ASIS&T AI Workshop 2022

AI IN THE REAL WORLD: STRENGTHENING CONNECTIONS BETWEEN LIS RESEARCH AND PRACTICE

ABSTRACTS

Wyndham Grand | Pittsburgh, PA
Saturday, October 29, 2022; 1-5pm ET

ORGANIZERS

Soo Young Rieh, University of Texas at Austin, USA (rieh@ischool.utexas.edu)
Clara M. Chu, University of Illinois at Urbana-Champaign, USA (cmchu@illinois.edu)
Dania Bilal, University of Tennessee-Knoxville, USA (dania@utk.edu)

asis&t
OCTOBER 29 - NOVEMBER 1, 2022
WYNDHAM GRAND, PITTSBURGH, PA
85TH ANNUAL MEETING

Table of Contents

Description 3

Artificial Intelligence (AI) in Data Curation: How Curators Re-Imagine Legacy Database Systems to Advance Equitable AI in Libraries, Archives and Museums (LAMs) 4
Sarah Bratt, Assistant Professor, University of Arizona

Computational Poetry Collection Analysis via Context-Dependent Language Models 5
Kahyun Choi, Assistant Professor, Indiana University at Bloomington

Knowledge Graph for Discovering Interdisciplinary Research Connections
Stanislava Gardasevic, PhD Candidate, Communication and Information Sciences Program; Teaching Assistant/Instructor at Library and Information Sciences University of Hawaii at Manoa 5

Images to integrated data: Digitizing and structuring historical records with deep learning 6
Sara Lafia, Postdoctoral Research Fellow, ICPSR, University of Michigan
David A. Bleckley, Data Project Manager, ICPSR, University of Michigan
J. Trent Alexander, Associate Director, ICPSR, University of Michigan

AI & Co-design in public libraries: Empowering underserved youth to cultivate symbiotic relationships between Artificial Intelligence (AI) and their communities 7
Hee Rin Lee, Assistant Professor, Michigan State University
Kahyun Choi, Assistant Professor, Indiana University
Selin Akgun, PhD Student, Michigan State University
Ji Youn Shin, Assistant Professor, University of Minnesota
Pooja Malvi, Master's Student, Michigan State University
Meredith Dedema, PhD Student, Indiana University

Historical Text Datafication and Loss: Computational Recovery of Typographical Layout Logic on an RDF Graph Featuring ML Methods 8
Huapu Liu, Doctoral Student, University of Alabama
Steven L. MacCall, Associate Professor, University of Alabama

Social Selection of Algorithms: The Unintended Consequences of Explainable AI 8
Alex Mayhew, PhD Candidate, University of Western Ontario

Digital Deep Redlining 9
Arcadio Matos, PhD Student, School of Communication and Information, Rutgers University
Vivek K. Singh, Associate Professor, School of Communication and Information, Rutgers University

Hypergraphing a Network of Inquiry, Search, and Retrieval 10
Alamir Novin, PhD Candidate, University of British Columbia

Mining scientific literature with Natural Language Processing to expand bibliometrics analysis 11
Gang Shao, Assistant Professor, Purdue Libraries and School of Information Studies, Purdue University
Joseph Eisenberg, Undergraduate Student, Purdue University

Description

Artificial Intelligence has been increasingly deployed in library and information environments and its application is rapidly expanding across user services, collection development, and library management, resulting in enhanced information search and discovery, more usable and accessible collections, and robust decision-making. The workshop builds on and extends the 2021 workshop to strengthen the ASIS&T AI community by bringing together researchers, educators, students, and practitioners interested in designing AI applications and conducting AI research in the LIS field. The goal of this workshop is to build an active community of researchers, educators, practitioners, and students who are committed to apply AI technologies effectively and innovatively in library and information environments.

This workshop covers a wide range of topics, including but not limited to, AI applications, solutions, empirical research findings, and perspectives in library and information environments. The participants will discuss lessons learned from using, designing, implementing, and/or evaluating AI applications and solutions in the context of library and information environments and work together to brainstorm potential solutions for making AI more effective and transferable for library users. Through the World Café method (i.e., collaborative dialogue with rotating multiple breakout sessions that build on each other so that issues are considered in-depth) and plenary discussion, participants will generate innovative ideas for advanced AI solutions and research agendas for future investigations. This workshop is in collaboration with SIG AI.

