Aprendizagem de máquina aplicada a indexação automática de documentos digitais

III Seminário do Portal de Periódicos - CAPES

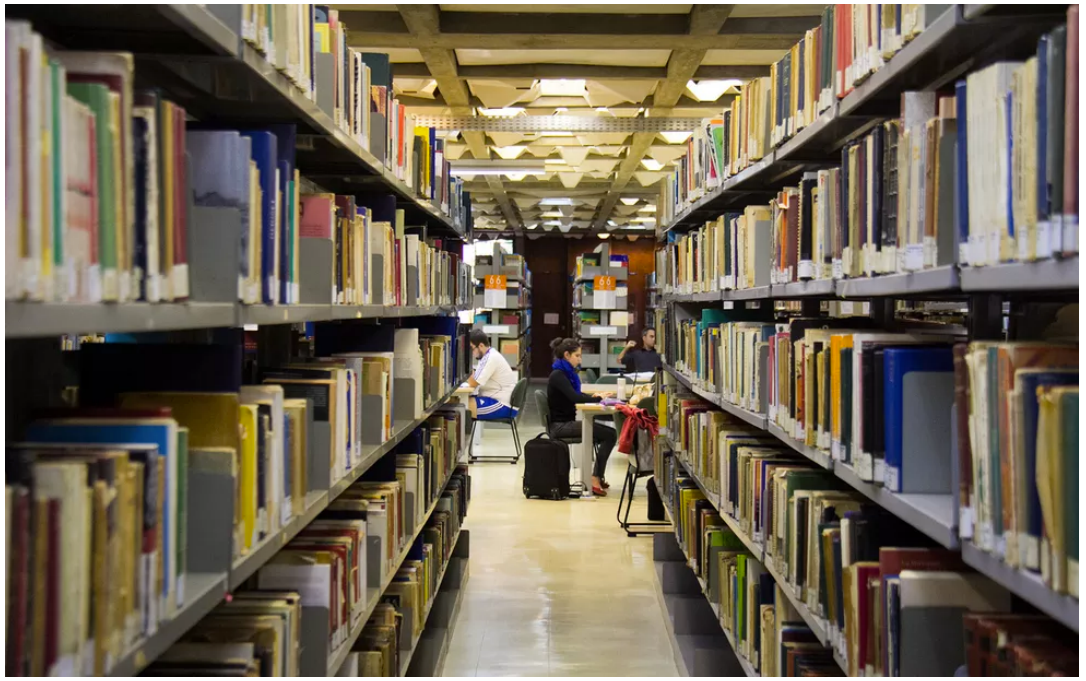
Dalton Martins
Faculdade de Ciência da Informação
Universidade de Brasília
daltonmartins@unb.br

UnB | FCI FACULDADE DE CIÊNCIA DA INFORMAÇÃO

UnB, sua linda
meu orgulho é você

Quem somos nós: institucionalidade

Universidade de Brasília
Faculdade de Ciência da Informação



Laboratório de Inteligência de Redes
Biblioteca Central

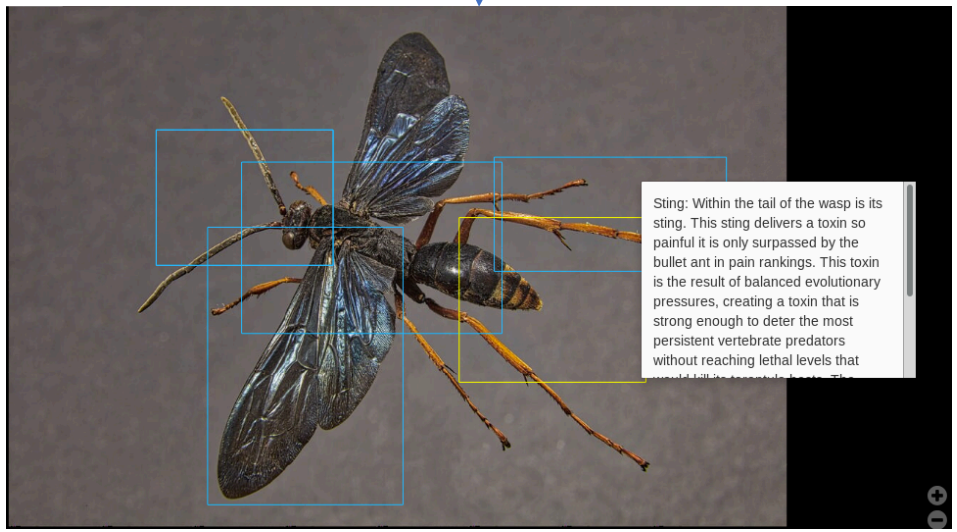
Quem somos nós: principais projetos

Tainacan

- **Pesquisa e desenvolvimento do Tainacan:** repositório digital para organização, gestão e difusão de acervos digitais em rede.

Quem somos nós: principais projetos

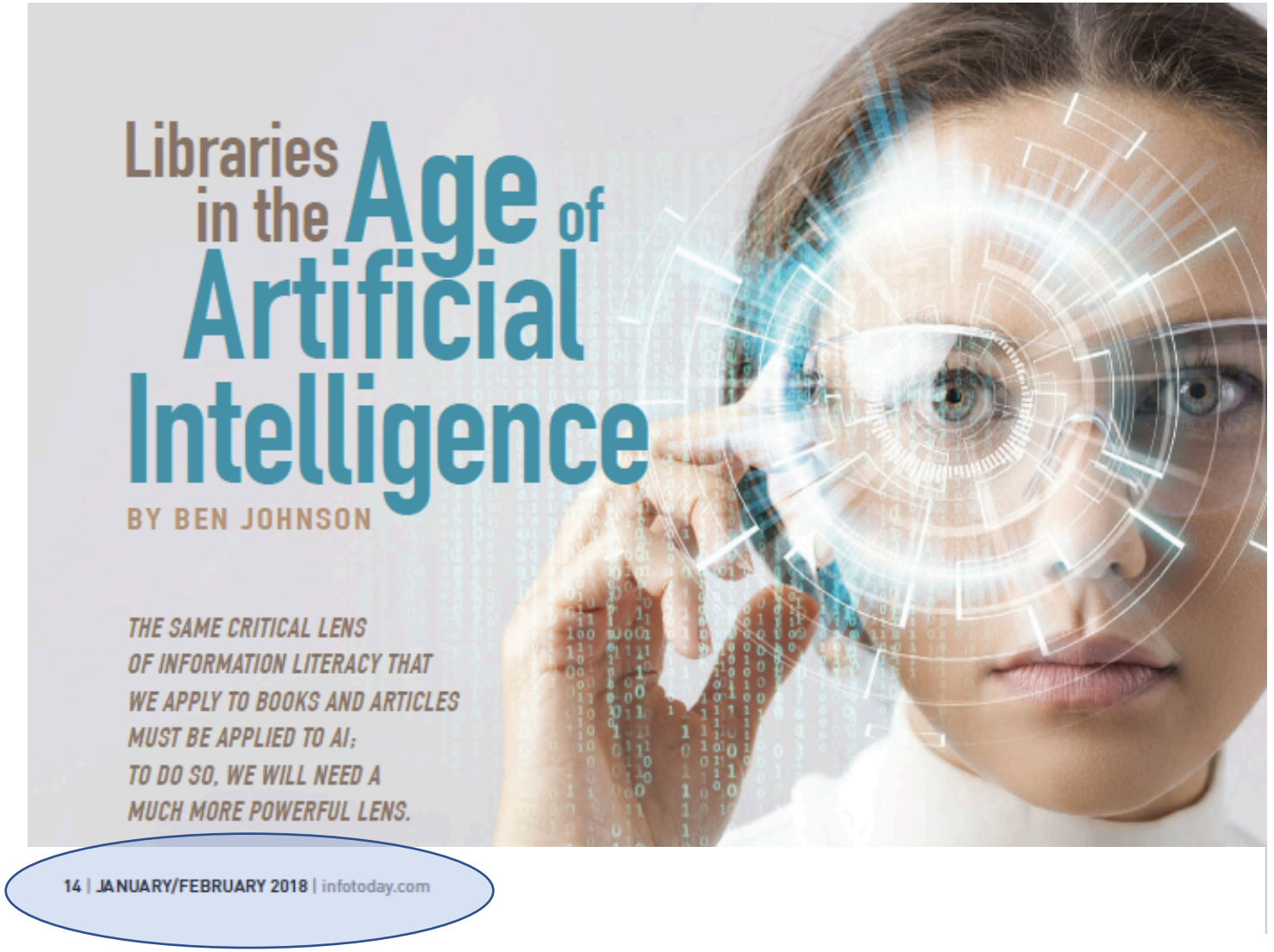
- Reconhecimento e reconciliação semântica de entidades em texto;
- Indexação automática e semi-automáica em imagens.



