GESTÃO DO PORTFÓLIO DE PRODUTOS DA EMBRAPA: aplicação de text mining como apoio para a análise de soluções tecnológicas agrícolas

Daniela Maciel Pinto
Mestre em Ciência da Informação. Doutoranda no Programa de Política Científica e Tecnológica da Unicamp, Campinas -SP, Brasil.
Analista, Empresa Brasileira de Pesquisa Agropecuária (Embrapa Territorial, Campinas, SP, Brasil).
Lattes: http://lattes.cnpq.br/9196370833242212
E-mail: daniela.maciel@embrapa.br

Resumo
O processo de Transferência de Tecnologias (TT) na Embrapa reúne um conjunto de ações, dentre as quais a organização de atividades focadas no fomento à geração de novas soluções tecnológicas e sua disponibilização aos agentes multiplicadores. Considerando a dinâmica existente nas unidades da Empresa, distribuídas pelo território, faz-se necessário, para uma boa orientação do processo de TT à Pesquisa e Desenvolvimento (P&D), o amplo conhecimento do conjunto de tecnologias já desenvolvidas. Assim, para amparar tais ações, busca-se realizar uma investigação na base de dados "Sistema de Gestão das Soluções Tecnológicas da Embrapa" (Gestec) a fim de: 1. identificar os principais tópicos e temas associados às tecnologias e 2. identificar a similaridade entre soluções tecnológicas. Para isso, será realizado o processamento de linguagem natural por meio do pacote TopicModels.

Palavras-chave: Agricultura, Gestão de portfólio, Transferência de Tecnologia, Modelagem de tópicos, Processamento de Linguagem Natural (PNL).

Introdução

A demanda global por produção agrícola está prevista para crescer significativamente até 2050, oferecendo uma oportunidade para o Brasil, dada sua vocação agrícola (FAO, 2018, 2014; United Nations, 2015). No entanto, para aproveitar essa oportunidade, serão necessárias mudanças estratégicas no campo da Ciência e Tecnologia (C&T) para garantir uma produção eficiente e sustentável de alimentos. Nesse sentido, as instituições de Pesquisa, Desenvolvimento e Inovação (PD&I) agrícola, como a Embrapa, precisam adotar uma gestão estratégica de inovação semelhante ao setor industrial (The Economist, 2016). Isso envolve a utilização de dados e informações para atender às demandas de pesquisa cada vez mais orientada para a racionalização e otimização de recursos.

No VII Plano Diretor (PDE), a Embrapa estabeleceu 11 objetivos e 29 metas estratégicas a serem alcançadas até 2030 (EMBRAPA, 2020). Para alcançá-los, é

essencial gerenciar de forma adequada os recursos internos e externos (Penrose, 2009; Teece, 2006, 2007) e posicionar os conhecimentos desenvolvidos ao longo do tempo como um diferencial competitivo (Tidd et al., 2015). Nesse contexto, a gestão estratégica da inovação é vista como um conjunto de ações sistêmicas e dinâmicas de aprendizado, integradas a diferentes áreas relacionadas à PD&I agrícola (Freeman & Soete, 2008; Kline & Rosenberg, 1986; Nelson, 1993; Quadros, 2008).

Como um processo dinâmico e em constante evolução, a gestão estratégica da inovação busca constantemente as melhores alternativas, práticas e técnicas existentes no mercado para apoiar o planejamento organizacional na busca de seus objetivos e metas. Com o aumento da importância do Big Data para a inovação estratégica (Cui, 2020; Muhammad et al., 2022), este trabalho visa aprimorar a gestão estratégica da inovação na Embrapa, utilizando técnicas de análise de dados não supervisionada aplicadas ao portfólio de produtos "Sistema de Gestão das Soluções Tecnológicas (Gestec)" por meio de processamento de linguagem natural e mineração de dados textuais (Silge & Robinson, 2017). O objetivo é identificar similaridades entre as soluções tecnológicas do Gestec e alinhá-las ao PDE atual.

