

Inteligência Artificial Generativa Aplicada na Análise da Produção Científica

Programa de Pós-Graduação em Informática e Gestão do Conhecimento (PPGI) - Uninove
Programa de Pós-Graduação em Gestão de Projetos (PPGP) - Uninove

Dr. Edson Melo de Souza

AULA 02

Gestão de Dados Científicos e as Leis de Bradford, Lotka e Zipf

Objetivos

- ▶ Compreender a importância da Gestão de Dados Científicos (GDC) na pesquisa moderna.
- ▶ Compreender os fundamentos das Leis de Bradford, Lotka e Zipf.
- ▶ Interpretar graficamente e matematicamente a distribuição da produção científica.
- ▶ Aplicar essas leis para solucionar problemas de gestão de coleções e organização de dados científicos (metadados).

O Que é Gestão de Dados Científicos?

A Gestão de Dados Científicos (GDC) é o processo ativo de coletar, organizar, armazenar, preservar e compartilhar dados gerados durante uma pesquisa.

- ▶ Garante a qualidade e integridade dos dados.
- ▶ Maximiza o impacto da pesquisa.
- ▶ Promove a transparência e reproduzibilidade.
- ▶ Atende a requisitos de agências de fomento.

Ciclo de Vida dos Dados

O ciclo de vida dos dados descreve as etapas fundamentais da gestão:

- 1. Planejamento e Coleta:** Definição do Plano de Gestão (PGD) e métodos de coleta.
- 2. Processamento e Análise:** Limpeza, organização e análise dos dados brutos.
- 3. Armazenamento:** Uso de repositórios e garantia de segurança a longo prazo.
- 4. Compartilhamento e Reúso:** Publicação dos dados, metadados e fomento a novas pesquisas.

O Plano de Gestão de Dados (PGD)

O PGD é um documento formal que descreve como os dados serão gerenciados ao longo do ciclo de vida da pesquisa. Elementos chave incluem:

- ▶ **Tipos de dados:** a serem gerados (ex: tabulares, imagens, código).
- ▶ **Padrões de metadados:** para descrever os dados.
- ▶ **Políticas de armazenamento:** e backup durante a pesquisa.
- ▶ **Questões éticas e de privacidade:** (ex: dados sensíveis, LGPD).
- ▶ **Estratégia de compartilhamento:** e preservação a longo prazo.

Princípios Fundamentais: FAIR

Os princípios FAIR guiam o compartilhamento e a reutilização de dados para que sejam:

- ▶ **Findable (Localizáveis):** Dados e metadados com identificadores únicos (ex: DOI) e indexados em mecanismos de busca.
- ▶ **Accessible (Acessíveis):** Dados recuperáveis por seus metadados, com protocolos abertos e autenticação clara.
- ▶ **Interoperable (Interoperáveis):** Usam vocabulários controlados e formatos padronizados para integração com outros dados.
- ▶ **Reusable (Reutilizáveis):** Metadados ricos, licença de uso clara e conformidade com padrões da comunidade.

"A Ciência Aberta visa tornar a pesquisa acessível a todos. A GDC fornece a infraestrutura para que isso aconteça de forma eficaz e ética."

Benefícios da GDC na Ciência Aberta:

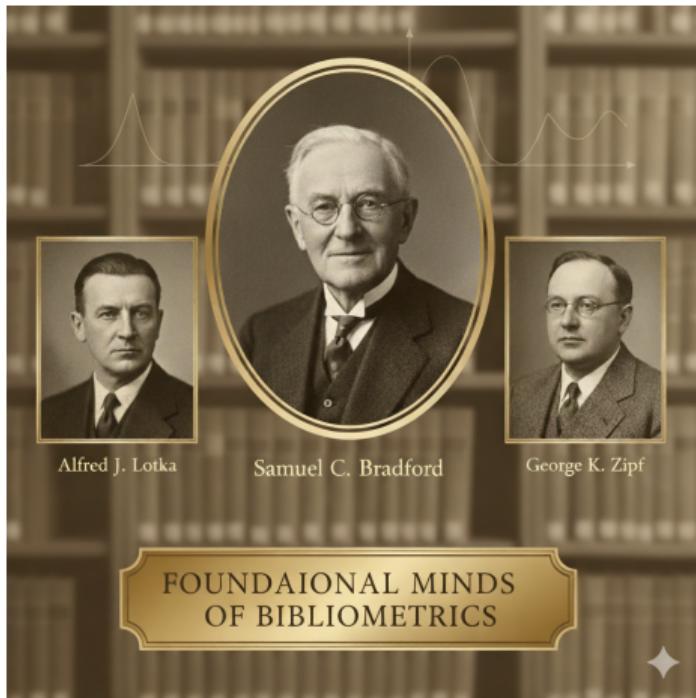
- ▶ **Impacto Ampliado:** Dados bem gerenciados e abertos são mais citados e reutilizados (potencial de +50% de aumento), ampliando o impacto do pesquisador e da instituição.
- ▶ **Otimização de Financiamento:** Evita a recoleta desnecessária de dados.