Pela primeira vez documentada no século XIII, Berlim foi sucessivamente a capital do Reino da Prússia (1701), do Império Alemão (1871-1918), da República de Weimar (1919-1932) e do Terceiro Reich (1933-1945). Depois da Segunda Guerra Mundial, a cidade foi dividida. Berlim Oriental se tornou a capital da República Democrática Alemã (RDA), enquanto Berlim Ocidental continuou sendo parte da República Federal da Alemanha (RFA).18 Com a reunificação alemã em 1990, a cidade passou a ser capital de toda a Alemanha.

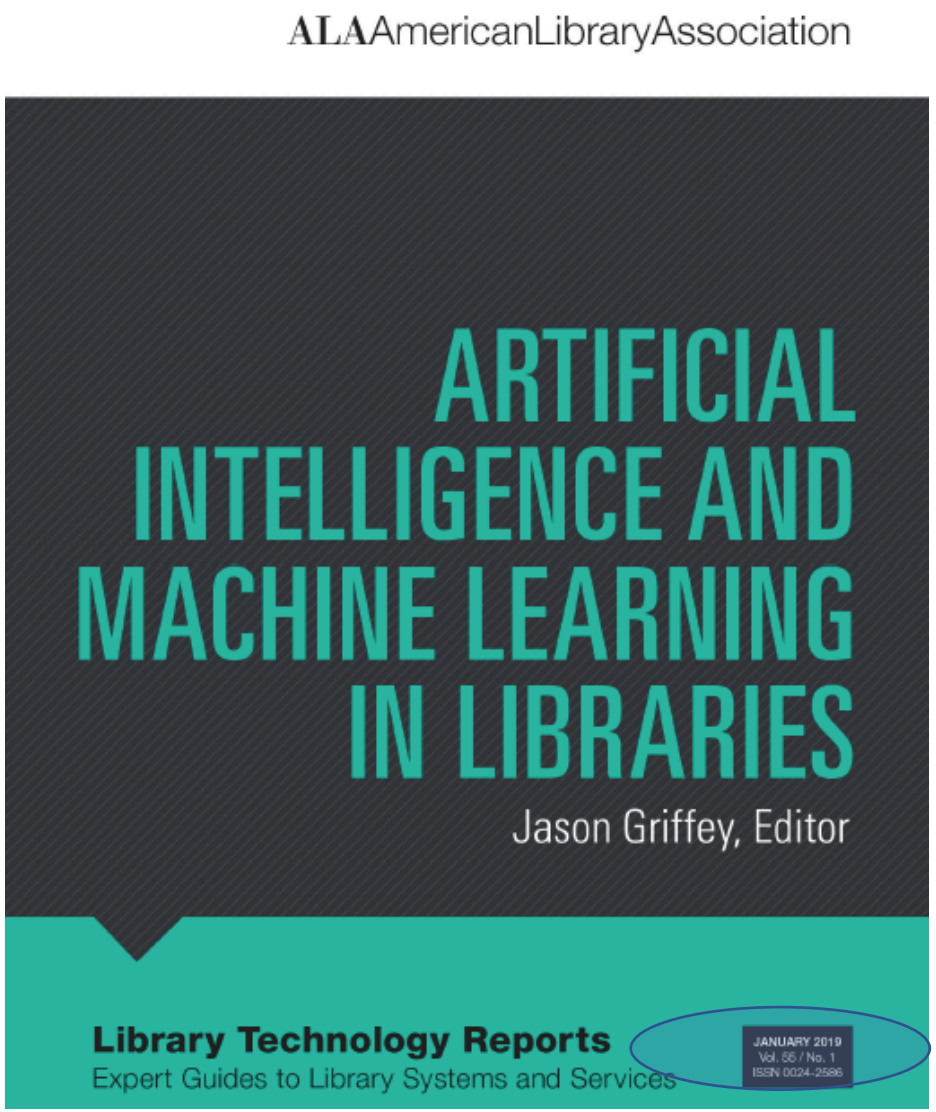
Pela primeira vez documentada no século XIII, Berlim foi sucessivamente a capital do Reino da Prússia (1701), do Império Alemão (1871-1918), da República de Weimar (1919-1932) e do Terceiro Reich (1933-1945). Depois da Segunda Guerra Mundial, a cidade foi dividida. Berlim Oriental se tornou a capital da República Democrática Alemã (RDA), enquanto Berlim Ocidental continuou sendo parte da República Federal da Alemanha (RFA).18 Com a reunificação alemã em 1990, a cidade passou a ser capital de toda a Alemanha.

Aprendizagem de máquina e inteligência artificial aplicadas às bibliotecas: um tema emergente



Revista InfoToday
2018

Aprendizagem de máquina e inteligência artificial aplicadas às bibliotecas: um tema emergente



American Library Association
2019
Caderno especial de tecnologia para as bibliotecas.

Aprendizagem de máquina e inteligência artificial aplicadas às bibliotecas: um tema emergente

from the PRESIDENT

Exploring AI

How libraries are starting to apply artificial intelligence in their work

Loida Garcia-Febo

I have recently started to hear more phrases such as, "I don't have to visit a library; I just ask Alexa [or Siri or Google Assistant] and it tells me everything I need to know. I speak to it all day."

The impact of even this early wave of artificial intelligence (AI)—including voice assistants and machine learning (ML)—is still uncertain in many fields, but it is time to include AI on our professional agenda and in our national conversation. In talking with librarians working in this area, it's clear that while AI can be useful, it also raises familiar concerns about privacy, intellectual freedom, authority, and access. And there are diversity considerations, as well, including access for people with different linguistic styles or abilities.

Fortunately, librarians are looking at AI from several perspectives. Some are using it to teach information literacy and critical-thinking skills to help patrons formulate questions for these devices and learn how to evaluate responses. University of Rhode Island, for example, is housing its collaborative efforts around AI in the library.

Cambridge (Mass.) Public Library (CPL) partnered with MIT Libraries and Harvard metaLAB to host the installation "Laughing Room," in which participants enter an artificially intelligent room that plays a laugh track whenever something is said that the room's algorithm deems funny. CPL Director Maria McCauley says this helped people to consider the impact of surveillance and AI on their lives. To further engage library users with big issues in science and technology shaping our society, the library will host a public dialogue about humor, culture, and AI with Harvard Law School's Cyberlaw Clinic this spring.

At MIT, Chris Bourg, director of libraries, is focusing on building a technical infrastructure so its collections are accessible by APIs and therefore can be used by machine-learning algorithms. MIT Libraries is working with AI/ML researchers at the university to analyze various library tasks and workflows that might be enhanced by AI. As Bourg says, it is important for academic libraries to make their collections accessible to AI tools like Alexa so that when someone asks a voice assistant for information, reputable scholarly literature is available. To make this successful, libraries will have to work to ensure scholarly information is openly accessible, not locked behind paywalls.

All this may be a lot of new information to process. But Catherine Nicole Coleman, digital research architect at Stanford Libraries in Palo Alto, California, has a good approach: Last year, Coleman conducted "Library AI Conversations" to help library workers familiarize themselves with the latest research and issues. She also worked mostly with bibliographers, archivists, and catalogers to explore the possibilities of AI for metadata and collection development. Additionally, they are collaborating with computer science faculty and faculty in the humanities and social sciences to explore human–machine collaboration, interaction, and interface (bit.ly/stanfordAI).

At ALA, we have resources to help library workers understand AI, these new devices, and the role of libraries. The Center for the Future of Libraries has written about voice-control devices (bit.ly/CFLvoice); the January issue of *Library Technology Reports* (bit.ly/LibTechAI) explores AI and ML; and many of our conferences—including the Library and Information Technology Association's forum and the Association of College and Research Libraries national conference—include sessions on AI.

My fellow library workers, the future of libraries will continue to be about the communities we serve. Librarians and library professionals will need to be at the forefront to support communities as these technologies transform our world. Let's continue the conversation and learn together. AL

Library professionals will need to continue to be at the forefront to support communities as emerging technologies transform our world.

LOIDA GARCIA-FEBO is an international library consultant.