This workshop is a result of the partnership with ASIS&T on the IDEA Institute on AI, a project made possible in part by the Institute of Museum and Library Services planning grant: RE-246419-OLS-20.

Abstracts

Artificial Intelligence (AI) in Data Curation: How Curators Re-Imagine Legacy Database Systems to Advance Equitable AI in Libraries, Archives and Museums (LAMs)

Sarah Bratt, Assistant Professor, University of Arizona

ABSTRACT

Data curation comes from the Latin root “to care.” Yet, information systems have not historically reflected marginalized voices, rendering them invisible across physical and digital platforms. While this area has received a great deal of attention in recent years, it has been in mainstream spaces (Safiya Nobel’s Algorithms of Oppression, Future of Work @MSR, Sabina Leonelli’s critiques of “Big Data Collections”). We still lack robust discussion and principled approaches to AI in data curation, especially in the liminal spaces and emergent areas of curatorial institutions where policy is still ambiguous or absent.

To address this gap, we use a speculative design approach to uncover the ways that curators are currently employing AI, the places where AI is amplifying marginalized voices in archives, and the speculations on how future AI applications can inform the design of equitable approaches to integrating AI into curation best practices. We ask: (RQ1) How do modern curatorial institutions – e.g., academic libraries, archives, and museums (LAMS) – currently employ AI to advance equitable systems? and (RQ2) How do organizational structures enable co-design of equitable AI futures in information systems between curators, organizational leadership, and the user community?

In this study, we draw from semi-structured qualitative interviews across 9 research sites across 3 curation contexts to identify the current uses, near-term plans, and speculative future possibilities of AI with a focus on curators of open research data. The 3 sites are: (1) cultural heritage archives, (2) academic library collections, and (3) research data management repositories. Curatorial institutions are an excellent site to examine AI in practice because many are in the process of transitioning from legacy systems to AI technologies.

Findings show that curators of open research data in LAMs tend to follow the lead of similar organizations in adopting AI innovations, which results in a type of institutional isomorphism. At the same time, other curators act as agents of resistance to corporate AI by deciding how legacy systems should incorporate AI. Curators choose equitable approaches AI, e.g., for research data harmonization of indigenous and queer archival materials.

We argue curation necessitates a reconciliation of the local infrastructure and data cultures with software products that are designed in industry, with a lack of customizability for LAMs contexts. We demonstrate that current curatorial practices are guided by future visions of AI – what we call the “teleos of AI” because teleos is Greek for the origins and provenance of records management while simultaneously suggesting an ""ultimate"" (future) goal.

This study opens new directions about the ethics of AI in the context of data curation. Our contribution to the field is a principled approach to ethics of AI in data curation and methods for doing caring curation and to inform regulation and policy and develop a body of knowledge that can serve to support curators to act upon as best practices in AI in curatorial institutions.

Computational Poetry Collection Analysis via Context-Dependent Language Models

Kahyun Choi, Assistant Professor, Indiana University at Bloomington

ABSTRACT

Poetry gained significant attention recently. For example, after Amanda Gorman's reading of her poem at the presidential inauguration ceremony, the traffic to poet.org, one of major online poem digital libraries, has dramatically increased, 250% from the previous year. In addition, traffic on other poetry websites increased remarkably during the Covid19 pandemic because people used poems for therapeutic purposes (Acim, 2021). Thanks to the Internet and digitization, gigantic poem collections are available to researchers and poem lovers through local libraries and online poem digital libraries.

Meanwhile, poem digital libraries suffer from a lack of variety in metadata, which hinders user's access to diverse poem collections. For example, existing digital libraries provide limited recommendation and retrieval services based only on a handful of metadata types, e.g., authors, titles, or publication year, while more advanced metadata types, such as themes, are known to be potentially useful in information retrieval. There are some services where themes are available in the form of auxiliary information, but only for a small portion of the collection due to the inherent difficulty in understanding poems. Furthermore, even though annotations are available, they are not scalable and prone to be biased. An automated annotation mechanism, such as an AI-based natural language processing (NLP) algorithm, is not a viable option yet, as it also needs annotated dataset for training. The challenge stems from the fact that poetry is considered as the most challenging text for humans and machines to understand due to its figurative language and multiple layers of meanings. Thus, poetry has been rarely explored in computational analysis, unlike prose, such as news articles and product reviews. Moreover, it is uncertain if existing NLP technologies for other genres of text are applicable to poems.