O Caos da Informação

A ciência produz dados em volume massivo. Sem padrões matemáticos, a gestão seria baseada em tentativas e erros, levando ao caos informacional.



Leis de Lotka, Bradford e Zipf



Alfred J. Lotka

Samuel C. Bradford

George K. Zipf

As Três Grandes Leis

Visão Geral

- ▶ **Lei de Lotka (1926):** Distribuição da produtividade dos autores. Poucos autores produzem a maioria dos artigos ([ALVARADO, 2002](#)).
- ▶ **Lei de Bradford (1934):** Distribuição da literatura científica em periódicos. Poucos periódicos concentram a maior parte dos artigos relevantes ([PINHEIRO, 1983](#)).
- ▶ **Lei de Zipf (1949):** Distribuição da frequência das palavras. Poucas palavras são usadas com muita frequência, enquanto muitas palavras são usadas raramente. ([BORTOLOSSI et al., 2011](#); [PINHEIRO; ALMEIDA, 2020](#))

Lei de Lotka

Produtividade dos Autores

A elite científica vs. a massa. O número de autores que publicam n trabalhos é aproximadamente $1/n^2$ daqueles que publicam apenas um ([ALVARADO, 2002](#)).

Exemplo: Para cada 100 autores com 1 artigo, apenas cerca de 25 terão 2 artigos ($100/2^2$), e apenas cerca de 11 terão 3 artigos ($100/3^2$)

Figura 1. Relação inversa entre o número de autores e sua produtividade, onde poucos autores publicam muito e muitos autores publicam pouco, seguindo uma distribuição do tipo potência.



Lei de Bradford

Dispersão da Literatura

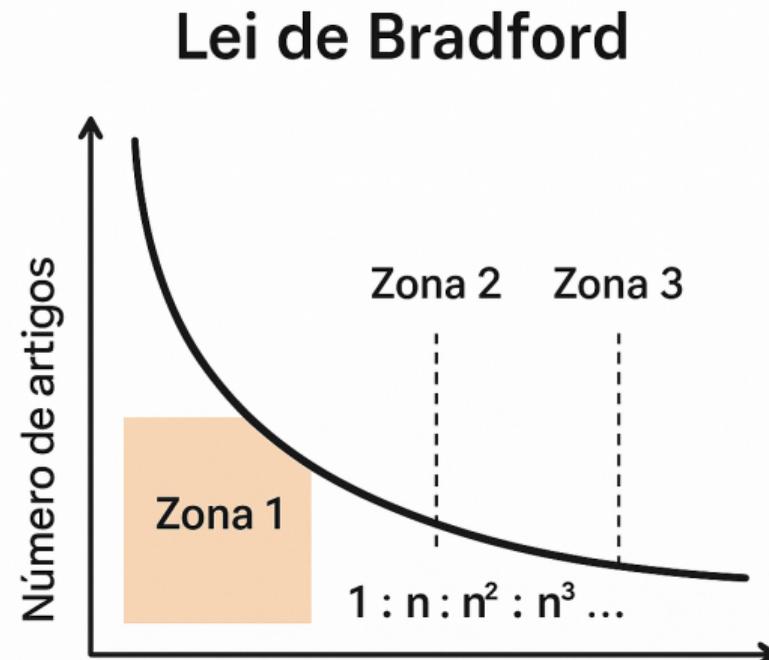
Se você tem revistas sobre um tema, um pequeno grupo de revistas (o núcleo) contém a maior parte dos artigos relevantes ([NARANAN, 1970](#); [BROOKES, 1985](#)).

