Análise de dados bibliográficos: estudo de caso no Catálogo Coletivo Nacional de Publicações Seriadas (CCN)

Bibliographic data analysis: case study in the National Collective Catalog of Serial Publications

Análisis de datos bibliográficos: estudio de caso en el Catálogo Colectivo Nacional de Publicaciones Seriadas

Bruno Carlos da Cunha Costa
Doutor, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brasil
Professor, Instituto Federal de Educação, Ciência e Tecnologia (IFRJ), Rio de Janeiro, RJ, Brasil
http://lattes.cnpq.br/8162384236816488

Ana Carolina Simionato Arakaki
Doutora, Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, Brasil.
Coordenadora de Serviços Bibliográficos, Instituto Brasileiro de Informação em Ciência e Tecnologia (COBIB/IBICT), Brasília, DF, Brasil
http://lattes.cnpq.br/9896600626524397
https://orcid.org/0000-0002-0140-9110

João Gabriel Grandotto Viana
Tecnólogo, Instituto Federal de Goiás (IFG), Goiânia, GO, Brasil
Bolsista, Instituto Brasileiro de Informação em Ciência e Tecnologia (COBIB/IBICT), Brasília, DF, Brasil
http://lattes.cnpq.br/5724937926178194

Gabriélle Helpis dos Santos
Graduação em andamento em Tecnologia em Ciência de Dados, Universidade Federal de Mato Grosso do Sul (UFMS), Campo Grande, MS, Brasil
Bolsista, Instituto Brasileiro de Informação em Ciência e Tecnologia (COBIB/IBICT), Brasília, DF, Brasil
http://lattes.cnpq.br/5724937926178194

Renan Barbosa dos Santos
Graduação em andamento em Matemática, Universidade de Brasília, UnB, Brasília, DF, Brasil
Bolsista, Instituto Brasileiro de Informação em Ciência e Tecnologia (COBIB/IBICT), Brasília, DF, Brasil
http://lattes.cnpq.br/1426164308581850

Renan Luiz da Silva Nascimento
Graduação em andamento em Estatística, Universidade de Brasília, UnB, Brasília, DF, Brasil
Bolsista, Instituto Brasileiro de Informação em Ciência e Tecnologia (COBIB/IBICT), Brasília, DF, Brasil
http://lattes.cnpq.br/9394350252301981

Anais do **VII Workshop de Informação, Dados e Tecnologia**, Porto Velho, 25-27 jun. 2024.
https://doi.org/10.22477/vii.widat.206

Resumo

Introdução: O Catálogo Coletivo Nacional de Publicações Seriadas (CCN) é uma base que reúne uma vasta variedade de publicações de bibliotecas de instituições acadêmicas e de pesquisa em todo o Brasil. O objetivo é explorar a convergência entre Biblioteconomia e Ciência de Dados, de maneira aplicada, destacando as extrações da base do CCN após sua modernização e reestruturação, isto é, demonstrando os modelos preditivos e componentes de inteligência artificial generativa. **Metodologia:** Utilizando uma abordagem empírica baseada no CCN, identifica-se a distribuição das publicações por conteúdo, ano e geografia, fornecendo insights sobre a pesquisa no Brasil. **Resultados:** Modelos preditivos foram desenvolvidos para estimar o volume de publicações, enquanto a exploração de componentes de Inteligência Artificial Generativa visa melhorar a interação com o CCN. **Conclusão:** A conclusão destaca o potencial da integração de tecnologias de IA para aprimorar o acesso e a interação com o CCN, impulsionando avanços na gestão e disseminação da informação acadêmica.

Palavras-chave: Ciência de Dados; Ciência da Informação; Análise Descritiva; Análise Preditiva.

Abstract

***Introduction**: The National Collective Catalog of Serial Publications (CCN) is a database that gathers a wide variety of publications from libraries of academic and research institutions across Brazil. The aim is to explore the convergence between Library Science and Data Science in an applied manner, highlighting the extractions from the CCN database after its modernization and restructuring, thereby demonstrating predictive models and generative artificial intelligence components. **Methodology**: Using an empirical approach based on the CCN, the distribution of publications by content, year, and geography is identified, providing insights into research trends in Brazil. **Results**: Predictive models were developed to estimate publication volume, while exploration of Generative Artificial Intelligence components aims to enhance interaction with the CCN. **Conclusion**: The conclusion emphasizes the potential of integrating AI technologies to further improve access and interaction with the CCN, driving advancements in academic information management and dissemination.*

