XXIII Seminário Nacional de Bibliotecas Universitárias
SNBU 2025
17 A 20 DE NOVEMBRO
SÃO PAULO - SP

**Eixo 4 - Produtos, Serviços, Tecnologias e Inovação**

# Avaliação de acervos com IA

*Collection assessment with AI*

**Ernesto Carlos Bode** – Biblioteca da Câmara dos Deputados – bod.ernesto@gmail.com

**Janice de Oliveira e Silva Silveira** – Biblioteca da Câmara dos Deputados – janice.silveira@camara.leg.br

**Cleber Zanchettin** – Universidade Federal de Pernambuco (UFPE) – cz@cin.ufpe.br

**Resumo**: [Objetivos] Investigar o potencial da clusterização para subsidiar a avaliação de acervos e coleções. [Fundamentação] A partir de dados de uma coleção específica de uma Biblioteca, foram produzidos *embeddings* com um modelo de Inteligência Artificial; a partir disso, foi aplicado um algoritmo de clusterização a fim de obter dados relevantes sobre o acervo. [Resultados] Foram obtidos dados qualitativos em classes CDU separadas. [Conclusões] O uso das técnicas aqui apresentadas, com apoio de modelos de inteligência artificial, possibilita produzir informações qualitativas sobre o acervo, subsidiando assim avaliações sobre a formação do acervo.

**Palavras-chave**: Acervos. Avaliação da coleção. Aprendizado de máquina. Inteligência artificial. Clusterização.

**Abstract**: [Objectives] This study investigates the potential of clustering techniques to support the evaluation of library collections. [Fundamentação] Using data from a Library, *embeddings* were generated through a Machine Learning model, followed by the application of a clustering algorithm to identify relevant information about the collection. [Findings] The results revealed qualitative patterns in CDU classes, allowing for a deeper analysis of the collection's composition. [Conclusion] The findings suggest that clustering techniques, supported by artificial intelligence models, provide consistent insights to enhance collection evaluation and management in libraries.

**Keywords**: *Keywords: library collections. Collection assessment. Machine Learning. Artificial intelligence. Clustering.*

## 1 INTRODUÇÃO

Efetuamos uma busca na base de teses e dissertações do Ibict[1] (sem especificar programa, ou universidade) e seus resultados são um bom ponto de partida para a discussão da relação entre Inteligência Artificial e Ciência da Informação, no contexto atual. Utilizamos na busca três termos que aparecem com frequência na pesquisa sobre IA: *Machine Learning*, *Redes Neurais* e o próprio termo *Inteligência Artificial.*

Os resultados da busca, como seria de se esperar, evidenciam a forte presença de programas na áreas Engenharias e Ciência Computação, naturalmente produtoras de pesquisa sobre IA, em consonância com suas respectivas grades curriculares. Os resultados indicam uma oportunidade de crescimento para a pesquisa em IA nos Programas de Pós Graduação em Ciência da Informação (PPGCI), ao se comparar com o volume de produção em áreas como Tecnologia da Informação. É claro que trata-se de uma busca bastante limitada, servindo aqui apenas como ponto de partida nas discussões. Deve-se considerar ainda que alguns trabalhos não listaram seus programas nos metadados de busca, como (MATTOS, 2025) do PPGCI da UFF[2] .

Dos resultados, foi possível notar também que há pesquisas sobre IA em domínios típicos da Ciência da Informação levadas a cabo por outros programas que não PPGCI's, como (FERREIRA, S. L., 2004)[3].

O objetivo desta discussão não é apontar diferenças quantitativas entre áreas, mas refletir sobre oportunidades para que a Ciência da Informação amplie sua atuação em pesquisa sobre IA. Mais relevante do que o número de trabalhos identificados nos PPGCI's, se comparado a outros programas, é explorar questões como: quais temas a área poderia investigar nesse campo?. Quais lacunas precisam ser preenchidas?

Além de compreender e aplicar tecnologias de IA em processos de trabalho em arquivos e bibliotecas, é importante considerar abordagens que vão além do uso de ferramentas prontas — como o ChatGPT e similares.

