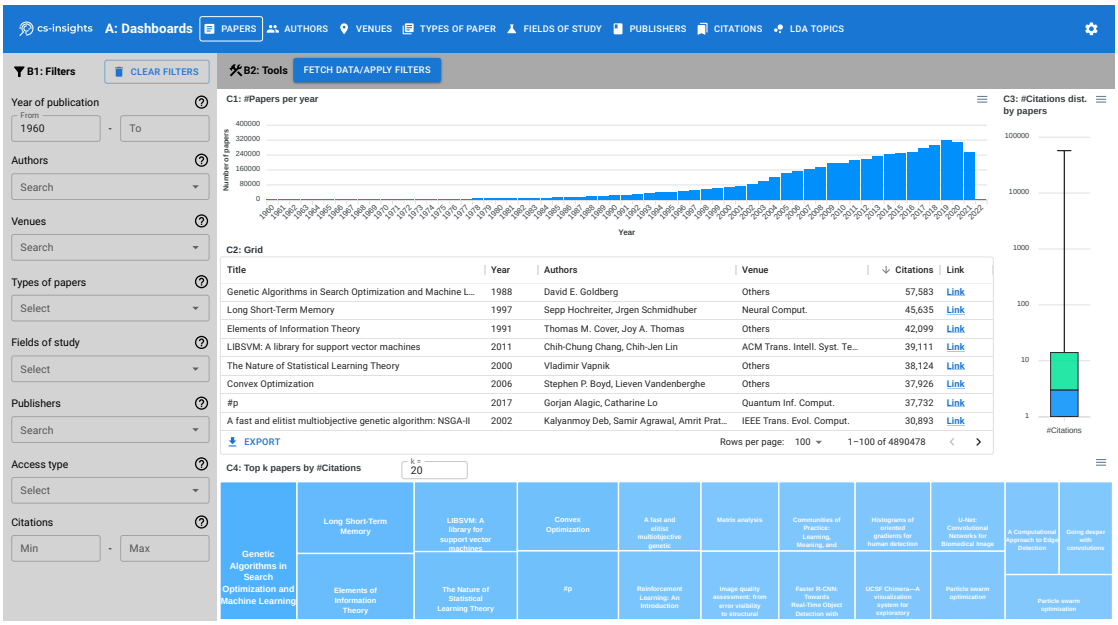

Figure 1: Frontend interface for CS-Insights.

and two for numeric ones. The "Fetch Data/Apply Filters" button loads a new data batch with the selected filters (B2).

All textual filters work with auto-completion and regular expressions, thus, while typing, the user is already presented with possible matching string suggestions. For the filters *Types of papers*, *Field of study*, and *Access type*, pre-set values are presented in a drop-down menu. For example, when clicking on *Types of papers*, the suggested article, proceedings, book, in collection, Ph.D., and master thesis appear. Both numerical filters (year of publication and citations) work by restricting minimum and maximum values. Different filters work together through a logical `AND` and values on the same filter with a logical `OR`. All filters can be cleared using the "Clear filters" button at the top left corner (B). To obtain more information about the filters and their match conditions, one can hover over the question mark icon next to their heading.

3.3 Visualizations

There are four common visualization elements for each dashboard, as Figure 1 (C) shows: *#[Dashboard] per year* (C1), *Paper Details Grid* (C2), *#Citations distribution* (C3), and *Top k by #Citations* (C4). The only exceptions are the *Citations* and *LDA Topics* dashboards, which have specific visualization elements. The former displays *Incoming* and *Outgoing* citations for the selected papers as a bar chart over time, as well as a box plot for both, respectively. The latter shows the semantic clusters and their list of frequent terms about the selected papers. The visualization elements can be exported in several formats (e.g., csv, png).

#[*Dashboard*] per year (C1) shows the amount of a given dashboard main element (e.g., papers) per year. For example, in the *Venues* dashboard, one can see a bar chart displaying the number of unique venues where the selected papers were published by year. Hovering over a bar reveals the exact number of entries for that year.

Paper Details Grid (C2) displays the available details for each dashboard choice in a table format. For example, in the *Papers* dashboard, the first column contains the paper's title followed by its year of publication, list of authors, venue, number of citations, and paper link (when available). For the *Authors* dashboard, the grid includes the name of the authors, the first and last year of publication, the number of papers, and citations.

#Citations distribution (C3) shows the distribution of citations for the selected papers. For all dashboards, except *Papers*, one can also select the number of papers as an alternative metric (B2). Hovering over the boxplot reveals the exact values for the minimum, 25% quartile, median, 75% quartile, and maximum.

Top k by #Citations (C4) displays the top k elements based on the number of papers (regardless of publication year) in a treemap format. As in C3, the *Papers* dashboard uses the number of citations as

Figure 2: Top k research fields by #Citations. The size of each tile is proportional to the #citations for that field.

a metric to generate its output. All the other dashboards also offer the option of selecting the number of papers. The value of k can be adjusted using a text field. When the text in C4's boxes is too large, we collapse them for readability purposes.

4 Architecture

The CS-Insights architecture consists of four main components[6]: *frontend*, *backend*, *prediction endpoint*, and *crawler*. Our solution is available as a free web application without the need for any installation as it runs in many web browsers[7]. To guarantee a flexible and modular setup, every component in CS-Insights runs on its own docker container. A more comprehensive list of technologies used in CS-Insights is detailed in Appendix A.1.

Frontend. It is responsible for presenting the main components of our tool, i.e., *dashboards*, *filters*, and *visualizations* (Figure 1). In the frontend's interface, one can filter, retrieve, and visualize computer science publications. We used TypeScript and React as web frameworks as they are open-access and have large community support. For visualization, we use ApexCharts and Material UI.