***Keywords:** Data Science; Information Science; Descriptive Analysis; Predictive Analysis.*

Resumen

***Introducción**: El Catálogo Colectivo Nacional de Publicaciones Seriadas (CCN) es una base de datos que reúne una amplia variedad de publicaciones de bibliotecas de instituciones académicas e investigativas en todo Brasil. El objetivo es explorar la convergencia entre Biblioteconomía y Ciencia de Datos de manera aplicada, resaltando las extracciones de la base de datos del CCN después de su modernización y reestructuración, es decir, demostrando los modelos predictivos y componentes de inteligencia artificial generativa. **Metodología**: Utilizando un enfoque empírico basado en el CCN, se identifica la distribución de las publicaciones por contenido, año y geografía, proporcionando información sobre las tendencias de investigación en Brasil. **Resultados**: Se desarrollaron modelos predictivos para estimar el volumen de publicaciones, mientras que la exploración de componentes de Inteligencia Artificial Generativa tiene como objetivo mejorar la interacción con el CCN. **Conclusión**: La conclusión enfatiza el potencial de integrar tecnologías de IA para mejorar aún más el acceso y la interacción con el CCN, impulsando avances en la gestión y difusión de la información académica.*

***Palabras clave:** Ciencia de Datos; Ciencia de la Información; Análisis Descriptivo; Análisis Predictivo.*

1 INTRODUÇÃO

A Biblioteconomia, dedicada à organização de informações em bibliotecas, e a Ciência de Dados, focada na extração de conhecimento de conjuntos complexos de dados (O'Regan, 2023), estão convergindo em resposta ao desafio representado pelo crescente volume de informações digitais. Essa interação dinâmica entre as disciplinas oferece oportunidades promissoras para aprimorar a gestão e o acesso à informação. Desse modo, este trabalho busca explorar a convergência entre Biblioteconomia e Ciência de Dados, de maneira aplicada, destacando as extrações da base do CCN após sua modernização e reestruturação, isto é, demonstrando os modelos preditivos e componentes de inteligência artificial generativa.

No âmbito do projeto Pinakes[1], conduzido pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), ocorre a modernização dos componentes de *software* e a reestruturação dos bancos de dados de serviços essenciais, como o Catálogo Coletivo Nacional de Publicações Seriadas (CCN), o Serviço Brasileiro de Localização (Comut) e a Rede Bibliodata. Essa iniciativa é motivada pela obsolescência tecnológica e pela necessidade de atender aos novos requisitos de interoperabilidade e segurança. O CCN é uma plataforma pública que reúne uma vasta gama de publicações, como periódicos e monografias. Com mais de 62 mil títulos de periódicos e mais de três milhões de transcrições de coleções, o CCN é composto principalmente por bibliotecas ligadas a instituições de ensino e pesquisa em todo o país.

Em Costa *et al.* (2023), foram apresentados os desafios e as estratégias aplicadas na modernização do sistema CCN sob a perspectiva da Engenharia de Dados, juntamente com as técnicas utilizadas na identificação de inconsistência nos dados com vistas ao desenvolvimento de protótipos para a correção de informações catalográficas.

[1] O projeto está disponível em: https://pinakes.tcti.ibict.br/. Acesso em: 19 jan. 2024.

Esta interlocução entre Biblioteconomia e Ciência da Informação com a Ciência de Dados, já teve discussões anteriores, a lembrar das publicações de Fernandes (2019), Matos; Conduru e Benchimol (2021), Freire e Freire (2019), Reis (2021), Paletta (2022). A interlocução está nas habilidades e competências para o ciclo de vida de dados, direcionados a conhecimentos tecnológicos e estatísticos em relação a produção de novos conhecimentos e métodos de recuperação e apresentação da informação.

Enfatizando estes estudos e apresentando uma forma empírica e pragmática, o trabalho utiliza-se como metodologia o estudo de caso (Yin, 2001), foram realizadas análises descritivas e preditivas, fornecendo melhorias na evolução e distribuição da pesquisa científica no Brasil. Os resultados demonstram os modelos preditivos que foram desenvolvidos, e componentes de Inteligência Artificial Generativa que estão em desenvolvimento para melhorar a interação com o CCN.