To address the gap, this study proposes AI-based NLP systems that achieve a high-level understanding of poetry, such as its theme. This project answers the following questions: 1) Can advanced deep learning-based language processing models, such as BERT, decipher themes of poems? 2) Do auxiliary data, such as authors' notes on their poems, provide additional information to the poem analysis systems? Particularly, this study explores about 12,600 poems and their authors' notes on the poems on poets.org for the experiments. Our approach shows promising results that BERT can analyze meanings from poems to some degree, and the auxiliary data is complementary to poems in the learning process.

Knowledge Graph for Discovering Interdisciplinary Research Connections

Stanislava Gardasevic, PhD Candidate, Communication and Information Sciences Program; Teaching Assistant/Instructor at Library and Information Sciences University of Hawaii at Manoa

ABSTRACT

The focus of library professionals was always directed toward knowledge organization. After the promise of Semantic Web technologies to be the next big driving force of AI research, we jumped on that bandwagon and adjusted our practices and standards to answer that call. However, the data-driven computational approaches represented by machine and deep learning outweighed as dominant AI stream, eliciting better, quicker, and cheaper results than the human-made categorization and knowledge organization.

The disadvantage of AI systems is that they are often referred to as 'black boxes', and the unknown can be the cause of doubt and distrust in the processes and results obtained, especially in the library and information

profession where transparency and informed decision-making are among the dominant values. About 10 years ago, a technology called Knowledge Graph (KG) emerged as a possible middle-way solution, where both sides of the spectrum can meet. KGs can facilitate both ML/AI applications as well as the human modeling of a given domain of knowledge. KG can be described as a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities [1].

In this presentation, I am introducing a project where a community of interdisciplinary Ph.D. students was engaged in creating a KG model. Based on this model, the knowledge graph was created with the aim to organize, capture, and aggregate information relevant to this population and their academic progress, while supporting interdisciplinary research. This approach supports quick access to the information on the entities of students' interest (such as professors, fellow students, their research, publishing, course taking/teaching activities, etc.), but also allows for the transfer of tacit/experiential knowledge for the community in question.

The knowledge graph modeled in the proposed way is a basis for applying machine learning techniques (embeddings, network analytics approaches, natural language processing) that can support consequential recommendation systems; while at the same time allowing for the user's critical thinking and discernment based on the evidence of the information provenance- contextualizing potential AI systems that could be in place.

Knowledge graphs were already utilized for the purpose of storing and retrieving research data; however, in this approach, we propose a multiplex network through which people and their activities can be represented and queried via 12 affiliations (actual and latent). The added value of social network layers included in this knowledge graph is intended to show connections of actors based on their co-authorship, dissertation mentorship, and course-taking activities, next to their current and potential affiliations based on the shared research topics, research areas, and domain of interest, to name a few.

If implemented in an information system, this approach would allow academic librarians to get more involved in the research activities of particular domains/departments they are serving, while facilitating interdisciplinary collaboration- as this value is embedded in the design efforts of the model.

Footnote:
[1] Hogan, A. et al. (2020). Knowledge graphs: A comprehensive introduction. http://arxiv.org/abs/2003.02320

Images to integrated data: Digitizing and structuring historical records with deep learning

Sara Lafia, Postdoctoral Research Fellow, ICPSR, University of Michigan
David A. Bleckley, Data Project Manager, ICPSR, University of Michigan
J. Trent Alexander, Associate Director, ICPSR, University of Michigan

ABSTRACT

Many libraries and archives maintain paper-based collections of research documents – such as administrative records – which are valuable to research communities but have limited re-analysis potential as they are not accessible to a broader scholarly community. As more collections are digitized, there is an opportunity to leverage Artificial Intelligence to increase the usability of information extracted and structured from research documents. In this talk, we describe our experiences using digital scanning, optical character recognition, and deep learning to create a digital archive of administrative records related to the Servicemen's Readjustment

Act of 1944, also known as the G.I. Bill. We explain how we use document image analysis with deep learning to structure information extracted from these scanned documents as part of a larger effort to digitize, parse, and link historical records related to G.I. Bill mortgages. These records offer insights into veterans who benefited from loans, impacts on the communities built by the loans, and the institutions that implemented them. This project serves as an example of how artificial intelligence can complement the digitization of administrative records to produce a structured resource for researchers and the public.