O presente experimento propõe que a área avance para além do uso de ferramentas prontas, explorando a vasta gama de tecnologias de IA de forma mais direta e aprofundada. Isso abre um campo fértil para a aplicação de linguagens de

[1] Pesquisa de busca no site do Ibict (https://bdtd.ibict.br) em 07.ago.2025.
[2] Pesquisa sobre IA no serviço de referência de bibliotecas universitárias.
[3] Sistema inteligente para Organização de Documentos.

programação, não necessariamente com a profundidade exigida em cursos de TI, mas focada em tarefas específicas da área CI. Seria suficiente dominar o uso de *scripts* de código, que significa executar tarefas como importar dados, ou fazer chamadas a modelos de IA. As facilidades de ambientes para uso de *scripts* como a solução Colab[4]. E a estrutura da linguagem *python*, bastante simplificada, a torna ideal para pesquisadores sem um profundo conhecimento em linguagens de programação. Ambas as soluções tem vasta documentação disponível. Este seria um 'letramento em IA' em um nível mais aplicado, capacitando o pesquisador a interagir diretamente com os modelos e dados.

O experimento que aqui apresentamos é uma exemplificação do preenchimento dessa lacuna. Trazendo tecnologias de IA para um problema de Ciência da Informação (Avaliação em Desenvolvimento de Coleções), interagindo diretamente com modelos de IA.

Apresentaremos, no entanto, a versão funcional do processo e suas etapas. Para aqueles que optarem pela replicação do experimento, inclusive aplicando-o em novos ambientes de trabalho, disponibilizamos em repositório público[5] todo o código de *scripts,* em linguagem *python*, extensivamente comentado. Antes de apresentar as etapas do processo, porém, vale destacar outro ponto sobre letramento em IA e que ajuda a posicionar melhor nosso o experimento.

A dissertação (FROTA, K. A. K., 2022) efetuou uma revisão sistemática de literatura[6] sobre IA "no contexto da Ciência da Informação". Dessa pesquisa, o ponto que gostaria de destacar são os elementos de IA que mais apareceram no universo pesquisado: "*Natural Language Processing (NLP)*, *Machine Learning*, *IA Geral* e *Search Engines*". Mas quase nada foi dito sobre *Large Language Models* (LLM's), o que deve ter ocorrido em função de como esta tecnologia é recente, mas importante.

O surgimento e aplicação de LLM's nos últimos anos é um ponto disruptivo para toda a área de IA e suas aplicações nas sociedades, inclusive em Arquivos e Bibliotecas. Foram essas tecnologias que causaram transformações tão relevantes a ponto de impulsionar tentativas para regulamentar a IA, como o **IAact** da União Européia que entrou em vigor recentemente (agosto de 2024), com propostas desde

[4] https://colab.google/ (ferramenta gratuita, com opções de versões pagas)
[5] https://github.com/ErnyBSB/clusteringForLibraries
[6] "no âmbito internacional" entre 2017 e 2021.

2021[7]. Por isso, seu uso e aplicação na pesquisa de Ciência da Informação deveria ser privilegiado. No experimento, abaixo descrito, utilizamos um LLM na última etapa do processo.

## 2 METODOLOGIA

Nos procedimentos metodológicos que descrevemos a seguir, o ponto de partida conceitual em Ciência da Informação é "Desenvolvimento de Coleções", mais especificamente seu processo de Avaliação (WEITZEL, 2012). O objetivo final é o de potencializar o processo de avaliação de coleções numa Biblioteca.

A hipótese é que o uso da técnica de *clusterização* (na categoria *unsupervised learning*), usando modelos de IA, e a análise de seus resultados com um modelo *LLM*, pode ser uma ferramenta útil para apoiar a Avaliação de coleções numa Biblioteca. Com subsídios que poderiam nortear o Desenvolvimento de Coleções.

### 2.1 Importação e tratamento de dados

A primeira tarefa do experimento foi o procedimento de obtenção dos metadados sobre um acervo de biblioteca. Utilizamos aqui o termo *dataset*, emprestado da área tecnologia da informação, normalmente usado no sentido de dados ainda sem os devidos tratamentos, disponibilizados para procedimentos de IA como treinamento de modelos[8].