Métodos

Este estudo descritivo-exploratório emprega o mapeamento sistemático na análise não supervisionada de dados textuais, visando identificar semelhanças nas soluções tecnológicas da Embrapa (Cervo et al., 2006). O processo seguiu duas etapas:

1. Definição do corpus e mineração textual: O corpus textual foi estruturado a partir de dados do Gestec, utilizando-se a variável "Descrição da Solução Tecnológica". As técnicas de limpeza e organização de dados foram realizadas através de data wrangling (Wickham, 2022), além de técnicas de text mining (Silge & Robinson, 2017) e ferramentas e pacotes de software.
2. Modelagem de tópicos e identificação de similaridade: realizou-se a análise de similaridade entre as soluções tecnológicas, conforme orientações de Grun e Hornik (Grün & Hornik, 2011) e Silge e Robinson (2017). A partir dos resultados, uma análise manual associou as soluções aos objetivos estratégicos do VII PDE. Para isso, foram utilizados o processamento de linguagem natural, modelagem de tópicos e outros pacotes de software.

Os instrumentos utilizados foram a linguagem R, o ambiente R Studio e os pacotes: "readxl", "topicmodels", "caret", "tidyr", "ggplot2", "stringr", "NLP", "curl", "tidytext", "wordcloud", "dplyr", "SnowballC" e "RColorBrewer".

Resultados

A base do Gestec continha 5 mil registros que, após limpeza e verificação de duplicidades, foi reduzida para 4.231 soluções tecnológicas. Os dados foram

classificados em 11 variáveis[1]. Dados anteriores à criação da Embrapa, embora representando apenas 2%, foram mantidos por ainda estarem ativos. 82% das tecnologias foram geradas nas primeiras duas décadas de 2000, 14% entre 1980-1990 e 2% na década de 1970 (Figura 1).

Principais Temas

As tecnologias abordam uma grande variedade de temas, como produção, solo, cultivares, sistemas de produção, sementes, regiões e qualidade produtiva. A análise textual (Nuvem de Palavras - Figura 2) identificou as palavras de maior frequência no *corpus* analisado, enquanto a Análise Fatorial de Correspondência (AFC) (Ratinaud, 2009) mostrou três grupos temáticos (Figura 3): Produção Animal (Vermelho); Produção Vegetal (Azul) e manejo produtivo/sistema de produção (Verde).

3.1 Similaridades entre as soluções tecnológicas, com vistas ao atual PDE

Com base em Grun e Hornik (2011), buscou-se identificar a similaridade entre as soluções tecnológicas existentes no Gestec por meio da técnica de Modelagem por Tópicos. Assim, a partir de uma matriz de frequência de termos, foram criados 15 tópicos, os quais foram classificados para, posteriormente, serem relacionados com os objetivos finalísticos do atual PDE. Após uma análise da base de dados, classificada por grupo de tópico, foi possível identificar tecnologias mais semelhantes tematicamente e, ainda, unidades da Embrapa com mais proximidade entre as soluções por cada tópico, possibilitando ações que se configuram pela complementaridade ou a evolução da solução tecnológica.

Com essa perspectiva, procedeu-se à análise das categorias, tendo em vista os oito objetivos finalísticos do atual PDE, criando-se a tabela 1, a qual demonstra as unidades com maior representatividade (quantidade de tecnologias) em cada um dos tópicos e o objetivo finalístico mais atinente ao tópico e à categoria. É necessário esclarecer que a quantidade de tópicos criada representou um número arbitrário, podendo ser substituída por qualquer número de grupos de tecnologias que se busca

[1] 1. Unidade Responsável; 2. Tema; 3. Nome da Solução Tecnológica; 4. Descrição da Solução Tecnológica; 5. Tipo de Solução Tecnológica; 6. Categoria da Solução Tecnológica; 7. Ano de Lançamento; 8. Estágio de Desenvolvimento; 9. Situação para Negócio; 10. Situação da Propriedade Intelectual; 11. Descontinuada.

constituir, ou mesmo a partir de algoritmos que auxiliam a definição de agrupamentos, como o método Elbow (Humaira; Rasydah, 2020).