American Library Magazine
2019
Editorial especial

Aprendizagem de máquina e inteligência artificial aplicadas às bibliotecas: um tema emergente

PERSPECTIVES

AI and Machine Learning

The challenges of artificial intelligence in libraries

dispatches

by Jason Griffey

Artificial intelligence (AI) and machine learning are everywhere, giving driving directions and identifying objects in photographs. They are so engrained in our technology that often people don't realize what they're experiencing is a machine learning system. Everyone with a smartphone has an AI system that uses machine learning.

For example, Google's Android operating system records, measures, and collects information and sends that data to servers. These servers use billions of data points collected from tens of millions of users as input for their machine learning systems. When you ask an Android phone to show you photos from the beach, a complex set of data moves back and forth between your phone and Google's servers, comparing your photos to the billions in its data set. The search results include pictures that the AI decided were most likely to be related.

Since Google has billions of photos to assess and millions of people helping it train its AI, the decisions that the AI makes are generally good. But AI is only as effective as its training data and the weighting given to the system as it learns to make decisions. If the data is biased, contains bad examples of decision making, or is simply collected in a way that doesn't represent the full problem set, the system will produce broken, unrepresentative, or bad outputs.

Apple, on the other hand, has chosen to model its AI and machine learning by analyzing and weighting your data locally on the iOS devices themselves. Your devices use the same machine learning algorithms to include your photos in Apple's preset weights, but they aren't pushed to Apple's servers. Because each data set is analyzed locally, there is no shared decision making as there is with Google. Each device must do heavy lifting itself, rather than rely on remote servers for the bulk of the work.

For data privacy and security concerns, localized machine learning has an advantage. If you don't need to send photos and data back and forth from server to client, and if providers don't need to store and host data, the data's vulnerability to attack is greatly reduced.

The examples above focus on object and image recognition in photos by a machine learning system. This is only one of dozens of uses for AI and machine learning systems.

It's also easy to see how an AI system is useful for libraries and archives in creating metadata from digitization projects. AI systems can be trained to recognize locations from a single photograph—including where the photographer was standing—based on angle, geography, and other factors. These systems can be enormously useful in making the processing and cataloging of archives and collections more discoverable.

As more libraries and library vendors move into developing AI and machine learning systems, we should be sensitive to the privacy implications of collecting and storing the data that's needed to train and update those systems. As with existing systems where we outsource data collection and retention to vendors, libraries need to be aware of the mechanisms by which that data is processed and how it may be shared with others through training sets. Where libraries can provide local analysis in the style of Apple and iOS, they should.

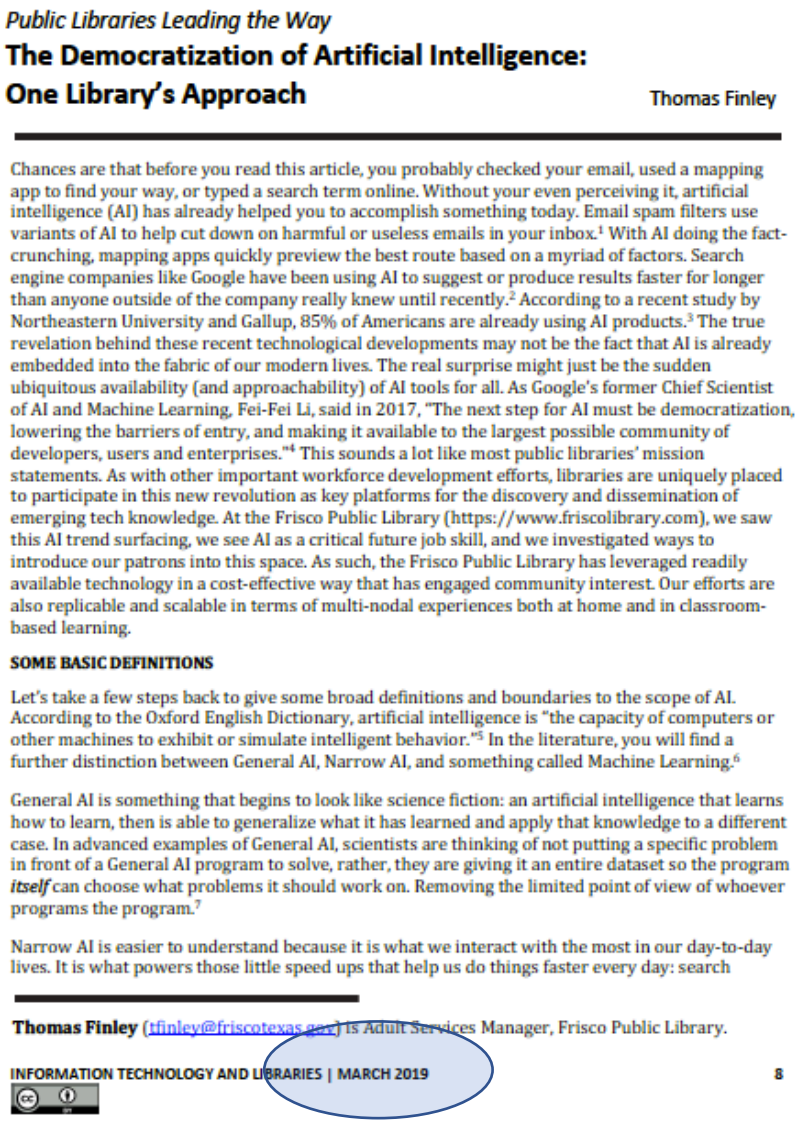
The opportunities associated with new machine learning systems to reform large portions of library activities will be rich and varied. While it will be some time before AI will conduct full conversations or reference interviews with students and patrons, the use of AI as an increasingly powerful lever inside other systems will progress quickly over the next three to five years. Libraries can watch these systems as they develop, work with vendors, and create their own services and systems so that our values and ethics are baked into the technology at the outset.

For data privacy and security concerns, localized machine learning has an advantage.

JASON GRIFFEY is a librarian and technologist and the founder and principal at Evenly Distributed. Adapted from "Artificial Intelligence and Machine Learning in Libraries," *Library Technology Reports* vol. 55, no. 1 (Jan. 2019).

American Library Magazine
2019
Análise de tendências

Aprendizagem de máquina e inteligência artificial aplicadas às bibliotecas: um tema emergente



Public Libraries Leading the Way

The Democratization of Artificial Intelligence: One Library's Approach

Thomas Finley

Chances are that before you read this article, you probably checked your email, used a mapping app to find your way, or typed a search term online. Without your even perceiving it, artificial intelligence (AI) has already helped you to accomplish something today. Email spam filters use variants of AI to help cut down on harmful or useless emails in your inbox.[1] With AI doing the fact-crunching, mapping apps quickly preview the best route based on a myriad of factors. Search engine companies like Google have been using AI to suggest or produce results faster for longer than anyone outside of the company really knew until recently.[2] According to a recent study by Northeastern University and Gallup, 85% of Americans are already using AI products.[3] The true revelation behind these recent technological developments may not be the fact that AI is already embedded into the fabric of our modern lives. The real surprise might just be the sudden ubiquitous availability (and approachability) of AI tools for all. As Google's former Chief Scientist of AI and Machine Learning, Fei-Fei Li, said in 2017, "The next step for AI must be democratization, lowering the barriers of entry, and making it available to the largest possible community of developers, users and enterprises."[4] This sounds a lot like most public libraries' mission statements. As with other important workforce development efforts, libraries are uniquely placed to participate in this new revolution as key platforms for the discovery and dissemination of emerging tech knowledge. At the Frisco Public Library (https://www.friscolibrary.com), we saw this AI trend surfacing, we see AI as a critical future job skill, and we investigated ways to introduce our patrons into this space. As such, the Frisco Public Library has leveraged readily available technology in a cost-effective way that has engaged community interest. Our efforts are also replicable and scalable in terms of multi-nodal experiences both at home and in classroom-based learning.