Backend. To access, retrieve, aggregate, and analyze data, we created a REST API *backend* that controls how to access data and performs computationally expensive tasks (e.g., aggregating citations of all authors for each paper available). CS-Insights uses TypeScript with Node.js as JavaScript runtime and Express.js for the HTTP requests. For persistent data storage, we use MongoDB, with mongoose providing the object document mapping.

Prediction endpoint. It is responsible for the training and prediction of models in the *LDA Topics* dashboard used to generate the semantic topics and the lists of the most frequent and salient terms. The project is implemented in python 3, and we use gensim's LDA (Blei et al., 2003) implementation for topic models. We implement visualizations using pyLDAvis (Sievert and Shirley, 2014). As training and inference typically require processing tens of thousands of documents, we create a dedicated service to maintain models, distribute them on the available compute infrastructure, assign them to compute jobs, and consolidate results.

Crawler. We use DBLP in our workflow as the main data source to feed CS-Insights with computer science publications. DBLP is currently the largest computer science repository with more than 6M documents. To keep CS-Insights up-to-date with the most recent publications, the *crawler* downloads the latest release from DBLP, corresponding full texts, and extracts their metadata. This preprocessing step uses the same process as in D3 (Wahle et al., 2022c), a dataset that extends DBLP with additional information.

5 Showcase experiments

To provide an overview of CS-Insights's core functionalities, we developed a collection of intuitive dashboards and filters that investigate broad trends in computer science publications.

5.1 Papers overview

Figure 1 shows the distribution of papers from 1960 to 2021 on the *visualization* (C) board. No filter (B) was selected to provide a high-level overview of the entire dataset. In the *#Papers per year* visualization (C1), the number of publications constantly grew, except for the last two years. A small decline over 2020 (and possibly 2021) can be explained by the COVID-19 pandemic that affected researchers worldwide. The current release includes data up to December 1st, 2021[11].

5.2 Fields of study

As the number of publications has been growing rapidly, it is natural to question in which areas these publications are increasing. Even though the number of research fields has not changed in the last couple of years (Figure 5), the number of papers between them differs greatly, as Figure 2 shows. Computer science is overrepresented when compared to other fields, such as mathematics.

[11] Our next data extraction will include 2021 and 2022.

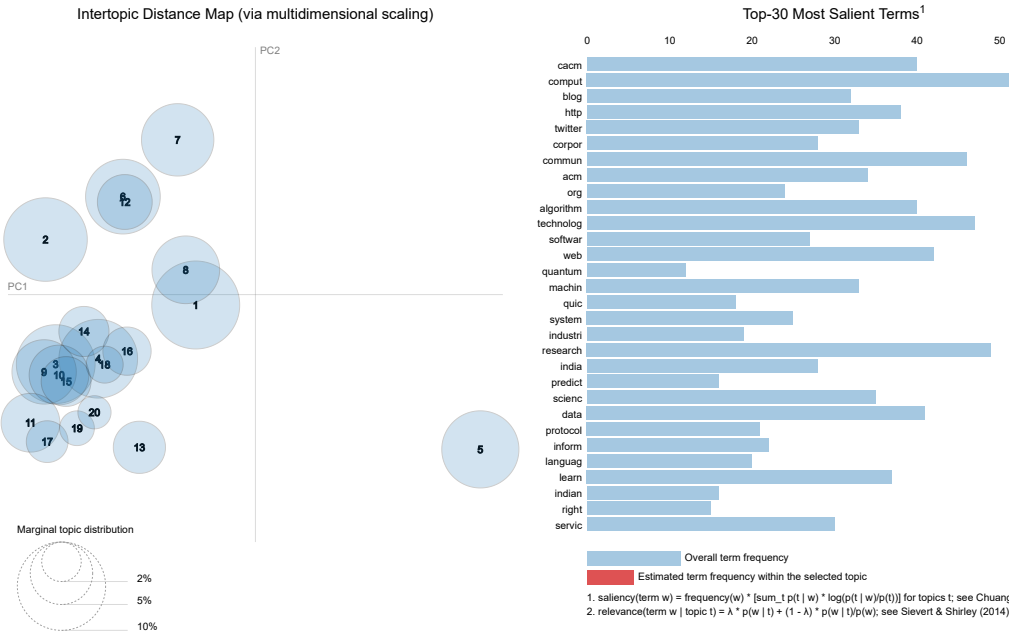

Figure 3: (Left) The cluster of the obtained topics. Each topic is represented by a circle whose radius/size is proportional to the marginal topic distribution. (Right) LDA topics for the Communications of the ACM in 2019. The list of terms for all clusters (current state) or the terms for one specific cluster when selected.

Table 1: Number of papers (#papers) and citations (#citations) for the top five fields of study in DBLP.

Field	#Papers	#Citations
Computer Science (CS)	4 189 349	96 935 856
Mathematics (MA)	695 351	21 716 666
Medicine (MD)	267 794	10 179 913
Engineering (EN)	323 866	7 550 670
Psychology (PS)	80 571	3 199 135

In absolute numbers, computer science has approximately twice as many papers and citations as all other fields combined[12]. In Table 1, we show a sample for the five top fields of study with their number of papers and citations. For a complete list of all fields of study and the boxplot of their distribution, see Table 4 and Figure 4 in Appendix A.