AI & Co-design in public libraries: Empowering underserved youth to cultivate symbiotic relationships between Artificial Intelligence (AI) and their communities

Hee Rin Lee, Assistant Professor, Michigan State University
Kahyun Choi, Assistant Professor, Indiana University
Selin Akgun, PhD Student, Michigan State University
Ji Youn Shin, Assistant Professor, University of Minnesota
Pooja Malvi, Master's Student, Michigan State University
Meredith Dedema, PhD Student, Indiana University

ABSTRACT

Since the internet emerged in the mid-1990s, public libraries, as early adopters, have long played a critical role in enhancing technology literacy in the US (Bertot et al., 2008; Ito et al., 2013; Braun and Visser, 2017). As we enter an era of increased AI technology in our society, libraries have tremendous potential for nurturing AI literacy. Building on the role of public libraries as facilitators of digital literacy (Ito et al., 2013; Subramaniam et al., 2012), the goal of this project is to explore how library-based AI literacy programs can serve underserved youth communities.

As an exploratory project, we examine the role of the library as a community catalyst to enable economically underserved youth to 1) have access to core knowledge about AI, and 2) play an active role in designing AI technologies for their communities. This project follows a pedagogy built upon a critical race theory that views students from socially underserved communities not as people with deficits but as people with “community cultural wealth” (Yosso, 2005). By using participatory design methodology, 10–14-year-old students in this project utilize their own assets to co-design AI technologies for their community. The education program has two main modules: 1) Module 1 - Understanding core concepts of AI, and 2) Module 2 - Envisioning AI for local industries. These two modules are developed based on the initial interviews of students that examined students’ assets and existing AI knowledge.

Module 1 - Understanding core concepts of AI: We focus on two main concepts of AI, which are “classification” and “generation.” In this module, students learn what classification and generation are and perform hands-on activities utilizing those two concepts. This module also focuses on incorporating local community issues into the curriculum and encouraging the youth participants to be co-designers rather than passive learners. By adopting a co-design approach, our education materials deliberately include design activities that position participants as active learners and co-designers in shaping future AI technologies. For example, students develop their own AI prototypes using crafting materials, arduino, and AI education platforms.

Module 2 - Envisioning AI for local industries: This module helps students envision how AI can be utilized in their local industries. These materials include videos and photos of the actual workplaces. For example, in Michigan, we focus on manufacturing because of the prevalence of manufacturing industries, and thus jobs, in Michigan (and the US) (U.S. Bureau of Labor Statistics, 2020). We work with the local healthcare industry in

San Diego, CA. Students get a sense of the job tasks and work environments of these businesses through the information we collect. For example, we present videos of workers' interviews illustrating their workflows and issues that can be addressed by AI. This enables them to envision what types of AI systems would support these environments.

Historical Text Datafication and Loss: Computational Recovery of Typographical Layout Logic on an RDF Graph Featuring ML Methods

Huapu Liu, Doctoral Student, University of Alabama
Steven L. MacCall, Associate Professor, University of Alabama

ABSTRACT

We report on the development and testing of a computational pipeline that utilizes cluster-oriented machine learning algorithms to recover the logical relationships embedded in the typographical layout of book indexes, structures that are lost when standard page-level OCR scanning techniques are deployed on the index pages of digitized historical books. Recovering typographical logic would aid in the construction of overlay indexes using RDF graphs that can provide granular access to digital texts thus improving textual navigation.

Social Selection of Algorithms: The Unintended Consequences of Explainable AI

Alex Mayhew, PhD Candidate, University of Western Ontario

ABSTRACT

Increasingly algorithms are being used to govern complex decisions, such as criminal sentencing and insurance premiums. The increasing influence of algorithms has brought the question of algorithmic bias to prominent attention. If the data we generate to power the algorithms captures our prejudices, then it is little surprise that algorithms themselves reproduce those same prejudices. Worse still, at the moment most algorithms are black-boxes, leaving this bias hidden.

One potential response to this challenge is Explainable AI (XAI). XAI are algorithms that analyze other algorithms and explain their 'reasoning', exposing the hidden bias and enabling us to respond. While this is a promising approach, it poses its own challenges. Any XAI system would itself be an algorithm, subject to prejudiced data and biased outcomes. But there may be even more subtle and pernicious failure modes.