SOME BASIC DEFINITIONS

Let's take a few steps back to give some broad definitions and boundaries to the scope of AI. According to the Oxford English Dictionary, artificial intelligence is "the capacity of computers or other machines to exhibit or simulate intelligent behavior."[5] In the literature, you will find a further distinction between General AI, Narrow AI, and something called Machine Learning.[6]

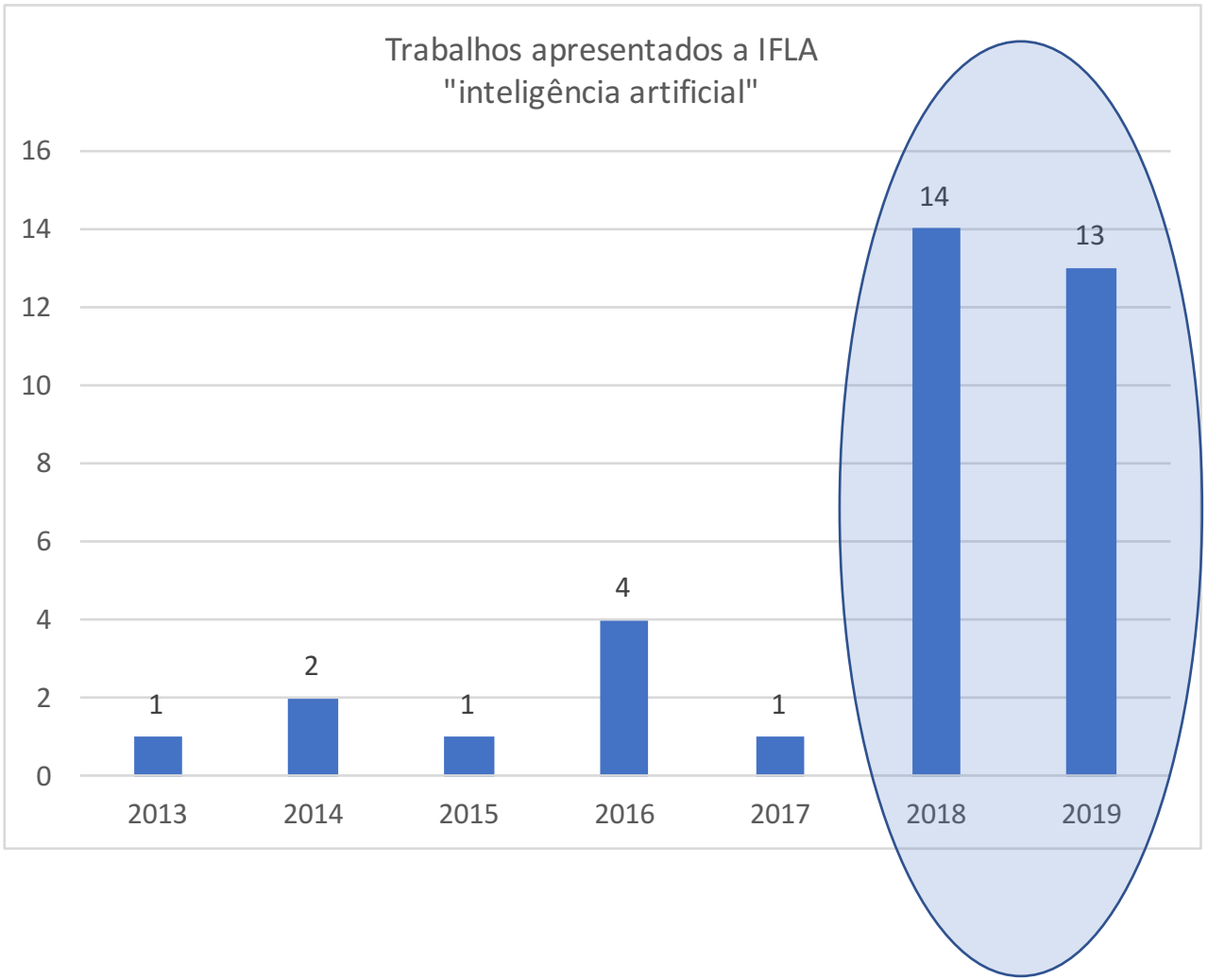
General AI is something that begins to look like science fiction: an artificial intelligence that learns how to learn, then is able to generalize what it has learned and apply that knowledge to a different case. In advanced examples of General AI, scientists are thinking of not putting a specific problem in front of a General AI program to solve, rather, they are giving it an entire dataset so the program ***itself*** can choose what problems it should work on. Removing the limited point of view of whoever programs the program.[7]

Narrow AI is easier to understand because it is what we interact with the most in our day-to-day lives. It is what powers those little speed ups that help us do things faster every day: search

Thomas Finley (tfinley@friscotexas.gov) is Adult Services Manager, Frisco Public Library.

Information Technology and Libraries
2019
Revista científica

Aprendizagem de máquina e inteligência artificial aplicadas às bibliotecas: um tema emergente



Trabalhos apresentados nos congressos da IFLA
Descritor: ”inteligência artificial”

Há um crescente e importante interesse nos temas aprendizagem de máquina e inteligência artificial no campo das bibliotecas nos últimos anos.

O que isso significa em termos de oportunidades de novos serviços, melhoria na qualidade das instituições e necessidade de qualificação profissional?

O que os bibliotecários estão pensando sobre isso?

A percepção dos bibliotecários sobre a aprendizagem de máquina e IA

LIBRARIANS' PERCEPTIONS AND ITS POTENTIAL

The subject of artificial intelligence (AI) is being discussed everywhere in the media. Stephen Hawking, Elon Musk, and Bill Gates regularly sound the alarm about AI as an existential threat to humankind. Open a newspaper, turn on the television, or log on to the internet, and you will find a plethora of information and opinions on AI and its potential impact on human endeavors. In addition to being a hot topic in the media, the scholarly literature in medicine and law is replete with AI research. It acknowledges AI as a transformative, if not disruptive, game changer. AI is being used today in the practice of law—in areas of contract review, billing, and jury selection. In the field of medicine, AI's ability to crunch massive datasets has allowed it to surpass humans in diagnostic capabilities. Educators in law and medicine have also acknowledged that AI is changing the way professionals are trained, and it will ultimately reduce the number of lawyers and doctors needed in the workforce of the future.

That said, the topic of AI is not everywhere—it's not in the library literature. Oddly, for a profession that has done more than its share of coping with disruptive technologies over the years, we librarians are not in any meaningful way discussing AI as compared to those in other professions. We have not developed any substantial research on this topic nor have our library schools acknowledged the possibility of reducing admissions. Capability for machine learning, natural language processing, and massive computing power are the three aspects of AI that impact the professions of law and medicine. These will likewise impact the profession of librarianship.

We were curious: Why is there a paucity of discussion about AI in our professional literature? In order to understand this, we surveyed our colleagues on their perception of AI, specifically as it relates to the future, the potential impact on our work, and the numbers in the workforce.

Methodology

For the purpose of the survey, we settled on IBM's Watson as our representation of AI. Watson is a fairly well-known product, along with similar AI-based systems, such as Amazon's Alexa and Apple's Siri. We felt that Watson could be easily conceptualized by our respondents, because it, like librarians, answers questions. Unlike librarians, Watson relies on machine learning and massive computing power to do so.

We received permission to model our survey on the one used by AI experts Müller and Bostrom (2016). Our survey, "Librarian Perception of Artificial Intelligence," was created using Qualtrics and received institutional review board approval. The instrument was distributed to professional library listservs during May and June 2017. The survey had 10 questions. Numerically, the responses ranged from 307 to 341.

Chart 1 shows that 56.3% of our respondents feel AI will have a transformative effect on librarianship. Surprisingly, 43.69% of our respondents—who are working professionals in a field that requires high levels of technological proficiency—feel AI will have little or no effect on librarianship. This is a departure from predictions (albeit limited) found in our library literature: "[T]he question is not so much what technology will be affected, but rather what technology, if any, will remain unaffected by AI" (Fernandez 2016). In addition, both the American Bar Association and the American Medical Association have recognized AI in their literature and at their annual meetings. In a 2016 roundtable discussion at the American Bar Association's annual meeting, it was concluded that "it is wise to embrace [AI] now so that it can be a tool as opposed to an impediment."

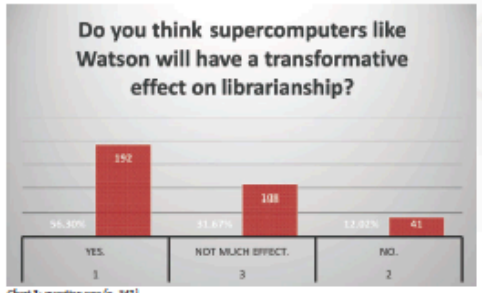


Chart 1: question one (n=341)

Pesquisa realizada durante os meses de maio e junho de 2017 via listas de emails de profissionais bibliotecários nos EUA.