Os metadados empregados nesta pesquisa referem-se à coleção "acervo geral" da Biblioteca escolhida, composta por mais de 182 mil obras.

A primeira etapa do processo se deu com a criação de um *dataframe*[9] com os dados originais. Trata-se de um objeto da biblioteca python Pandas e se refere a um recurso que permite uma série de operações de dados. No nosso caso, foi executado um pré-processamento dos metadados[10] como a (I) remoção de caracteres inválidos, (II) espaços desnecessários, (III) remoção de títulos repetidos, (IV) remoção de *stop-words*, (V) *tokenização* e (VI) *stemmização*. Com a remoção de títulos repetidos, a etapa resultou em 131.296 obras no *dataframe.*

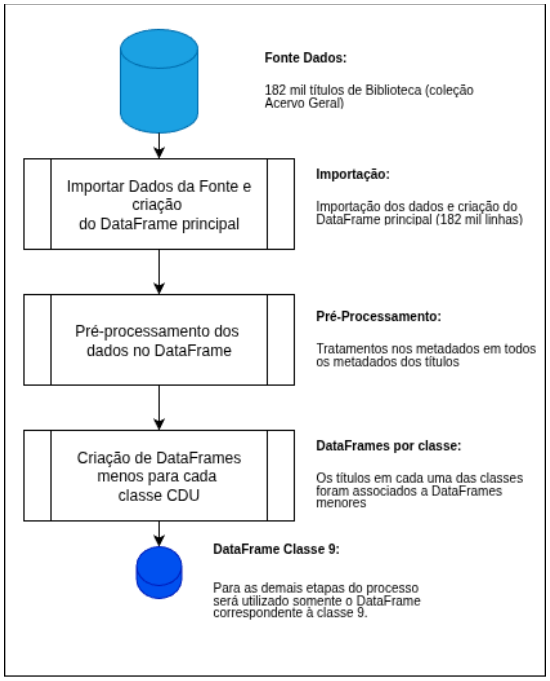
[7] https://www.consilium.europa.eu/pt/policies/artificial-intelligence/
[8] https://www.kaggle.com/datasets/ernestocarlosbode/books-metadata
[9] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html
[10] Procedimentos comuns de NLP (natural language processing).

## 2.2 Divisão das classes CDU

A segunda etapa procedimental foi a criação de *dataframes* menores correspondentes a cada uma das classes CDU[11] do total de dados. Esta estratégia foi adotada, pois observamos nos testes inicias que executar os procedimentos em uma classe CDU por vez (de 0 a 9) torna o processo mais preciso e coerente.

Importante ressaltar que estamos apresentando neste artigo apenas os resultados relativos à **classe nove** (história, geografia e biografias) da coleção considerada, porém os procedimentos são aplicáveis, às demais classes, com a alteração de algumas variáveis que discutiremos mais adiante. O *dataframe* específico para a classe 9 contém 11.859 obras (considerando títulos únicos). A *figura 1* permite visualizar essas primeiras etapas.

**Figura 1** - Etapas inicias do processo




Fonte: elaborado pelos autores

## 2.3 Criação dos Embeddings

A próxima etapa corresponde à tarefa ***embeddings*** dos metadados para cada título na biblioteca. A necessidade dessa etapa baseia-se no fato de que o processamento de dados nos modelos de IA é feito a partir de números. Assim, tomemos a coluna do dataframe com os títulos. Os textos de cada título serão

[11] O acervo em questão utiliza a Classificação Decimal Universal (CDU).

convertidos para uma representação tipo vetorial, basicamente, um número com coordenadas. Os modelos disponíveis para *embeddings* foram treinados para criar vetores em espaços geométricos correspondentes a conceitos no sentido semântico, sendo que vetores similares correspondem a conceitos similares semanticamente. Assim, os vetores correspondentes aos títulos que tenham palavras como "história da América" serão similares a aqueles com texto "história da Ásia" e estariam bem distantes (no sentido matemático) de títulos como "Química Orgânica".