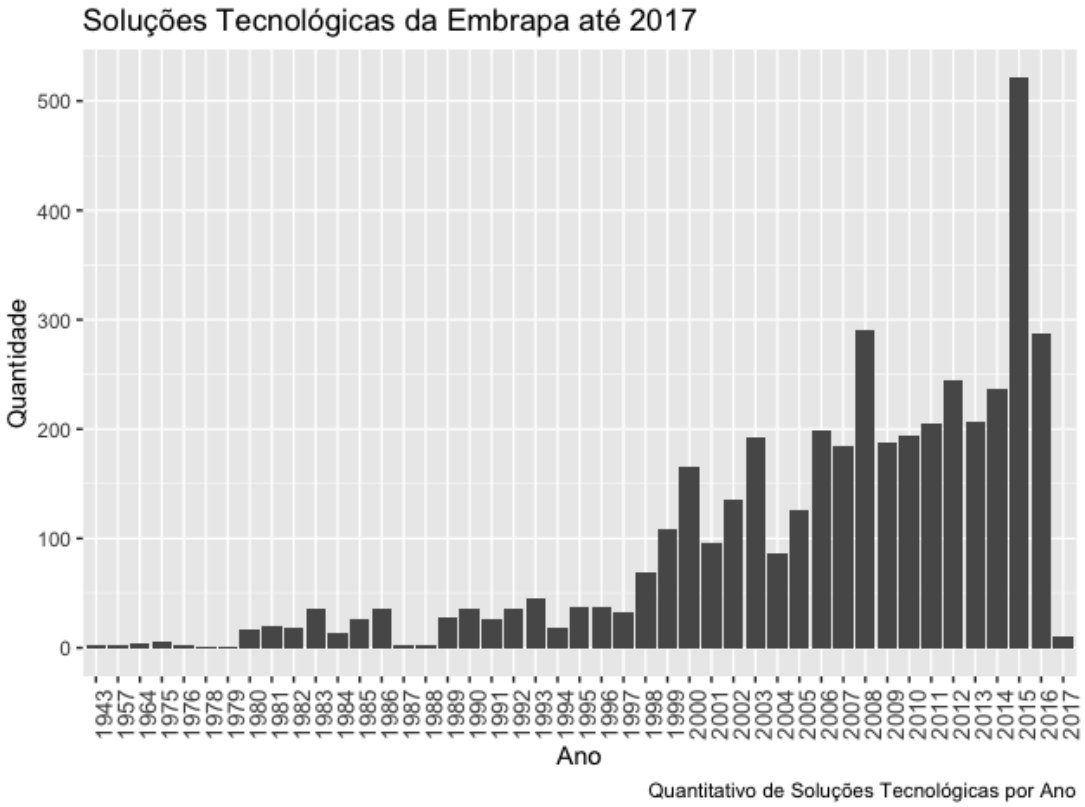

Figura 1. Relação das soluções tecnológicas após processamento.
Fonte: Elaborado pela autora.

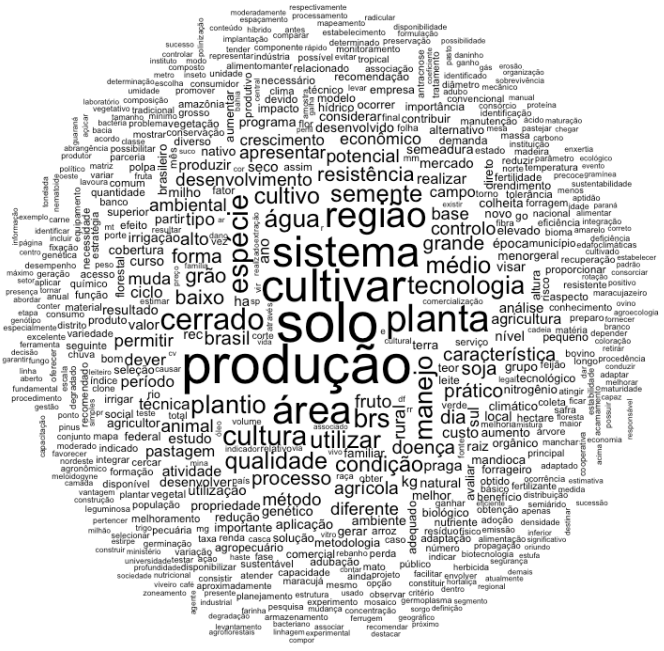

Figura 2. Principais palavras relacionadas às soluções tecnológicas da Embrapa

Fonte: Elaborado pela autora com uso do Iramuteq.