5.3 LDA topics

In the *LDA topics* dashboard, users can explore the most frequent and salient (Chuang et al., 2012) terms (stemmed words) of a given collection of documents through an LDA implementation for topic modeling (Sievert and Shirley, 2014). Figure 3 shows the topic distribution for the "Communications of the ACM" venue in 2019. The output in this dashboard is divided into the semantic cluster topics (left) and the list of the most frequent and salient terms (right). Both parts are produced based on the text in the titles and abstracts of papers, which are parsed, stemmed, and cleaned[13] (e.g., stopword removal, punctuation removal, punctuation removal). In Figure 3, we see an overlap of clusters, indicating their semantic proximity. We also include the same experiment for the years 1999 (Figure 6) and 2009 (Figure 7) for comparison in the Appendix A. When no cluster is selected, the plots consider all titles and abstracts to compose their list of terms. When hovering over a cluster, the 30 most relevant terms of the selected cluster are shown on the right as red bars while continuing to show the overall frequency of those 30 terms in all clusters as blue bars. One can also identify clusters associated with a term by clicking on the desired terms directly.

5.4 Additional experiments

In the following, we provide additional examples of analyses that can be performed with CS-Insights. The analytical capabilities of the tool reach far beyond the portraited examples, and with active development, more features will become available to answer more questions about CS research. For space reasons, the tables and figures are included in the Appendix A, and further analysis is included in (Küll, 2022).

[12] Documents can be associated with more than one field.

[13] https://radimrehurek.com/gensim/parsing/preprocessing.html

Table 2 states the top 30 most salient terms for the top 5 most cited authors (combined) and venues (individual) in the field. This table provides insight into the research areas and topics that have been most prominent among the most cited authors and venues in the field. For example, the terms in this table reveal that the topics related to "image", "segmentation", or "object" "detection" are the most prominent among the computer vision conference CVPR.

Table 3 shows the number of papers and number of citations of the top-5 most cited authors ranked by the number of papers and number of citations. This table provides information about the most cited authors in the field and how their work has been received by the research community. The data of this CS-Insights export reveals which authors have published the most papers and received the most citations and how their work has been received by the research community. For example, among all CS researchers, Ross B. Girshick (a computer vision researcher at Meta AI) has received the most citations (both on average and in total), and one of his close collaborators, Kaiming He (also at Meta AI), comes right after him.

Figure 8 visualizes venues considering their total number of papers. The grid in this figure provides insight into the conferences and journals with the most relevant papers in the field, e.g., which venues are most active and have the highest impact and how long they have been taking place.

Figure 9 illustrated venues considering their total number of citations. This figure provides insight into the conferences and journals that have received the most citations in the field. By looking at the data in this figure, it is possible to see which venues have received the most citations and how citations are generally distributed.

Figure 10 plots the LDA topics for Communications of the ACM in 2019 with cluster 1 selected. This figure provides insight into the topics that have been discussed in Communications of the ACM in 2019. The most salient terms are, for example, "latency" and "quick", or web "protocols" like "tcp" and "http".

Figure 11 again shows the LDA topics for Communications of the ACM in 2019 with the term "comput" selected. This figure additionally provides insights into the topics related to the term "comput" over Figure 10. Not only does the distribution of clusters change in accordance with the selected keyword, but also the order of terms changes. For example, "http" and "twitter" achieve much higher relative relevance when "comput" is selected.

6 Conclusion

We presented CS-Insights, an interactive, open-source, and web-based visualization tool to facilitate the exploration of computer science publications. Our tool crawls and processes papers from DBLP in a modular architecture, facilitating the maintenance and incorporation of more efficient components in the future. In future work, we plan to incorporate improvements to our tools, such as the visualization of authors' affiliations and their correlation with existing *dashboards* and collaboration graph visualizations to spot authors' and institutions' collaborations easily. The project is currently actively developed by four contributors on GitHub, with new features shipping every month. The currently planned features are available through the project roadmap[14] and currently include for example, a keyword search for papers, a split view to compare filters (e.g., authors, venues and their topics), additional impact measures (h-index, i10-index), or data expansion to 131m+ open articles on the Internet Archive Scholar[15] and FatCat[16].

Limitations

As CS-Insights is a work in progress and it relies on external resources, there are a few limitations that should be mentioned. Even though DBLP is the largest computer science repository, with an extensive list of features at its disposal, it does not contain all publications about computer science.

Not all characteristics available in our dataset are used for our analysis yet (e.g., the author's affiliation). Further, some features are sparse, such as the publisher's name, which is due to missing entries in DBLP.

Due to hardware constraints, we cannot perform topic modeling for the entire dataset at runtime. Hence, we currently cap the number of documents to 100K that can be used for training and prediction in the *LDA Topics* dashboard. In future work, we plan to precompute common chunks of documents to allow for analyses of millions of documents.

[14] https://github.com/users/jpwahle/projects/1/
[15] https://scholar.archive.org/
[16] https://fatcat.wiki/

Ethics Statement

CS-Insights (like other open-access analysis tools) can be misused to provide a biased and unilateral view of computer science research or related areas. As D3 and DBLP are our data sources, all the documents used in CS-Insights are in English; thus, a variety of other languages missing in our analysis. Currently, our experiments and showcases hold truth for DBLP, which is an expressive subset of computer science publications, but by no means complete. Consequently, local events and underrepresented languages are also not included in CS-Insights. However, researchers can use CS-Insights to analyze trends of research about low-resource languages (including sign language) whenever it is mentioned in the title or abstract.

Another possible concern is with the non-anonymization in our visualizations. We build CS-Insights intending to facilitate the overview of computer science publications, their authors, venues, and topics of interest. Therefore, we do not omit authors' names and their number of publications or citations, which can be misused. For example, one can use available APIs, such as *genderize*[17], to infer an author's gender and propagate a false correlation between productivity and gender.