Like any software, XAI systems will increasingly exist as a population. Additionally, future versions of XAI systems will be preferentially based on particularly 'well performing' versions of XAIs under use in the previous generation. This creates an evolutionary environment where the selection of each generation is influenced by nebulous social measures, like user satisfaction or mollification. Cognitive Science has shown us that humans typically prefer coherence over truth. This could result in the XAI optimizing for what is convincing instead of what is true, without anyone intending such an outcome.

As computer system designers well know, the machine does what you tell it to do, not what you want it to do. In this case the 'telling' is not an intentional act. Compounding the problem, in this case it is possible to still generate results that are superficially acceptable to the stakeholders of the system. The evolutionary perspective can be helpful in framing and understanding some of the challenges surrounding XAI.

In crudest terms, we can imagine a scenario where successive generations of XAI are developed, each successive generation selected for an increased ability to communicate superficial coherence, rather than the accuracy of their model of the algorithm being explained. Absent awareness of this possibility, this malignant failure mode may be the default outcome.

Digital Deep Redlining

Arcadio Matos, PhD Student, School of Communication and Information, Rutgers University
Vivek K. Singh, Associate Professor, School of Communication and Information, Rutgers University

ABSTRACT

Digital redlining is a phenomenon where real world discrimination is replicated and amplified in digital spaces. We define digital deep redlining as a phenomenon where such discrimination is exacerbated due to the ever-increasing growth in data capture and deeper algorithms.

(1) Richer Data: Technological advances have caused more information resources to become mediated by immersive technologies, such as virtual reality (VR), capable of capturing 100-300 times the amount of data of traditional smartphones and wearables (Strachan, 2022).
(2) Deeper Algorithms: Emerging deep learning and other AI approaches have tremendous power at identifying trends which are not easily observable to humans or traditional algorithms. A combination of AI algorithms and immersive technologies threaten to erode privacy and compound inequalities. For instance, rich VR data and deep algorithms allow for humans to be identified with just five minutes of data (Strachan, 2022).

Challenges:
(1) Disproportionate impact of early backers: Information technology has the potential to concentrate power among the designers and early adopters. For instance, reinforcement learning in recommendation engines results in the early inputs and interaction behaviors to have an outsized impact on the algorithm. For instance, more people are likely to be recommended library books which are “liked” by the early users than the later ones. Often, by the time these algorithms are ready for mass consumption the early adopters move on to newer technologies resulting in a mismatch between who the technology is trained on and who utilizes it. In the recent past this has manifested in the form of bias in multiple information algorithms.
(2) Indirect identification: Loss of privacy has been an issue in digital information environments. However, with the emergence of web tracking, mobile tracking, and now VR tracking, the “digital fingerprints” identifying an individual occur in such varied forms that they are not easily interpreted by a lay-person or even traditional algorithms. For instance, de Montjoye et al. (2015) showed that locational data at just 4 unique points in credit card swipes may be sufficient to uniquely identify someone despite not including any traditional personally identifying information (PII).
(3) Lagging legislation: While there is clear legislation to prevent physical redlining, there is much less legal support to prevent digital redlining. Although aspects of digital media like IP-address and bandwidth are easily connected with traditional redlining concepts, indirect aspects which require sophisticated deep learning to relate to an individual’s socio-economic characteristics are rarely discussed. For instance, Nair et al. (2022) describes numerous socio-economic characteristics that can be derived from short VR session logs. Prevention of redlining based on such characteristics (e.g., time taken to render a VR object in one’s game) is needed today, i.e., before such characteristics becomes entrenched in the financial substrate of the emerging information eco-system.

References

Angwin, J., & Parris Jr., T. (2016). Facebook lets advertisers exclude users by race. ProPublica. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race

De Montjoye, Y. A., Radaelli, L., Singh, V. K., & Pentland, A. S. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. Science, 347(6221), 536-539.