É possível dividi-los em **zonas** contendo aproximadamente o mesmo número de artigos, mas com quantidades crescentes de periódicos tendendo a proporções próximas a $(k, kn, kn^2, kn^3, \dots)$.

- ▶ A **primeira (núcleo)** contém poucos periódicos altamente produtivos.
- ▶ A **segunda** contém mais periódicos que, juntos, publicam número semelhante de artigos.
- ▶ A **terceira** contém ainda mais periódicos, mas juntos publicam o mesmo número de artigos que as zonas anteriores.
- ▶ E assim sucessivamente.

Visualizando a Lei de Bradford

Figura 2. Distribuição da literatura em zonas de produtividade, onde poucos periódicos concentram muitos artigos e zonas sucessivas crescem na proporção (k, kn, kn^2, kn^3, \dots).



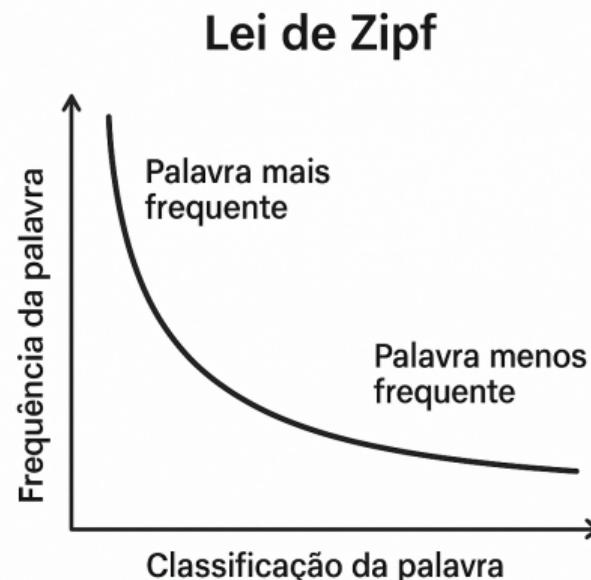
Lei de Zipf

Frequência de Palavras

Em um texto, a frequência de uma palavra é inversamente proporcional ao seu ranking. A palavra mais comum aparece o dobro da segunda, o triplo da terceira, etc. ([BORTOLOSSI et al., 2011](#); [PINHEIRO; ALMEIDA, 2020](#))

Equação: $f(r) = \frac{k}{r^\alpha}$, onde $f(r)$ é a frequência da palavra na posição r , k é uma constante e α é geralmente próximo de 1.

Figura 3. Relação inversa entre a frequência de uma palavra e sua posição no ranking, onde poucas palavras são muito frequentes e a maioria ocorre raramente, seguindo uma distribuição do tipo potência.



Intervalo de 15 minutos

Aplicando as Leis na Análise da Produção Científica na Prática

- ▶ **Python:** linguagem de programação de alto nível, versátil e amplamente utilizada em ciência de dados, machine learning e automação, com sintaxe clara e grande ecossistema de bibliotecas.
- ▶ **Dataset:** conjunto estruturado de dados, normalmente organizado em linhas (registros) e colunas (variáveis), podendo vir em formatos como CSV, Excel, JSON ou SQL.
- ▶ **pandas:** biblioteca do Python para manipulação eficiente de dados tabulares, oferecendo estruturas como *DataFrame* e *Series* para importar, limpar, transformar e analisar datasets.
- ▶ **matplotlib.pyplot:** módulo da biblioteca Matplotlib para criação de gráficos (linhas, barras, histogramas e dispersão).
- ▶ **Google Colab:** ambiente online gratuito fornecido pelo Google que permite executar códigos Python na nuvem, facilitando experimentos, compartilhamento e análise de dados.

- ▶ **Dataset:** Disponibilizado no repositório do GitHub (<https://github.com/usuario/repositorio>).
- ▶ **Google Colab:** Acesse via <https://colab.research.google.com/> e faça login com sua conta Google.