The (un)intentional one-sided report of computer science publications can be used for specific political agendas. The connection between the author's affiliations and potential cross-reference with their country can propagate a limited and unfounded view of a country or institution's scientific status. One should be aware that there is no unique repository for all computer publications (or any other research field) worldwide. Hence, specific and collaborative efforts should be encouraged to obtain a more accurate perspective in our analysis.

CS-Insights, its components, and data are licensed to the general public under a copyright policy that allows unlimited reproduction, distribution, and hosting on any website or medium[18,19]. Hence, anyone accessing our tool can exploit its limitations and inherited biases to propagate and amplify societal problems.

In the retrieval and pre-processing of our data, there are a few string-parsing-matching inconsistencies (e.g., umlauts in authors' names and multiple name variations for the same author). As CS-Insights is an ongoing project with regular releases, we hope to address the shortcomings in the future. A roadmap and issue board are available through the project's repository for more details.

[17] https://genderize.io/

[18] https://dblp.org/db/about/copyright

[19] https://github.com/jpwahle/cs-insights/blob/main/LICENSE

Acknowledgements

This work was partially supported by the IFI program of the German Academic Exchange Service (DAAD) under grant no. 57515252. We also thank Tom Neuschulten and Alexander von Tottleben for developing the prediction endpoint.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):224.

Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012. Termite: visualization techniques for assessing textual topic models. In *AVI*, pages 74–77. ACM.

Michael Gusenbauer. 2019. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1):177–214.

Lennart Küll. 2022. Analyzing the state of computer science research with the dblp discovery dataset.

Saif M. Mohammad. 2020. NLP Scholar: An Interactive Visual Explorer for Natural Language Processing Literature. *arXiv:2006.01131 [cs]*.

Monarch Parmar, Naman Jain, Pranjali Jain, P. Jayakrishna Sahit, Soham Pachpande, Shruti Singh, and Mayank Singh. 2020. NLPExplorer: Exploring the Universe of NLP Papers. In *Advances in Information Retrieval*, volume 12036, pages 476–480. Springer.

Terry Lima Ruas and Luciana Pereira. 2014. Como construir indicadores de Ciência, Tecnologia e Inovação usando Web of Science, Derwent

World Patent Index, Bibexcel e Pajek? *Perspectivas em Ciência da Informação*, 19(3):52–81.

Tahereh Saheb, Tayebeh Saheb, and David O. Carpenter. 2021. Mapping research strands of ethics of artificial intelligence in healthcare: A bibliometric and content analysis. *Computers in Biology and Medicine*, 135:104660.

Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70. ACL.

Timo Spinde, Jan-David Krieger, Terry Ruas, Jelena Mitrović, Franz Götz-Hahn, Akiko Aizawa, and Bela Gipp. 2022. Exploiting Transformer-Based Multitask Learning for the Detection of Media Bias in News Articles. In Malte Smits, editor, *Information for a Better World: Shaping the Global Future*, volume 13192, pages 225–235. Springer.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177. ACL.

Nees Jan van Eck and Ludo Waltman. 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538.

Jan Philip Wahle, Terry Ruas, Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2022a. Identifying Machine-Paraphrased Plagiarism. In Malte Smits, editor, *Information for a Better World: Shaping the Global Future*, volume 13192, pages 393–413. Springer International Publishing, Cham.

Jan Philip Wahle, Terry Ruas, Frederic Kirstein, and Bela Gipp. 2022b. How large language models are transforming machine-paraphrased plagiarism. *arXiv preprint arXiv:2210.03568*.

Jan Philip Wahle, Terry Ruas, Saif Mohammad, and Bela Gipp. 2022c. D3: A massive dataset of scholarly metadata for analyzing the state of computer science research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2642–2651, Marseille, France. European Language Resources Association.

Jan Philip Wahle, Terry Ruas, Saif M. Mohammad, and Bela Gipp. 2022d. D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research. *arXiv:2204.13384 [cs]*.

A Appendix

A.1 Technologies

Here we list relevant technologies used in CS-Insights:

Environment

- Docker - https://www.docker.com

Frontend

- TypeScript - https://www.typescriptlang.org
- React - https://reactjs.org/
- ApexCharts - https://apexcharts.com/react-chart-demos/
- Material UI - https://mui.com/

Backend

- MongoDB - https://www.mongodb.com
- mongoose - https://mongoosejs.com/
- Node.js - https://nodejs.org/en/
- TypeScript - https://www.typescriptlang.org
- Javbascript - https://www.w3.org/standards/webdesign/script
- Express.js - https://expressjs.com/
- express-restify-mongoose - https://florianholzapfel.github.io/express-restify-mongoose/

Prediction endpoint

- gensim - https://radimrehurek.com/gensim/models/ldamodel.html

- pyLDAvis - https://pyldavis.readthedocs.io/en/latest/readme.html

Crawler

- aiohttp - https://docs.aiohttp.org/en/stable/
- GROBID - https://grobid.readthedocs.io/en/latest/

A.2 Additional Figures & Tables

Tables 2 and 3 and Figures 8 to 11 (in the pages ahead) show example interactions with CS-Insights that were mentioned in Section 5.4.

Table 2: Top-30 most salient topics for top-5 most cited authors (combined) and conferences (individual). *Italic* terms are only present in the author's topcis. **Bold** terms are common between authors and venues.