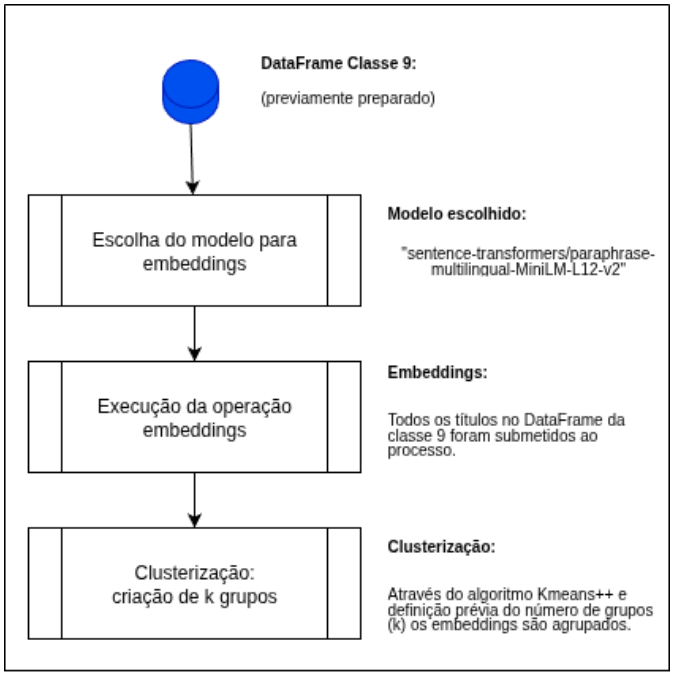
No processo final da experimentação de *embedding* utilizamos modelos da família BERT[12] (*Bidirectional Encoder Representations from Transformers*). Apesar de introduzido em 2018[13], pela Google AI, possui várias versões atualizadas, inclusive multilíngues baseadas na versão original, disponíveis publicamente. Adotamos o modelo: *'sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2'* no repositório *Hugging Face*[14]. Essa etapa gerou outras quatro colunas com dados na forma numérica que serão utilizadas na etapa seguinte.

### 2.4 Processo de Clustering

A próxima etapa do processo: ***Clustering*** é executada a partir das quatro colunas que já passaram pelo processo de *embeddings*. Basicamente, um algoritmo de *clustering* agrupa os vetores semelhantes (que correspondem aos metadados das obras) de acordo com um número predefinido de *clusters,* que é definido e informado ao algoritmo pelo pesquisador. Para esta etapa, escolhemos o algoritmo *K-means++*[15]. O critério para escolha do melhor número k foram análises gráficas que mostraram a proximidade (ou dispersão) gráfica dos pontos, cada ponto correspondendo a uma versão simplificada dos vetores. Vale destacar que para a execução do processo em outras classes CDU haverá a necessidade de repetir a análise gráfica. A figura 2 apresenta graficamente até agora descritas.

[12] https://huggingface.co/docs/transformers/en/model_doc/bert
[13] https://pt.wikipedia.org/wiki/BERT_(modelo_de_linguagem)
[14] https://huggingface.co/
[15] https://en.wikipedia.org/wiki/K-means%2B%2B

**Figura 2** - Etapas de embeddings e clustering




Fonte: elaborado pelos autores

Visualizar os dados no *dataframe* como tabela é uma forma prática de assimilar o andamento das etapas do processo até agora. Observar na tabela da figura 3 que para cada linha e para cada coluna, cria-se uma coluna correspondente aos respectivos *embeddings*[16]. A identificação de qual *cluster* cada obra/linha pertence aparece na coluna "*cluster*" (tabela na figura 3), leia-se: *pertence ao agrupamento x, no nosso caso x de 0 a 5*.