2 Resultados e Análise: Descritiva, Preditiva e Inteligência Artificial

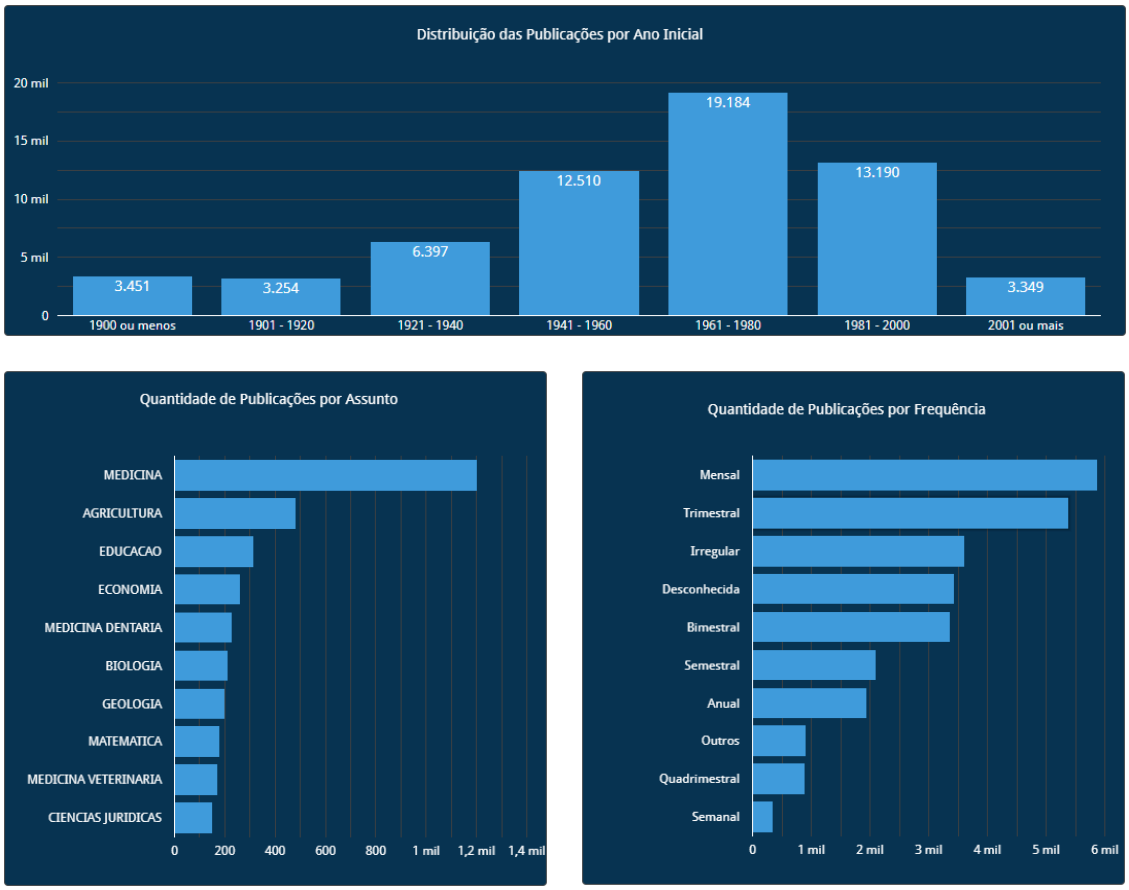
Considerando que a Ciência de Dados é definida como o estudo da extração generalizável de conhecimento a partir de dados (Dhar, 2013), torna-se necessário enfatizar o requisito epistêmico comum na avaliação da aplicabilidade de um novo conhecimento para tomada de decisão, seja descrevendo e contextualizando os dados (análise descritiva), sua capacidade de descrever o passado (análise preditiva) e conversacional por meio da inteligência artificial.

Desse modo, durante a análise descritiva, os dados foram categorizados em um *dashboard*[2] com vias a fornecer informações relevantes acerca das pesquisas científicas brasileiras. Foram definidas sete visões, a saber, (i) Distribuição de Publicações por Conteúdo; (ii) Distribuição das Publicações por Ano Inicial; (iii) Distribuição das Publicações por Ano Final; (iv) Nacionalidade de autoria das

[2] Disponível em: https://lookerstudio.google.com/reporting/8c58534d-22fe-4937-9cdd-d6741878a68f/page/p_qq31vshib. Acesso em: 19 jan. 2024.

publicações; (v) Ranking de Países; (vi) Frequência de Publicações; (vii) Quantitativo de Publicações por Assunto; (viii) Quantitativo de Publicações por Área do Conhecimento, e; (ix) Número de publicações por biblioteca. Exemplos de algumas visões são apresentadas abaixo na Figura 1.