Cited Authors	CVPR	NeuroImage	IT. PA. M. Int.	IT. Inf. Theory	Com. ACM
network	**imag**	connect	paper	code	program
data	**result**	**network**	**propos**	channel	**data**
imag	**featur**	activ	**result**	bound	**algorithm**
pattern	**network**	function	**imag**	decod	comput
mine	**video**	respons	model	**algorithm**	softwar
problem	**segment**	**imag**	**learn**	paper	**inform**
face	**propos**	**data**	object	**inform**	**problem**
cluster	detect	**result**	match	sequenc	develop
object	**shape**	cortic	**algorithm**	**network**	languag
video	model	model	surfac	capac	time
propos	**object**	ag	**problem**	error	**user**
queri	method	task	**face**	sourc	research
segment	pose	visual	**segment**	function	system
inform	local	area	recognit	signal	number
user	state	method	**featur**	estim	commun
train	**learn**	brain	**data**	rate	provid
result	point	process	**shape**	construct	new
shape	**face**	matter	structur	given	acm
algorithm	camera	predict	motion	**problem**	**result**
present	perform	region	estim	spl	gener
text	art	cortex	**network**	sub	**network**
classif	**train**	state	**cluster**	time	technolog
featur	scene	diffus	**video**	scheme	paper
databas	**problem**	suggest	label	nois	web
graph	recognit	tempor	camera	multipl	**object**
differ	track	motor	work	model	design
human	achiev	associ	**train**	lower	includ
learn	demonstr	eeg	class	**result**	scienc
type	scale	left	function	case	year
latent	motion	right	requir	upper	project
Common Topics	13	4	16	4	8

Table 3: Productivity and popularity for top 5 authors ranked by the number of papers and number of citations.

Rank	#Papers	#Citations	Year	Author	Citation avg	Venues
Paper	1 649	74 467	1977	H. Vincent Poor	45.16	648
	1 445	39 300	1997	Mohamed-Slim Alouini	27.20	462
	1 382	35 887	1989	Lajos Hanzo	25.97	462
	1 287	73 436	1980	Philip S. Yu	57.06	732
	1 260	30 704	1982	Victor C. M. Leung	24.37	702
Average	1 405	50 759	1985		35.95	601
Citation	69	146 867	2004	Ross B. Girshick	2 128.51	35
	662	123 682	1974	Anil K. Jain 0001	186.83	345
	66	114 330	2009	Kaiming He	1 732.27	33
	231	109 821	1987	Jitendra Malik	475.42	131
	454	105 025	1985	Andrew Zisserman	231.33	250
Average	296	119 945	1992		950.87	159

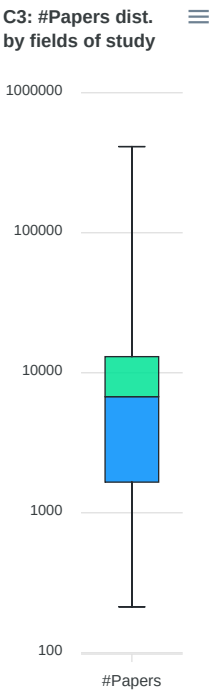

Figure 4: #Papers distribution for the field of study.

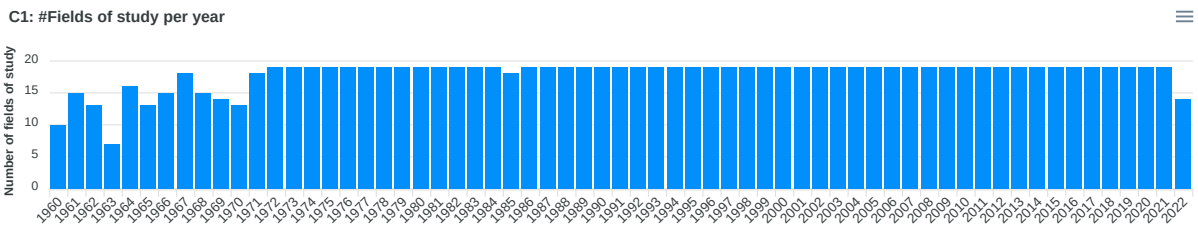

Figure 5: #Fields of study per year.

Table 4: Fields of study ranked by #Papers

Field	#Papers	#Citations	First year	Last year	Avg. Citation
Computer Science	4 189 349	96 935 856	1960	2022	23.14
Mathematics	695 351	21 716 666	1960	2022	31.23
Engineering	323 866	7 550 670	1960	2022	23.31
Medicine	267 794	10 179 913	1960	2022	38.01
Psychology	80 571	3 199 135	1961	2022	39.71
Physics	67 156	1 399 100	1960	2022	20.83
Business	57 273	1 496 619	1960	2022	26.13
Materials Science	50 013	618 573	1960	2022	12.37
Biology	30 751	985 557	1961	2022	32.05
Economics	30 076	1 062 187	1961	2022	35.32
Sociology	27 714	762 877	1962	2022	27.53
Environmental Science	26 354	410 721	1964	2022	15.58
Chemistry	17 179	547 508	1961	2021	31.87
Geology	16 926	304 257	1962	2022	17.98
Geography	16 007	407 229	1961	2021	25.44
Political Science	15 979	245 088	1960	2022	15.34
Philosophy	5 680	60 718	1960	2021	10.69
Art	4 332	18 519	1960	2021	4.27
History	2 756	49 622	1961	2021	18.01
Others	698 383	781	1960	2022	0.00