Publicada na revista InfoToday em fevereiro de 2018

A percepção dos bibliotecários sobre a aprendizagem de máquina e IA

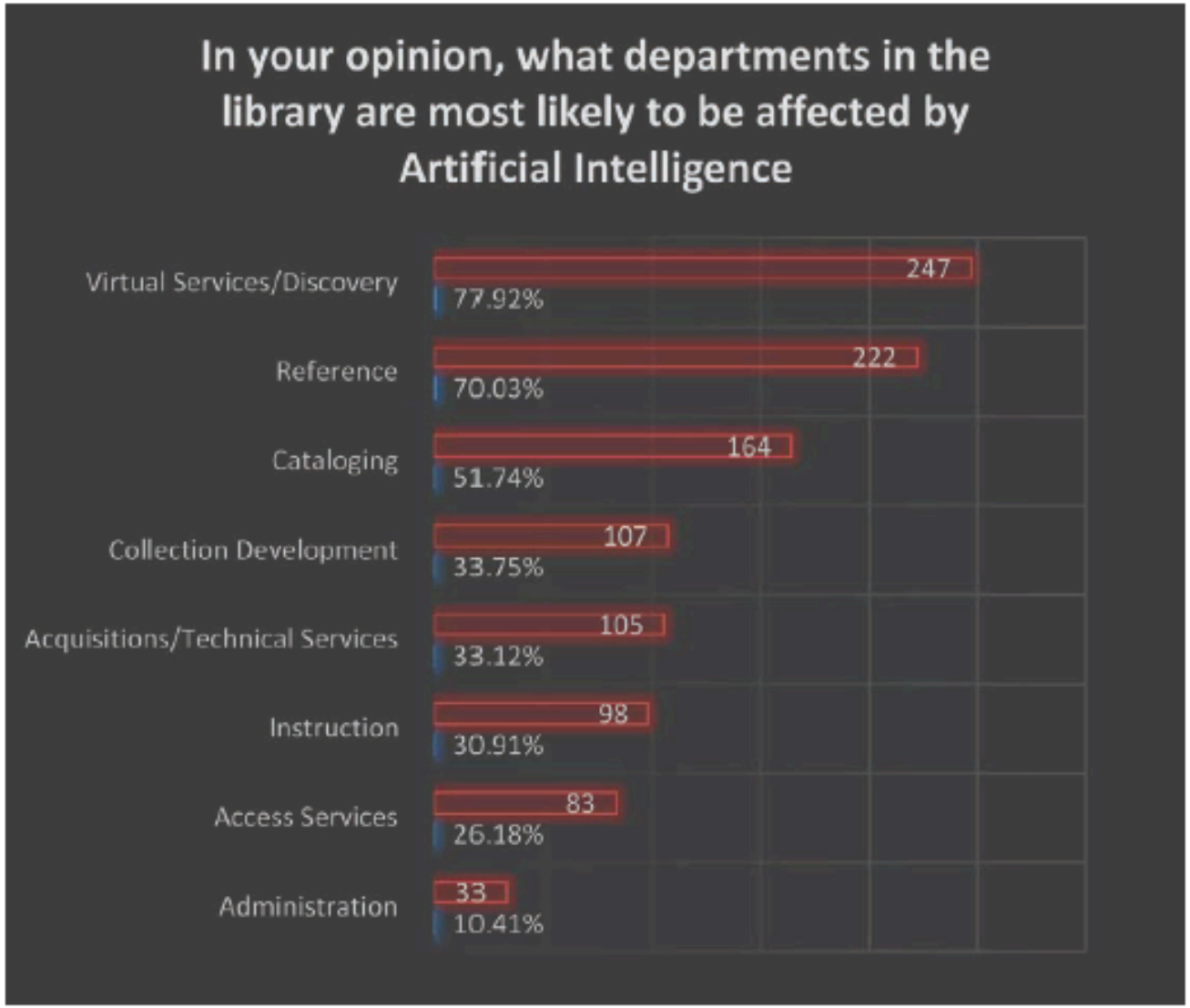


Chart 2: question two (n=317)

A percepção dos bibliotecários sobre a aprendizagem de máquina e IA

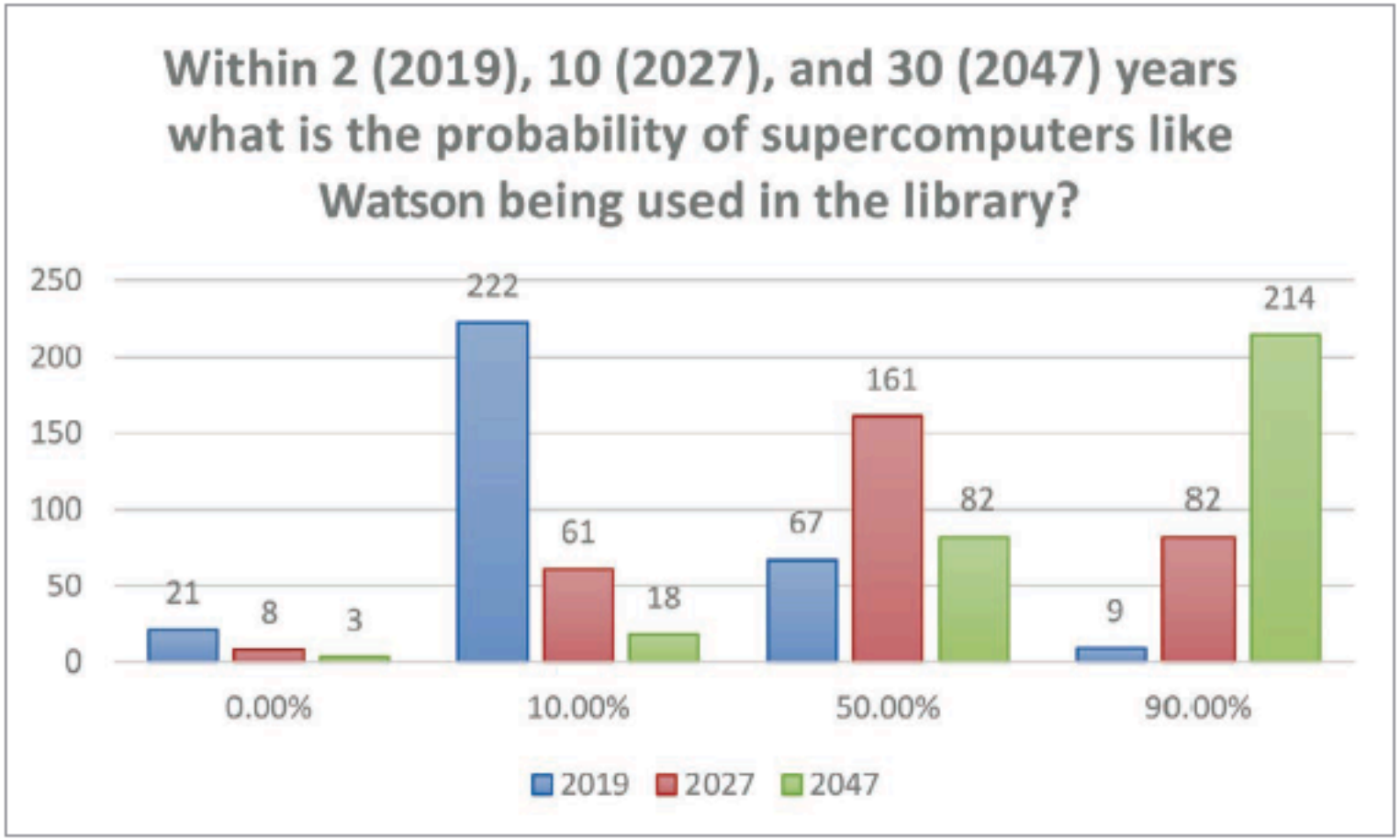
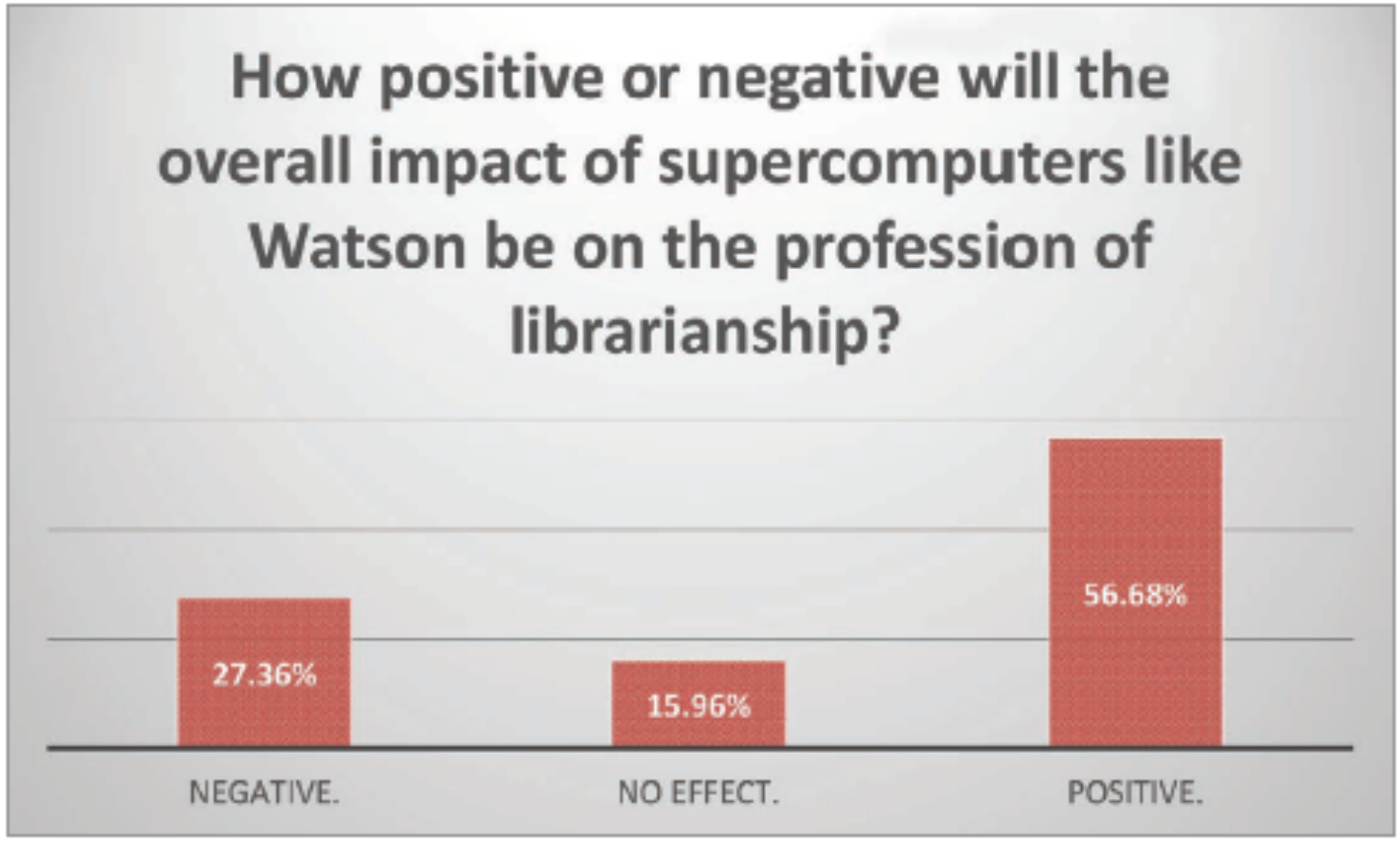