Figura 1 - Visões do Dashboard CCN.



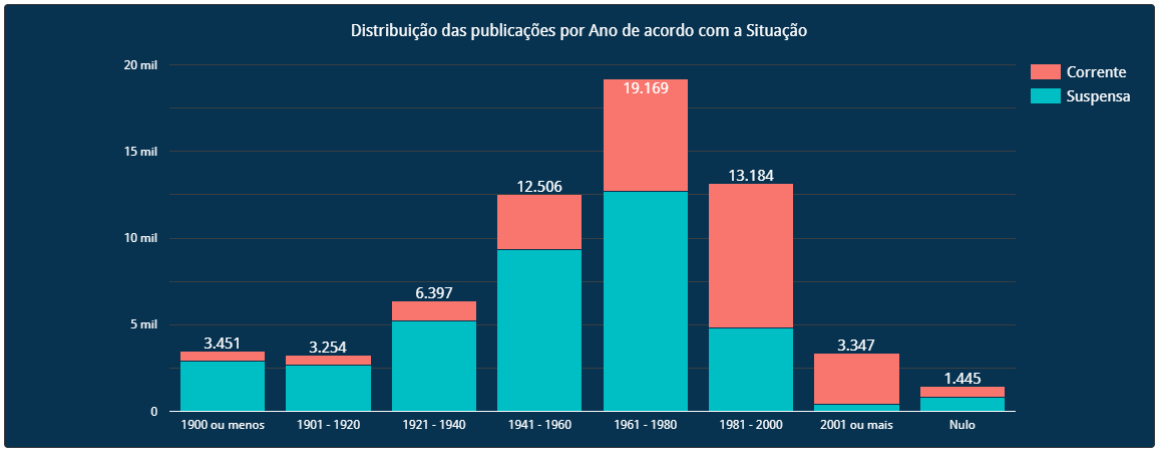
Fonte: Dados da Pesquisa.

Com os dados organizados, identificamos informações importantes para a Biblioteconomia e Ciência da Informação no Brasil. Por exemplo, a maioria das publicações seriadas iniciou entre 1941 e os anos 2000, com mais de 60% das suspensas entre 1961 e os anos 2000. Publicações estão presentes em todos os continentes, com destaque para Estados Unidos, Brasil e França. Predominam publicações mensais ou trimestrais, especialmente em Medicina e Agricultura.

Áreas como ciências da saúde e agrárias têm maior volume. Cerca de 60% das publicações estão em até três bibliotecas, principalmente as centrais. *Softwares* como *Pergamum*, *Aleph* e *SophiA* são amplamente utilizados, com adesão aos serviços do Ibict.

De maneira a complementar, também foi realizado um estudo das principais correlações não-evidentes entre os diversos atributos e classes no modelo de dados. Foi possível identificar as principais relações por meio de testes estatísticos, como o teste de Cramer e o teste qui-quadrado. Assim, foi acrescido ao *dashboard*, uma página com os principais resultados deste estudo. Um exemplo de visualização é apresentado na Figura 2:

Figura 2 - Publicações por Ano e Situação.



Fonte: Dados da Pesquisa.

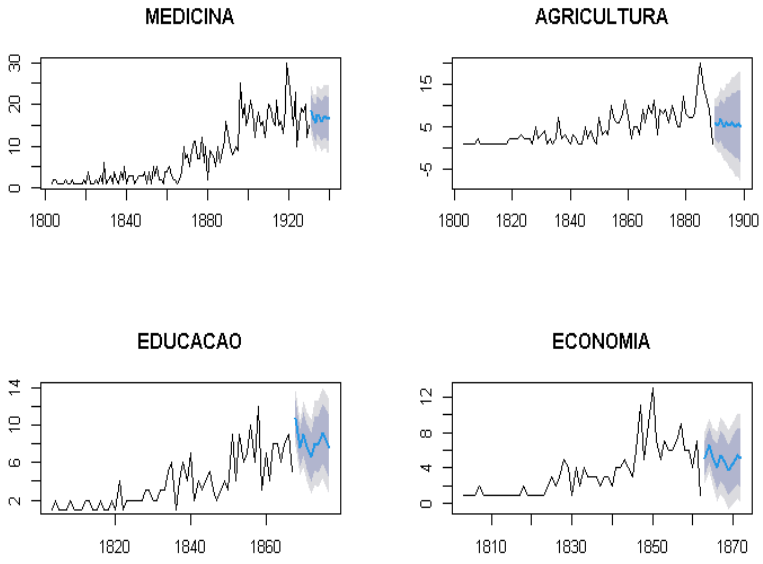
Ao analisar a situação das publicações, observou-se que o período com maior número de publicações também registrou o maior número de revistas suspensas. Utilizando um modelo preditivo baseado em dados do CCN, aplicamos a teoria das séries temporais para prever a quantidade de publicações por assunto entre 1981 e 1990, utilizando dados históricos disponíveis desde 1800 até 1980 para treinar o modelo. O estudo focou em áreas de conhecimento com maior número de publicações, como Medicina, Agricultura, Educação e Economia. A teoria das séries