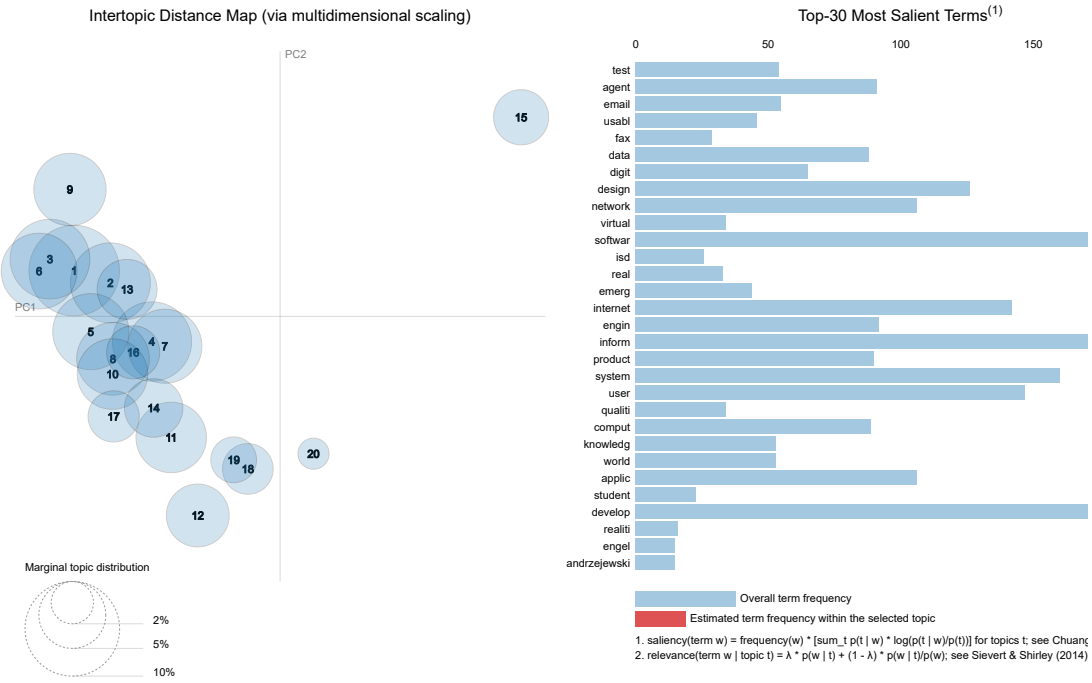

Figure 6: LDA topics for the Communications of the ACM in 1999. (Left) A cluster of the obtained topics. (Right) The list of terms for all clusters, or the terms for one specific cluster when selected.

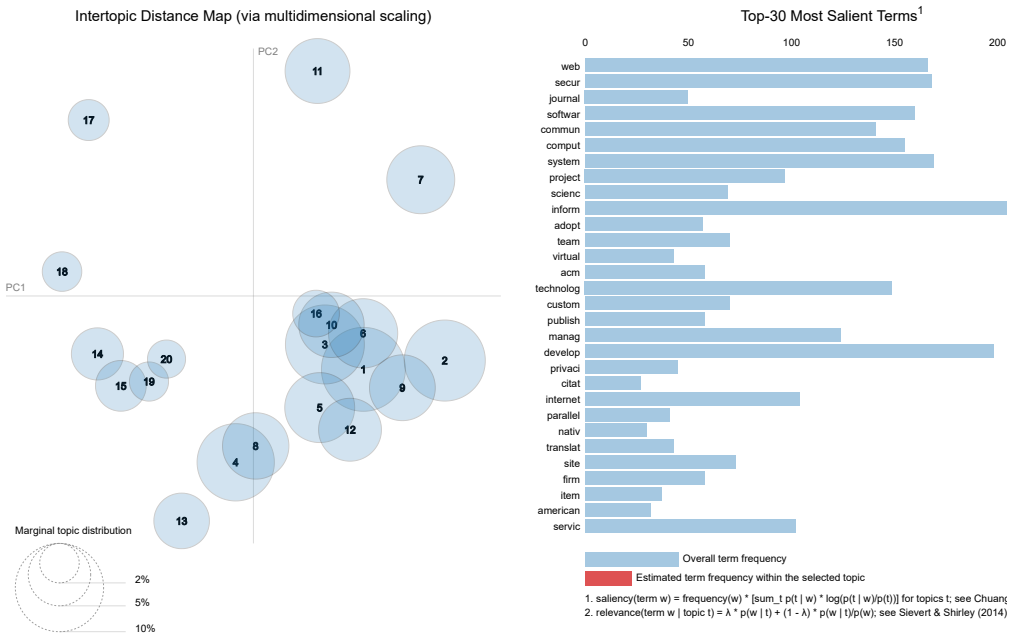

Figure 7: LDA topics for the Communications of the ACM in 2009. (Left) The cluster of the obtained topics. (Right) The list of terms for all clusters, or the terms for one specific cluster when selected.