Chart 3: questions three (n=319), four (n=318), and five (n=317)

A percepção dos bibliotecários sobre a aprendizagem de máquina e IA



Charts 4, 5, and 6: questions six (n=315), seven (n=315), and eight (n=307)

A percepção dos bibliotecários sobre a aprendizagem de máquina e IA

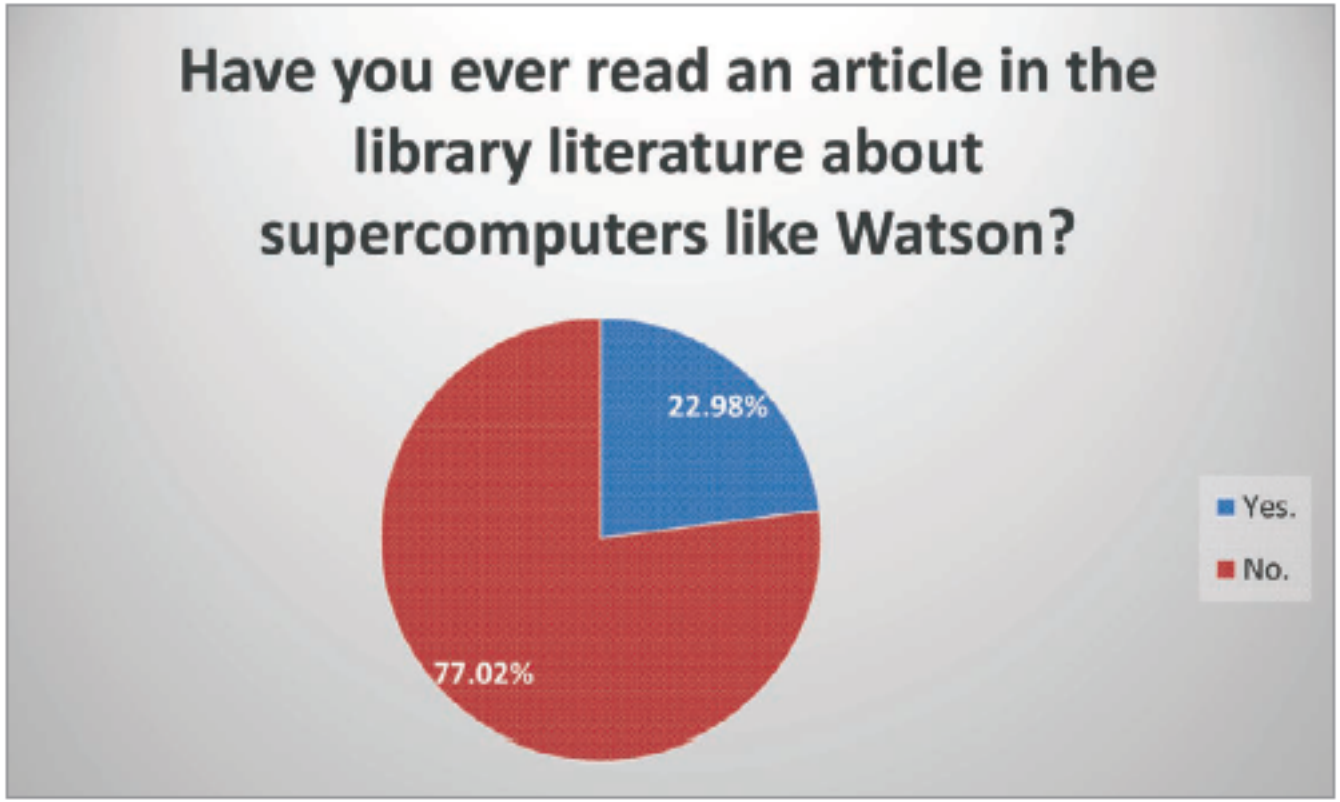


Chart 7: question nine (n=309)