Nair, V., Garrido, G. M., & Song, D. (2022). Exploring the unprecedented privacy risks of the metaverse. http://arxiv.org/abs/2207.13176

Shaheen, M. Y. (2021). AI in healthcare: medical and socio-economic benefits and challenges [Preprint]. https://doi.org/10.14293/S2199-1006.1.SOR-.PPRQNI1.v1

Strachan, M. (2022). Metaverse company to offer immortality through 'live forever' mode. Vice. https://www.vice.com/en/article/pkp47y/metaverse-company-to-offer-immortality-through-live-forever-mode"

Hypergraphing a Network of Inquiry, Search, and Retrieval

Alamir Novin, PhD Candidate, University of British Columbia

ABSTRACT

Currently, A.I. has difficulty framing relevant information in their environments like humans (i.e., the "framing problem") (Chow, 2013). To do so, A.I needs to know which network of questions are most central to frame a problem. Scholars suggest observing how humans frame situations via inquiry, but more robust methods are required to understand human inquiry (Fodor, 2008). This paper puts forward a novel method for measuring and analyzing how humans cognitive frame situations: Tracking people's syntax for inquiry (i.e., who, what, where, when, which, how and how much?) within a network – referred to as a Network-of-Inquiry (NOI) henceforth. An NOI allows researcher to understand how central each of the syntax for inquiry (i.e., who, what, where, when, which, how and how much?) are when different groups of people frame a situation.

To demonstrate, suppose researchers observe John Searle (1999) in his famous Chinese Room (where he does not understand Chinese). Searle is asked ""Where is 村?"" on a map. He replies ""The 村 is here."" To an external observer, the agent's Information Search Process (ISP) is completed using only the inquiry of "Where." However, questions other than "where" are more central to Searle, such as answering "what" (as in ""what is 村?"). In other words, even though Searle completed the task using "where," his need for "what" was not exhibited to the outside observer. With more complicated search tasks, researchers require more nuanced measurements to elicit this internal network of questions. Fortunately, unlike Searle's Chinese room, people can "think-aloud" their Network-of-Inquiry so that it can be graphed and measured for central questions. The NOI can also be visualized in a hypergraph. Hypergraphs requires minimal parameters (Wolfram, 2021) and can capture the complex dynamics of inquiry.

The remainder of the paper provides an experiment where the method of NOI was used to distinguish how two groups framed their ISPs differently:

Participants (N=40) were asked to complete a research task using a mock search engine while thinking-aloud.

Two groups of students were asked to search about cellphones on campus. The mock search engine retrieved the same search results for both groups (regardless of what they queried). However, at the start of the task one group was also informed that the university cared about the "health" of the students. The search results were then framed so that any results about "Health" appeared at the bottom. Similar to the example in Searle's Chinese room, the question was whether this frame affected the NOIs of the groups differently, so that one group would have the question of "where" play an important role, such as "Where is the relationship of cellphones and health"). Participants conducted their search task using a think-aloud and 4854 utterances were collected and graphed based on which inquiries were used (e.g., who what where when which how and how much).

For the first group without the framed task, the utterances were centralized around the question of "what." However, for the group of participants with the health framed task, the question of 'where' became more central. Intuitively, this is sound: Participants looking for results will make utterances involving the question "where" to guide their search. However, this shows how even though the output and frequency of questions remained similar (i.e., both groups still inquired "What" most often), the hypergraph revealed the centrality of the questions. The NOI provided a useful understanding and way of measuring how different frames influence people. It also detected bias in an agent's inquiry when other measurements could not (e.g., frequency). Subsequently, this method may assist researchers with understanding how framing problems emerge.

Mining scientific literature with Natural Language Processing to expand bibliometrics analysis

Gang Shao, Assistant Professor, Purdue Libraries and School of Information Studies, Purdue University
Joseph Eisenberg, Undergraduate Student, Purdue University

ABSTRACT

The number of scientific publications has grown exponentially in many research fields over the past 10-20 years. With the research on Natural Language Processing (NLP) as an example, the Scopus database recorded 11,385 NLP publications in 2021, which is four times as in 2011 (2,855 papers) and 18 times as in 2001 (642 papers). The rapidly growing publications make big data challenges for bibliometrics analysis and literature studies. Therefore, it is necessary to explore innovative approaches to conduct bibliometric analysis using large-scale text analysis. NLP is an active and emerging AI field, providing advanced methodologies and frameworks for mining text data at a large scale. The open science initiatives in NLP offer excellent opportunities to librarians and information science researchers to easily adopt established NLP pipelines and transform pre-trained models with customized datasets. In this presentation, we will briefly review the history of NLP and discuss two case studies using NLP and transformers for large-scale bibliometrics analysis and mining scientific articles in full text.