ATIVIDADE PRÁTICA

Atividade Prática

Analisando um Dataset de Produção Científica

Objetivo: Aplicar as Leis de Bradford, Lotka e Zipf para analisar um dataset real de produção científica.

Passos:

1. Carregar o dataset no Google Colab usando pandas.
2. Limpar e preparar os dados (remover duplicatas, lidar com valores ausentes).
3. Aplicar a Lei de Bradford para identificar os periódicos mais produtivos.
4. Aplicar a Lei de Lotka para analisar a produtividade dos autores.
5. Aplicar a Lei de Zipf para examinar a frequência das palavras nos títulos dos artigos.
6. Visualizar os resultados com gráficos usando matplotlib.pyplot.

Duração: 45 minutos.

- ▶ **Leitura 1:** As Leis da Bibliometria em Diferentes Bases de Dados Científicos
<http://dx.doi.org/10.5007/2175-8077.2016v18n44p111>
- ▶ **Leitura 2:** How to design bibliometric research: an overview and a framework proposal
<https://doi.org/10.1007/s11846-024-00738-0>
- ▶ **Leitura 3:** A relational database for bibliometric analysis
<https://doi.org/10.1016/j.joi.2010.06.007>

Resumo da Aula

Nessa aula estudamos os principais conceitos relacionados à Gestão de Dados Científicos (GDC) e as Leis de Bradford, Lotka e Zipf. Abordamos:

- ▶ Gestão de Dados Científicos: importância, ciclo de vida e princípios FAIR.
- ▶ Relação entre GDC e Ciência Aberta.
- ▶ O caos da informação e a necessidade de leis matemáticas.
- ▶ As Três Grandes Leis: Lotka, Bradford e Zipf.
- ▶ Fundamentos da comunicação científica e indicadores bibliométricos.
- ▶ Aplicações práticas usando Python e Google Colab.

As Leis da Produção Científica: Organizando o Caos da Informação



Lei de Lotka:
Poucos pesquisadores de elite produzem a maioria dos artigos científicos em uma área.

Poucos pesquisadores de elite produzem a maioria dos artigos científicos em uma área.



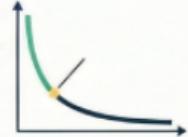
Lei de Bradford:
A dispersão da literatura a maior parte dos artigos relevantes sobre um tema.

Um pequeno núcleo de periódicos concentra a maior parte dos artigos relevantes sobre um tema.



Lei de Zipf:
A frequência das palavras com muita frequência, enquanto a maioria é rara.

Em qualquer texto, poucas palavras são usadas com muita frequência, enquanto a maioria é rara.



Referências

- ALVARADO, R. U. **A Lei de Lotka na bibliometria brasileira.** *Ciência da Informação*, SciELO Brasil, v. 31, p. 14–20, 2002. doi:[10.1590/S0100-19652002000200002](https://doi.org/10.1590/S0100-19652002000200002).
- BORTOLOSSI, H. J.; QUEIROZ, J. J. D. B.; SILVA, M. M. da. **A Lei de Zipf e Outras Leis de Potência em Dados Empíricos.** 2011.
- BROOKES, B. "Sources of information on specific subjects by sc bradford. *Journal of information science*, Sage Publications Sage CA: Thousand Oaks, CA, v. 10, n. 4, p. 173–175, 1985. doi:[10.1177/016555158501000406](https://doi.org/10.1177/016555158501000406).
- NARANAN, S. **Bradford's law of bibliography of science: an interpretation.** *Nature*, Nature Publishing Group UK London, v. 227, n. 5258, p. 631–632, 1970. doi:[10.1038/227631a0](https://doi.org/10.1038/227631a0).
- PINHEIRO, L. V. R. **Lei de Bradford: uma reformulação conceitual.** Ibit, 1983.
- PINHEIRO, R. G.; ALMEIDA, B. d. **As estratégias de internacionalização: um estudo bibliométrico aplicando as leis de Lotka, Bradford e Zipf na base SPELL no período de 2008 A 2018.** *Revista de Administração, Contabilidade e Economia da Fundace, Ribeirão Preto*, v. 11, n. 1, p. 60–79, 2020. doi:[10.13059/RACEF.V11I1.656](https://doi.org/10.13059/RACEF.V11I1.656).