A percepção dos bibliotecários sobre a aprendizagem de máquina e IA

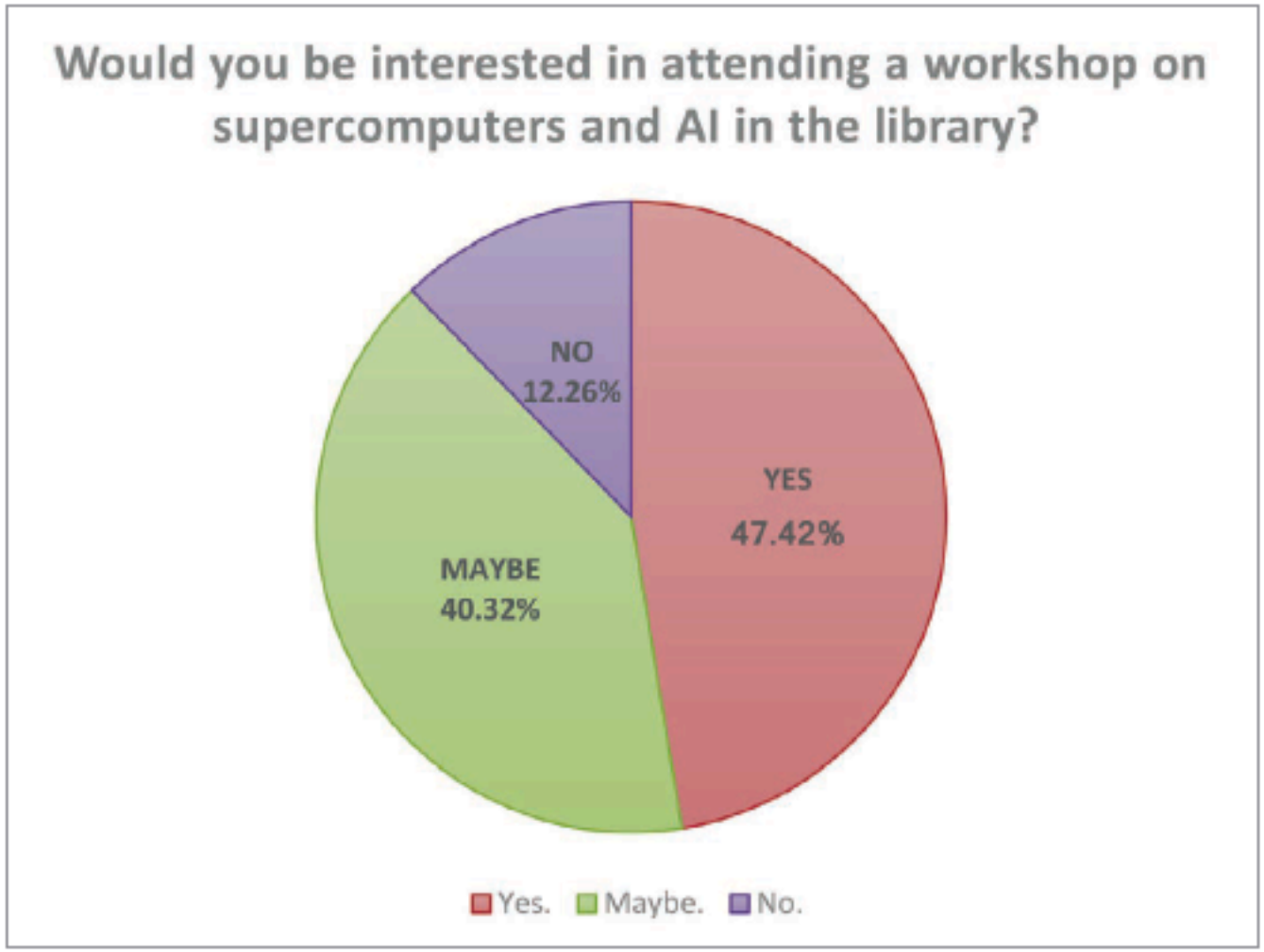
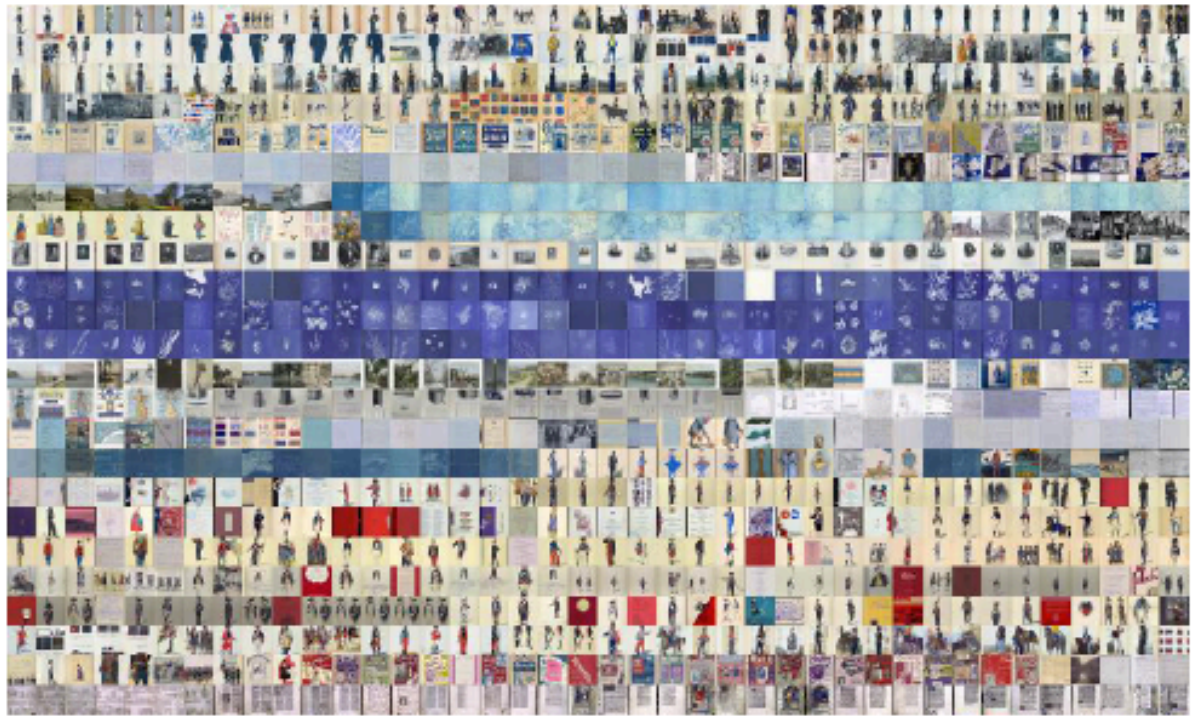


Chart 8: question 10 (n=310)

Como esse cenário impacta as bibliotecas?

Dados de qualidade



A grande riqueza dos dados das bibliotecas está em suas coleções, na catalogação, na indexação, na relação com o uso da informação!

São insumos fundamentais para a geração de novos serviços e produtos informacionais!

Novos serviços que usem esses dados são um próximo estágio fundamental ao desenvolvimento das bibliotecas.

Tipos de serviços
(mais rapidamente impactados pela IA)

- Referência
- Catalogação e indexação automática ou semi-automática
- Busca e recuperação da informação
- Aquisição
- Processamento de linguagem natural, imagem, áudio e vídeo
- Reconhecimento de padrões e redes semânticas de relacionamento entre documentos.

Potencial de aplicação da IA em Bibliotecas

Article:
Cox, A.M. orcid.org/0000-0002-2587-245X, Pinfield, S. and Rutter, S. (2018) The intelligent library: Thought leaders' views on the likely impact of artificial intelligence on academic libraries. Library Hi Tech. ISSN 0737-8831

https://doi.org/10.1108/LHT-08-2018-0105

Potencial de aplicação da IA em Bibliotecas

Library roles in AI	Competencies that need to be extended	Alternative providers of service/ function
Procuring content for AI to work from (including both licensing and through open access)	Procurement and licensing of e-content	Publishers and other new intermediaries
Providing content	Digitisation, metadata provision	Publishers and other new intermediaries
Data quality control	Collection management	
Procuring AI tools	Procurement and licensing of software and services	IT departments, academic departments
Data curation (e.g. of derived data)	Collection management, digital preservation	Publishers and other new intermediaries
Designing data infrastructure to enable AI	Design of information discovery infrastructure	IT departments
Explaining how to navigate the new information environment	Understanding of the scholarly publishing landscape, including data creation processes	
Teaching critical data literacy: understanding how to evaluate AI tools and their results, and also protect one's own privacy	Information literacy	IT departments
Designing AI tools	N/A – outside normal library professional work	Academic departments, Publishers
Data analysis and writing algorithms	N/A	IT departments, academic departments