## 2.5 Análises com LLM

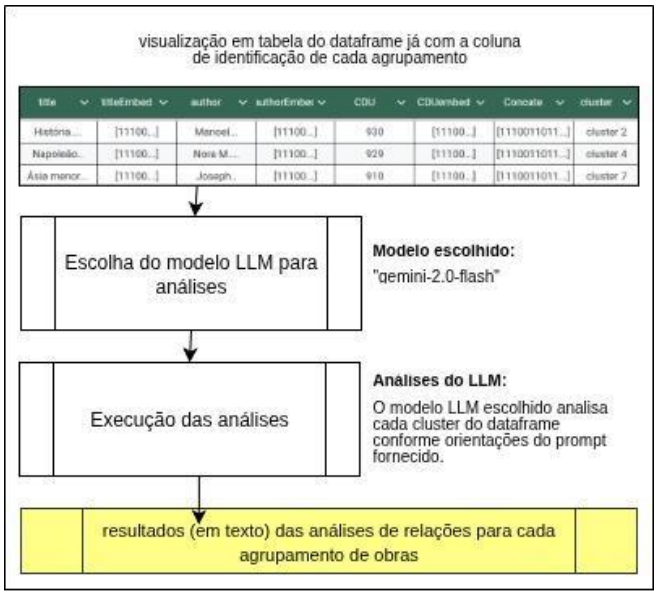
A última etapa do processo, antes da análise dos resultados finais obtidos, utiliza novamente um modelo de IA, desta vez um Large Language Model (LLM), mais especificamente o *gemini-2.0-flash*[17]. Esse modelo foi escolhido por ser gratuito, nos limites de uso dessa experimentação, além de ser um dos mais potentes disponíveis no momento dos testes.

O uso desse modelo LLM aqui se dá através da criação de um "*prompt*", ou seja, um comando que inserido no código permite analisar o *dataframe* que contém os dados de cada obra (*author, title, publish, CDU*) já identificados em relação ao

[16] A coluna 'Concate' corresponde à concatenação das colunas utilizadas. O algoritmo de *clustering* precisa apenas da coluna 'Concate' para criar os agrupamentos.
[17] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash

respectivo *cluster*. Importante aqui compreender que as etapas de *embeddings* e *clustering* foram utilizadas para agrupar as obras em *clusters*. Mas a análise do modelo LLM ocorre diretamente nos dados das colunas originais, ou seja, o texto dos títulos, o texto do código CDU e etc.

**Figura 3** - Última etapa do processo e suas entregas




Fonte: elaborado pelos autores

Na forma como desenhamos o *prompt,* foi solicitado ao modelo "identificar e analisar" relações hierárquicas nos dados fornecidos. O objetivo foi compreender porque os títulos de livros foram agrupados da forma como o algoritmo agrupou. O trabalho do LLM nesta etapa é identificar o porquê dessa proximidade semântica, assim permitindo conhecer e visualizar relações entre as obras da classe 9 na coleção.

## 3 RESULTADOS E DISCUSSÕES

O agrupamento de obras da coleção em *clusters* é apenas o ponto de partida para compreender porquê há proximidade semântica em cada *cluster*. A classe CDU utilizada corresponde a três áreas, mas foi possível agrupar, com coerência, as obras em seis *clusters*. A quantidade de obras em cada *cluster* criado é bastante diferente, variando entre uma centena até 3 mil obras. Essa variação já fornece um panorama da distribuição do acervo dentro da classe, indicando áreas de maior ou menor concentração de obras, no nosso caso, a forte presença de Biografias. O *cluster* com

menos obras da coleção analisada *"representa um conjunto coeso de obras sobre a história da antiguidade clássica, com foco principal em Roma e Grécia",* conforme nos informou a análise do LLM.

Em todas as tentativas de execução da clusterização, destacou-se consistentemente a predominância de Biografias na coleção, o que reforça a validade do processo e a correta associação de assuntos. Mas, o aspecto mais relevante dos resultados são as informações **qualitativas**. O *cluster* 3, por exemplo, conseguiu detectar uma associação entre diferentes áreas da Classe 9, entre história e geografia.