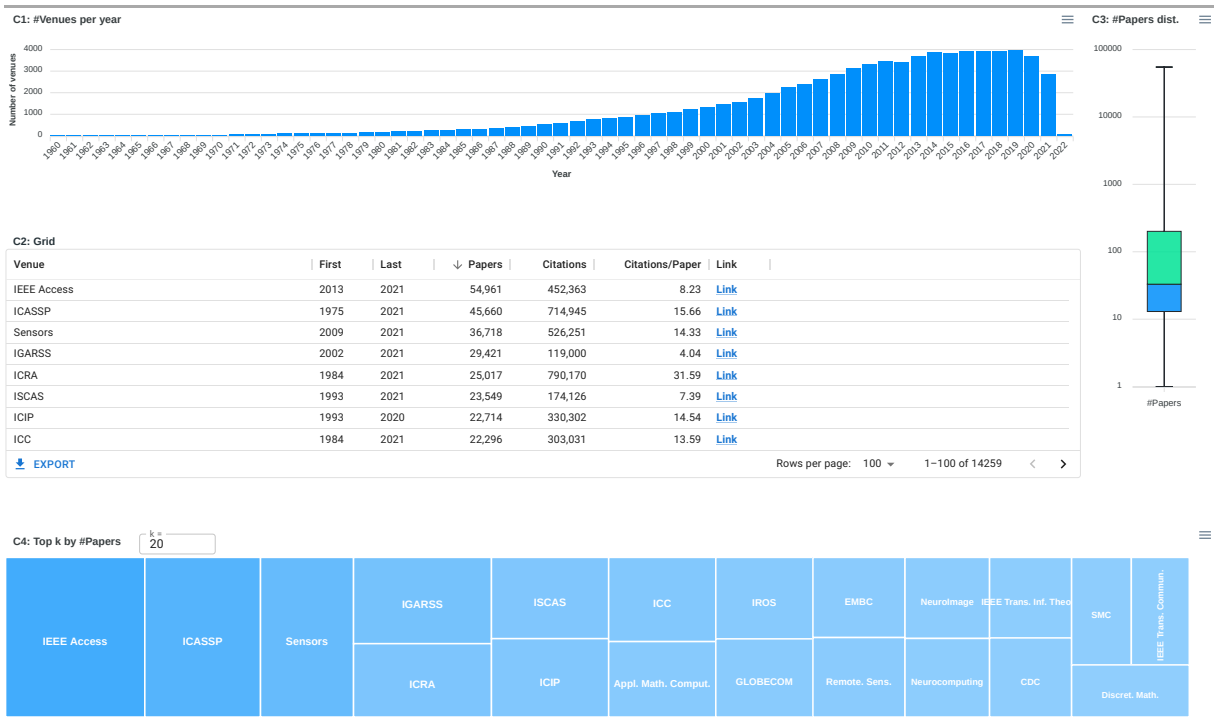

Figure 8: List of venues starting from 1960 ordered by #papers.

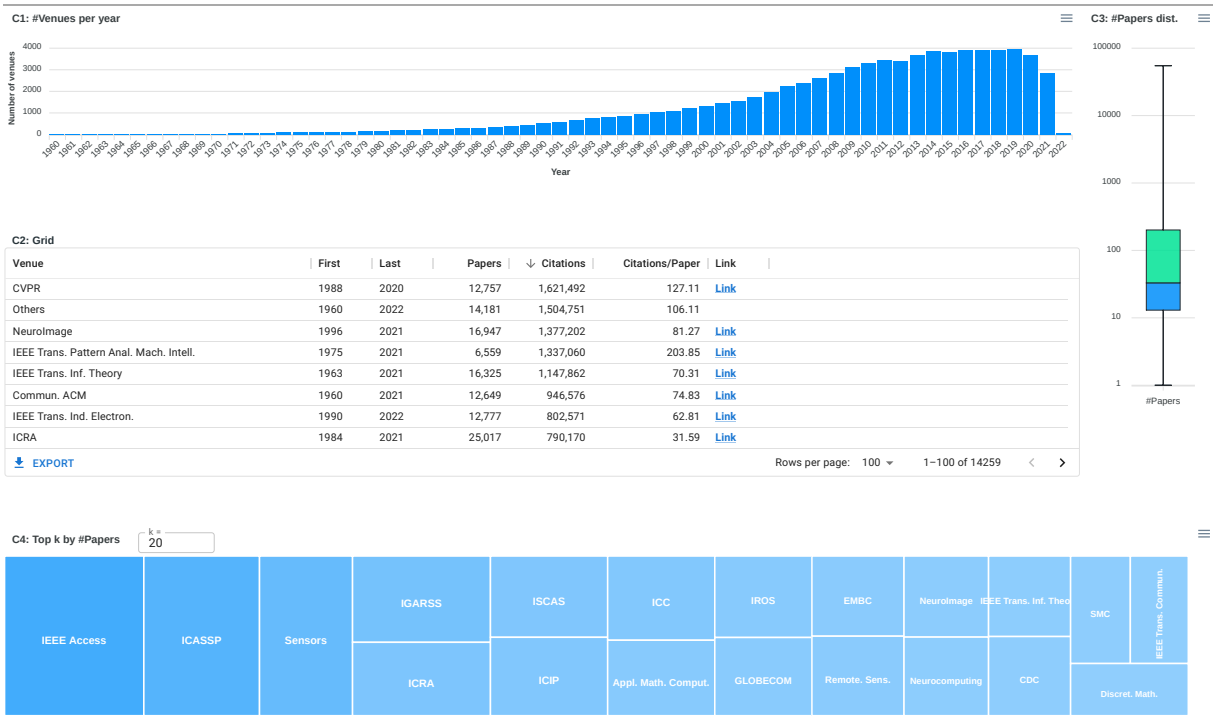

Figure 9: List of venues starting from 1960 ordered by #citations.