temporais é uma abordagem amplamente reconhecida na análise estatística para prever eventos futuros com base em padrões e tendências identificados em dados coletados ao longo do tempo.

A análise inicial revelou que, visualmente, as projeções para a quantidade de publicações em cada assunto são consistentes, vale ressaltar a projeção referente ao tema "Agricultura", que foi notavelmente influenciada por um aumento repentino na quantidade de publicações, seguido por um retorno aos valores considerados normais, através de uma queda acentuada no final da série.

Ao confrontar as previsões com os dados reais, observou-se uma concordância satisfatória, com diferenças consideravelmente pequenas, em torno de 2,94 publicações. O período específico escolhido para as previsões corresponde ao último intervalo de tempo para o qual dados reais estão disponíveis de forma consistente, permitindo comparações diretas. No entanto, este estudo sugere a viabilidade de expandir a análise para incluir mais assuntos e com acesso a dados mais recentes, estendendo as previsões até períodos mais atuais.

Figura 3 - Resultado do modelo preditivo para os assuntos Medicina, Agricultura, Educação e Economia.



Fonte: Dados da Pesquisa.

Com o advento dos Grandes Modelos de Linguagem, do inglês *Large Language Models* (LLM), um direcionamento do projeto voltou-se para a aplicação de LLMs no contexto da Biblioteconomia e catalogação. O objetivo é permitir a consulta ao catálogo CCN utilizando os recursos de linguagem, obtendo, assim, uma catalogação conversacional, ou seja, permitir atividades de catalogação não apenas com os recursos de formulários e dados tabulados, mas, através de interação dialógica. A primeira atividade foi a análise exploratória dos LLMs. A tabela abaixo resume as principais características.

Foi realizada uma análise de várias LLMs disponíveis com licença open-source, selecionadas com base em requisitos como capacidade de resposta em português, número de parâmetros e viabilidade de execução em computadores pessoais. Entre essas LLMs estão: Llama, *Alpaca*, *Cabrita*, *Mistral*, *Tinyllama*, *Starling*, *Zephyr*, *Falcon* e *Deepseek*. Destaca-se o LLM Llama, desenvolvido pela Meta AI, que acessa 48,9% do conhecimento mundial *on-line*. O modelo Alpaca é uma versão reduzida do *Llama*, enquanto o Cabrita é otimizado para tradução para o português. O *Deepseek*, do MIT, é treinado em inglês e chinês, e o Falcon (7B), do *Technology Innovation Institute,* possui uma versão adaptada para rodar sem GPU. O Starling é desenvolvido pela Universidade da Califórnia.

Tabela 1 - Comparação dos LLMs.

Nome	GPU	QTD Parâmetros	Res. Português	Tecnologia	Rodou?
Llama	16GB	7B	Não	Llama	Não
Alpaca	16GB	7B	Não	Llama	Não
Cabrita	16GB	7B	Sim	Llama	Não
Mistral	16GB	7B	Não	Llama	Sim
Tinyllama	--	1.1B	Não	Llama	Sim
Starling	16GB	7B	+/-	Openchat 3.5	Sim
Zephyr	16GB	7B	+/-	Mistral	Sim
Falcon	--	1B	Sim (com tradução)	OpenOrca	Sim
Falcon	16GB	7B	Sim	OpenOrca	Não
Deepseek	8GB	1.3B	Sim	Llama	Não

Fonte: Dados da Pesquisa.