As implicações para o Desenvolvimento de Coleções são a possibilidade de automatizar sua Avaliação, através de uma técnica que permite sua execução frequentemente, a um custo baixo. Além disso, as informações obtidas nos *clusters* com as análises e, de fato, interpretações do LLM, permitem obter respostas **qualitativas** acerca de suas coleções. Por exemplo, as análises no *cluster 9 e cluster 3* pode, ou não, estar de encontro com as diretrizes da Política de Desenvolvimento da biblioteca considerada.

Podemos prever, no entanto, limitações. Na clusterização, a escolha do número de grupos é um problema, há técnicas estatísticas que podem ser utilizadas para mitigar o problema, mas o fator "negócio", ou seja o que cada coleção efetivamente tem em cada classe sempre será fundamental e portanto não parece ser possível ignorar múltiplas tentativas de agrupamento. Com relação aos LLM's, não se trata de um recurso que pode, facilmente, ser "adquirido" e por isso dependemos de sua disponibilidade via redes. Possíveis custos de uso pode ser uma barreira.

Para todos os tipos de modelos utilizados aqui, a questão da língua portuguesa pode ser um limitador de sua efetividade. Já que a presença dessa língua não é majoritária no treinamento dos modelos. Mesmo assim, a efetividade dos modelos para a língua portuguesa é um campo em contínuo desenvolvimento, e pesquisas como esta contribuem para impulsionar e validar novas ferramentas

## 4 CONSIDERAÇÕES FINAIS

A reprodução do experimento aqui descrito pode se beneficiar de customizações com alteração das variáveis disponíveis: (I) desenhos de "prompt", (II)

uso de diferentes modelos LLM, (III)modelos de *embeddings*, (IV) diferentes quantidades de clusters(k), ou até outros (V) algoritmos de clusterização.

As Bibliotecas estão em constante renovação de serviços a fim de "gerar valor a quem busca o conhecimento" (LIRA; JACINTHO, 2023). O uso básico de *scripts* de código entre bibliotecários, pode ir ao encontro do necessário *letramento com IA* para os novos serviços que estão surgindo o que soma-se ao *letramento informacional* já em marcha. (PINHEIRO; COSTA; VITORIANO, 2025). Esse caminho pode abrir portas também para uma maior colaboração interdisciplinar com áreas de tecnologia.

Nossa proposta com o experimento aqui abordado pretende não apenas apontar um caminho, mas permitir que os profissionais da informação se posicionem como protagonistas na criação de serviços inteligentes e alinhados às novas demandas no nosso século.

## REFERÊNCIAS


FERREIRA, S. L. **Sistema inteligente de organização de documentos**. 2004. Dissertação (Mestrado) – Instituto Tecnológico da Aeronáutica, São José dos Campos, 2004.

FROTA, K. A. K. **Inteligência artificial no contexto da ciência da informação**: uma revisão sistemática de literatura. 2022. Dissertação (Mestrado) – Universidade do Rio Grande do Sul, Porto Alegre, 2022.

LIRA, Edna K.; JACINTHO, Eleiana M. dos Santos. Tendências de serviços para biblioteca e as competências do profissional bibliotecário: um olhar para o futuro. **Transinformação**, [*S. l.*], v. 35, 2023. Disponível em: https://periodicos.puc-campinas.edu.br/transinfo/article/view/6953/7413. Acesso em: 5 ago. 2025.

MATTOS, J. M. M. **Inteligência artificial no serviço de referência de bibliotecas universitárias**: colóquios e aplicações. 2025. Dissertação (Mestrado) – Universidade Federal Fluminense, Niterói, 2025.

PINHEIRO, Maria H. B.; COSTA, Marcos R. M.; VITORIANO, Maria Albeti Vieira. A interface entre inteligência artificial e letramento informacional no ensino superior: contextos, avanços e desafios. **Encontros Bibli**, [*S. l.*], v. 30, 2025.

WEITZEL, Simone da Rocha. Desenvolvimento de coleções: origem dos fundamentos contemporâneos. **Transinformação**, [*S. l.*], v. 24, n. 3, p. 213-228, set./dez. 2012. Disponível em: https://www.scielo.br/j/tinf/a/PMK9FqgDj9rMs9WtmYKd5nb/. Acesso em: 2 abr. 2025.