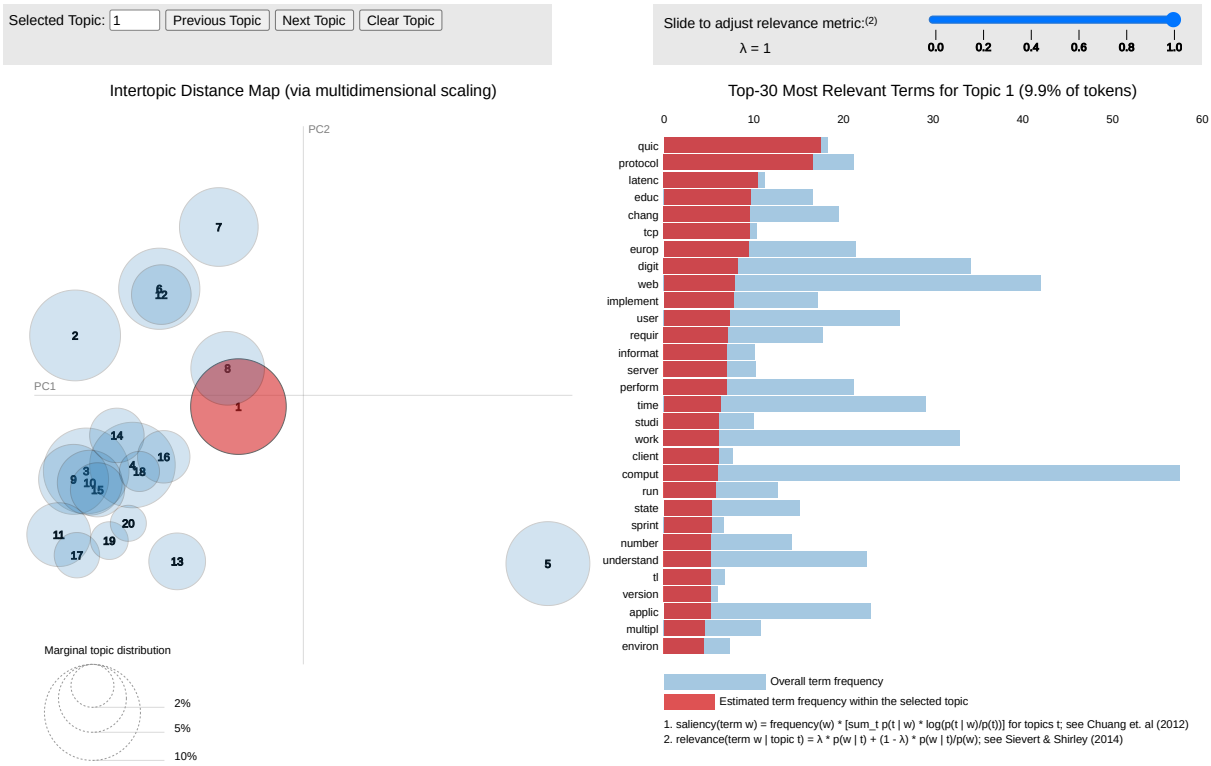

Figure 10: LDA topics for the Communications of the ACM in 2019 with cluster 1 selected.

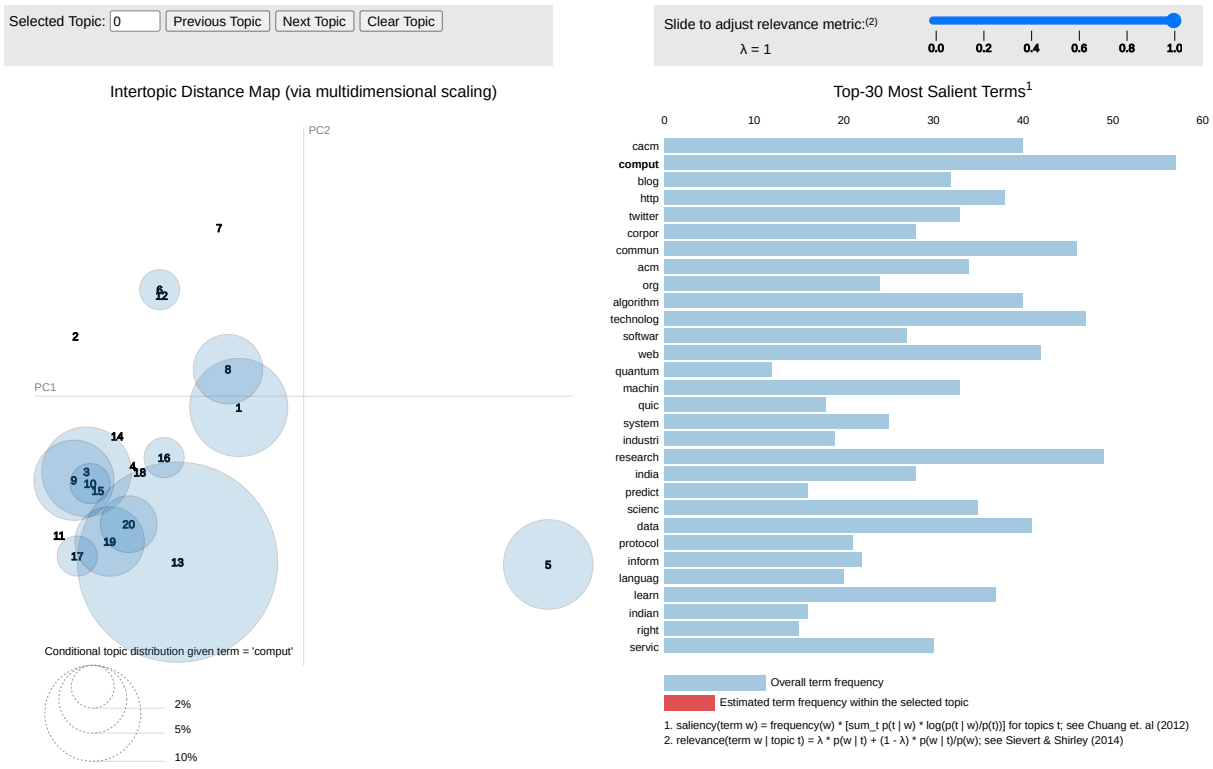

Figure 11: LDA topics for the Communications of the ACM in 2019 with the word “comput" selected.