Após a análise exploratória dos modelos de linguagem, foi feito um estudo nas técnicas a serem utilizadas na personalização dos modelos para o contexto do CCN. Foram utilizadas duas técnicas, *Fune-Tuning* e *RAG*.

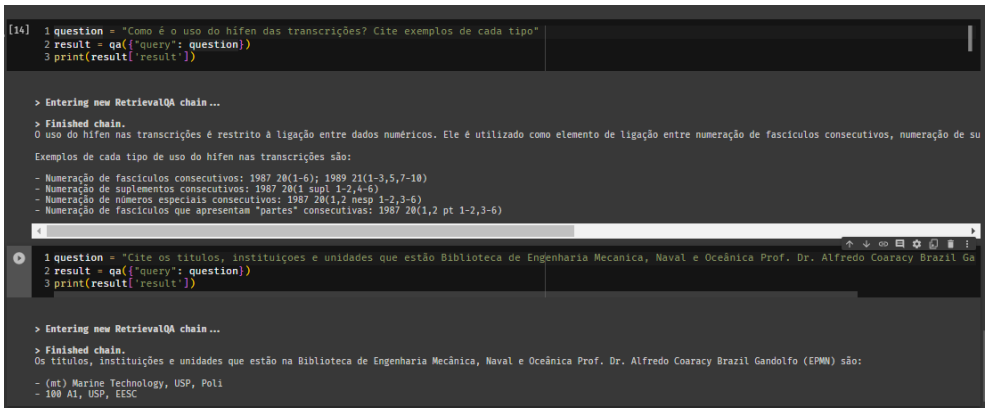
Fine-tuning é um processo de ajuste refinado em modelos de linguagem. Durante o treinamento inicial, LLM's são alimentados com grandes volumes de dados para aprender a estrutura da linguagem e a relação entre palavras e conceitos. No entanto, para aplicá-los a tarefas específicas ou domínios particulares, como Biblioteconomia, é necessário realizar um processo de *fine-tuning.* Neste processo, o modelo pré-treinado é ajustado com um conjunto de dados mais específico do domínio de interesse. Isso permite que o modelo aprenda padrões, nuances e terminologias específicas do domínio, melhorando sua capacidade de gerar respostas ou realizar tarefas de maneira mais precisa e relevante para o contexto desejado.

Por sua vez, o processo de Geração Aprimorada por Recuperação, do inglês *Retrieval-Augmented Generation* (RAG), envolve aprimorar as respostas de um modelo de linguagem de grande escala, assegurando que ele consulte uma fonte de conhecimento externa e confiável, além de seus dados iniciais de treinamento, antes de produzir uma resposta. A principal diferença entre as técnicas entende-se por: o *fine tuning* altera o próprio modelo, o RAG não há alteração no modelo, ele é somente utilizado para a montagem da resposta final. O RAG foi a técnica mais indicada no contexto do CCN pois os dados do CCN são atualizados, frequentemente e pela alta quantidade de dados, o *fine tuning* tem a tendência de alucinar nas suas repostas que seria quando o modelo não sabe a resposta, mas responde mesmo assim.

O CCN foi aprimorado com um conjunto reduzido de dados, que incluía uma tabela de publicações, suas respectivas bibliotecas, instituições e unidades. Após a primeira etapa, o processo de RAG foi concluído, utilizando os frameworks

LangChain[3] e *LlamaIndex*[4]. Foi realizado um teste com uma seção menor do documento de regras de transcrição, onde o usuário pode fazer perguntas a respeito do documento e os procedimentos para os dois testes foram os mesmos. Assim, um exemplo de resposta utilizando o *Llama LLM* é apresentado na Figura 4.

Figura 4 - Exemplo de resposta do CCN utilizando o Llama LLM.



Fonte: Dados da Pesquisa.

3 CONSIDERAÇÕES FINAIS

As considerações finais do estudo destacam a importância dos dados organizados para fornecer insights valiosos nas áreas de Biblioteconomia e Ciência da Informação no contexto nacional. Os padrões identificados, como a distribuição temporal das publicações e a predominância de certos assuntos, fornecem uma compreensão mais profunda da produção de conhecimento nessas áreas.