Table 1 Potential library roles in AI

Impactos em potencial para as bibliotecas

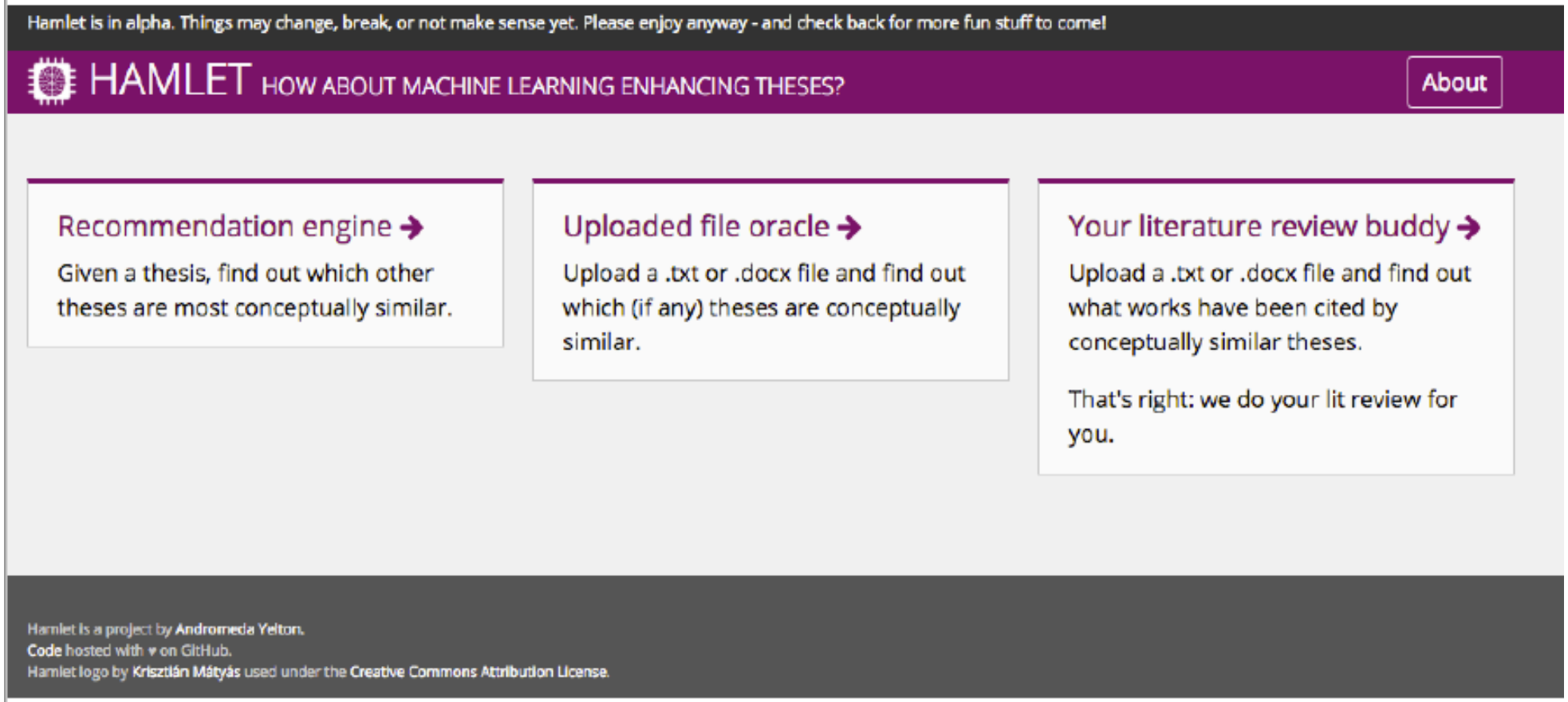
Others can only be guessed at. Yet bringing our data together with the literature, it emerges that AI should be seen as defining a nexus of change (Pinfield et al., 2017) that has "wide and deep" ramifications (JISC reference) in terms of:

1. What a library is, what a collection is and how to search for material. The library may increasingly be seen as data, accessed through AI, the scope of the collection as framed by the AI;
2. How established services are delivered, for example by chatbots and other intelligent agents;
3. What users expect of libraries: through expectations learned in other areas;
4. What information literacy is: the ability to navigate a new space of AI tools and data, and data literacies, including critical awareness of how to protect one's own privacy;
5. Who users are: some users will be AI tools; human access to content will be remediated through content being summarised and partially analysed for them by machines;
6. What libraries know about users and so how the library is managed: because of management decisions based on use data, combined with other learning and research analytics;
7. How the library works with other internal and external partners and competitors, especially IT services and new third-party commercial services;
8. How library services are evaluated: again through wider and deeper data;
9. What skills librarians need: be that for licensing, evaluation of data analysis and visualisation tools or using such tools themselves;
10. Whether the library community can operate effectively at different levels beyond the institution: in order to design and deliver services which will serve international communities of scholars and students;
11. Indeed, whether we need librarians (because of chatbots, automated metadata creation tools etc) or libraries (because of alternative intermediaries) at all, at least as currently conceived.

Exemplos e aplicações

http://www.akiwi.eu/

Alguns serviços (exemplos)



https://hamlet.andromedayelton.com/

Alguns serviços (exemplos)

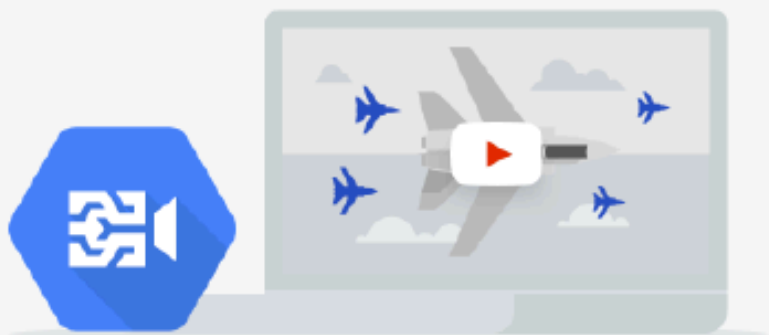
API CLOUD VIDEO INTELLIGENCE

Pesquise e descubra seu conteúdo de mídia com a API Cloud Video Intelligence

FAÇA UMA AVALIAÇÃO GRATUITA

Análise precisa de vídeo

Com a API Google Cloud Video Intelligence, a pesquisa e a descoberta de vídeos ficaram ainda melhores ao coletar metadados com uma API REST muito fácil de usar. Agora é possível pesquisar cada momento de cada arquivo de vídeo no seu catálogo. Com essa API, você pode fazer anotações em vídeos armazenados no Google Cloud Storage



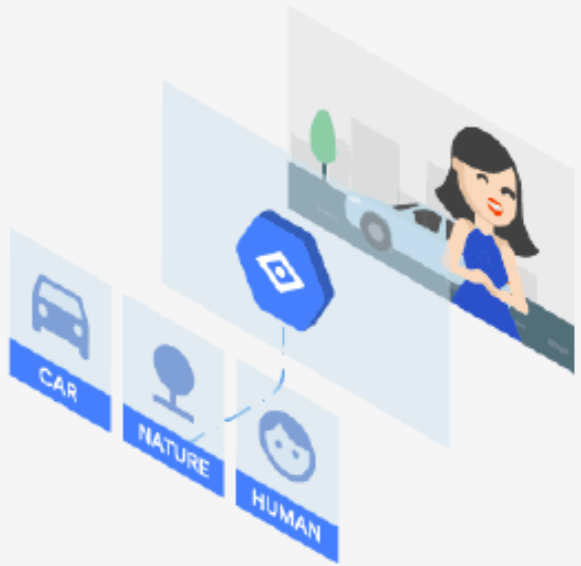
https://cloud.google.com/video-intelligence/

Alguns serviços (exemplos)

Análise avançada de imagens

O Cloud Vision oferece modelos pré-treinados por meio de uma API ou a capacidade de criar modelos personalizados usando o AutoML Vision. Assim, você tem a flexibilidade que precisar, dependendo do seu caso de uso.

A **API do Cloud Vision** encapsula modelos avançados de machine learning em uma API REST fácil de usar, o que permite aos desenvolvedores entender o conteúdo de imagens. Essa API rapidamente classifica as imagens em milhares de categorias, (por exemplo: "veleiro"), detecta objetos e rostos individuais e extrai palavras impressas contidas nas imagens. Crie metadados no seu catálogo de imagens, modere conteúdo ofensivo ou ative novos cenários de marketing usando a análise de sentimento das imagens.



https://cloud.google.com/vision/

Uma agenda de P&D para o desenvolvimento da IA nas bibliotecas

- Formação:
  - Incorporar disciplinas e projetos de pesquisa na formação de graduação dos bibliotecários;
  - Capacitar profissionais atuando na área em cursos de média e longa duração, como especialização;
- Desenvolvimento:
  - Diagnosticar conjuntos de dados que podem ser abertos para novos serviços pelas bibliotecas;
  - Implementar descrição semântica dos dados para melhorar qualidade descritiva e facilitar consumo por algoritmos computacionais;
  - Selecionar e priorizar serviços de maior interesse para a geração de novos produtos informacionais automatizando o acesso aos dados;
- Pesquisa:
  - Incorporar algoritmos de aprendizagem de máquina na análise e tratamento de dados oriundos das bibliotecas;
  - Propor novos modelos conceituais para automatizar serviços de busca, recuperação, indexação, catalogação e descoberta de novos conhecimentos nos conjuntos de dados.
  - Realizar projetos experimentais com as instituições de interesse.

Obrigado!

daltonmartins@unb.br