Pode-se enfatizar que a interação entre a Biblioteconomia e Ciência de Dados é fundamental para superar os desafios impostos pelo crescente volume de informações digitais. A convergência dessas disciplinas, como ilustrada pelo estudo do Catálogo Coletivo Nacional de Publicações Seriadas (CCN), demonstra potencial significativo para aprimorar tanto a gestão quanto o acesso à informação

[3] Disponível em: https://www.langchain.com/. Acesso em: 19 jan. 2024.
[4] Disponível em: https://www.llamaindex.ai/. Acesso em: 19 jan. 2024.

acadêmica. Por meio da aplicação de modelos preditivos e componentes de Inteligência Artificial Generativa, foi possível não apenas estimar o volume de futuras publicações, mas também melhorar a interação com o CCN, evidenciando a relevância da integração de tecnologias avançadas no campo da biblioteconomia.

Além disso, a utilização de Grandes Modelos de Linguagem (LLMs) para catalogação conversacional aponta para a inovação contínua no acesso e na organização da informação, permitindo uma abordagem mais dinâmica e interativa na pesquisa acadêmica. Esse avanço tecnológico, ao lado das análises descritivas e preditivas, enfatiza a necessidade de bibliotecários e profissionais da informação se adaptarem às novas ferramentas e métodos analíticos, garantindo que as práticas de biblioteconomia permaneçam relevantes e eficazes na era digital. Portanto, este estudo sublinha a importância de explorar e expandir a sinergia entre Biblioteconomia e Ciência de Dados para melhorar a gestão da informação e promover o desenvolvimento de estratégias inovadoras de disseminação do conhecimento.

REFERÊNCIAS

COSTA, B.; VIANA, J. G.; SANTOS, G.; SANTOS, G.; ASSIS, T. B. Engenharia de Dados e Biblioteconomia: A modernização do Catálogo Coletivo Nacional de Publicações Seriadas. *In*: TRILHA DA INDÚSTRIA - CONGRESSO BRASILEIRO DE SOFTWARE: TEORIA E PRÁTICA (CBSOFT), 14., 2023, Campo Grande/MS. *Anais* [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 17-20. Disponível em: https://cbsoft2023.ufms.br/. Acesso em: 03 mar. 2024.

DHAR, V. Data science and prediction. *Communications of the ACM*, [*s.l.*], v. 56, n. 12, p. 64-73, 2013. Disponível em: https://doi.org/10.1145/2500499. Acesso em: 29 mar. 2024.

FERNANDES, J. H. C. Interlocuções bibliográficas e epistemológicas entre a ciência de dados e a ciência da informação. *Ciência da Informação*, [*s.l.*], v. 49, n., 2019. Disponível em: https://doi.org/10.18225/ci.inf.v49i3.5655. Acesso em: 03 mar. 2024.

FREIRE, G. H. A.; FREIRE, I. M. Ciência de dados e ciência da informação. *Informação & Sociedade*: Estudos, [*s.l.*], v. 29, n. 3, 2019. Disponível em: https://cip.brapci.inf.br/download/147968. Acesso em: 03 mar. 2024.

MATOS, M. T.; CONDURU, M. T.; BENCHIMOL, A. C. A produção científica e o acesso aberto sobre a ciência de dados no contexto da ciência da informação: estudo bibliométrico. *Páginas A&B*: Arquivos e Bibliotecas, Portugal, v., n. esp, 2021. Disponível em: https://ojs.letras.up.pt/index.php/paginasaeb/article/view/10228/9636. Acesso em: 03 mar. 2024.

O'REGAN, G. Introduction to Data Science. *In*: *Mathematical Foundations of Software Engineering:* Texts in Computer Science. Springer, 2023. Disponível em: https://link.springer.com/chapter/10.1007/978-3-031-26212-8_24. Acesso em: 25 mar. 2024

PALETTA, F. C. Fundamentos de ciência de dados e inteligência artificial: conexões com a ciência da informação. *Revista Fontes Documentais*, [*s.l.*], v. 5, n. ed., 2022. Disponível em: https://periodicos.ifs.edu.br/periodicos/fontesdocumentais/article/view/1446. Acesso em: 03 mar. 2024.

REIS, M. J.; SENA, N. C. S. Biblioteconomia de dados e ciência de dados no contexto da e-science. *Revista Fontes Documentais*, [*s.l.*], v. 4, n. ed., 2021. Disponível em: https://periodicos.ifs.edu.br/periodicos/fontesdocumentais/article/view/1310. Acesso em: 03 mar. 2024.

YIN, R. K. *Estudo de caso*: planejamento e métodos. Bookman editora, 2001.