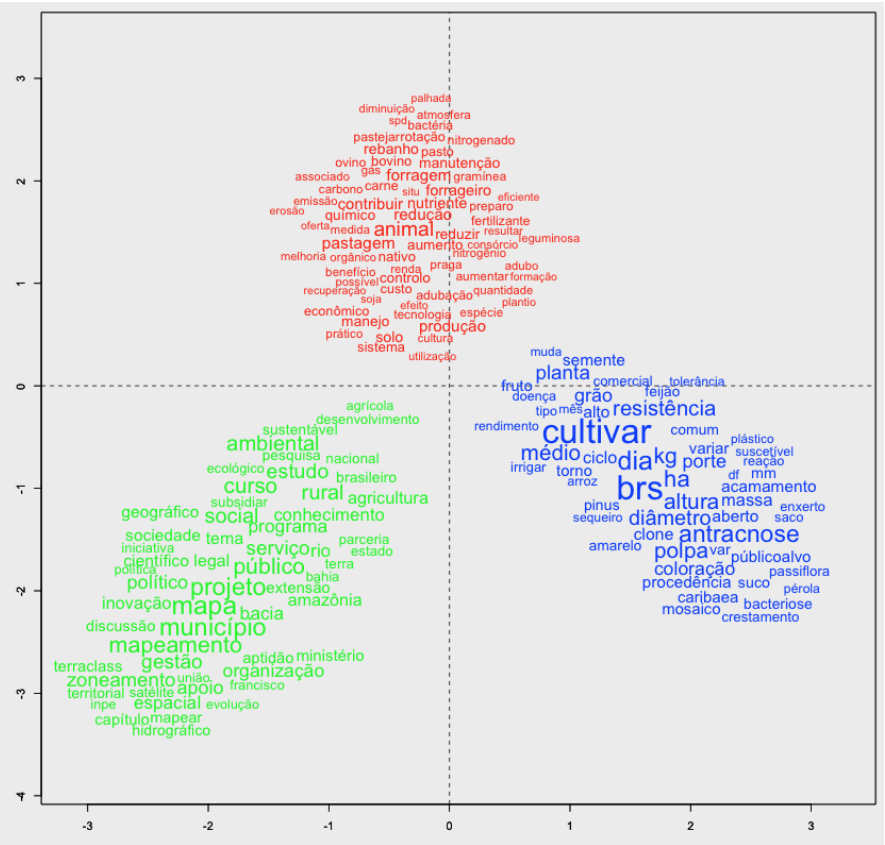

Figura 3. Análise fatorial de correspondência (AFC) das palavras mais frequentes
Fonte: Elaborado pela autora com uso do Iramuteq.

Considerações Finais

A gestão dinâmica de portfólio de produtos tem se beneficiado da análise de grandes volumes de dados. Neste trabalho, aplicamos o processamento de linguagem natural em análises não supervisionadas dos dados do Gestec para entender as relações temáticas entre as soluções tecnológicas da Embrapa. Observamos um foco na sustentabilidade, com temas como "produção", "solos" e "água" frequentemente presentes. Contudo, não identificamos relação com o objetivo "Automação de processos, agricultura de precisão e digital", sugerindo a necessidade de maior investimento nessa direção. Essa investigação preliminar, que apresenta limitações, requer validação de especialistas da área de Transferência de Tecnologias da Embrapa para confirmar os resultados.

Referências

Cervo, A. L., Bervian, A., & Silva, R. (2006). *Metodologia científica. São Paulo: Pearson Prentice Hall*.

Cui, M. (2020). Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, *1*(1), Artigo 1.

EMBRAPA. (2020). *VII Plano Diretor da Embrapa 2020-2030.* http://www.infoteca.cnptia.embrapa.br/handle/doc/1126091

FAO. (2018). *The future of Food and Agriculture: Alternative pathways to 2050*. https://knowledge4policy.ec.europa.eu/publication/future-food-agriculture-alternative-pathways-2050_en

FAO, T. (2014). The state of food and agriculture: Innovation in family farming. *Rome FAO*.

Freeman, C., & Soete, L. (2008). *A Economia da inovação industrial*. Editora da UNICAMP.

Grün, B., & Hornik, K. (2011). **topicmodels**: An *R* Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13), Artigo 13. https://doi.org/10.18637/jss.v040.i13

Humaira, H., & Rasyidah, R. (2020). Determining the appropiate cluster number using Elbow method for K-Means algorithm. *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA)*.

Kline, S. J., & Rosenberg, N. (1986). An Overview of Innovation. Em N. Rosenberg, *Studies on Science and the Innovation Process* (p. 173–203). WORLD SCIENTIFIC. http://www.worldscientific.com/doi/abs/10.1142/9789814273596_0009

Muhammad, A., Yu, C. K., Qadir, A., Ahmed, W., Yousuf, Z., & Fan, G. (2022). Big data analytics capability as a major antecedent of firm innovation performance. *The International Journal of Entrepreneurship and Innovation*, *23*(4), Artigo 4. https://doi.org/10.1177/14657503211050809

Nelson, R. R. (1993). *National Innovation Systems: A Comparative Analysis* (SSRN Scholarly Paper 1496195; Número 1496195). https://papers.ssrn.com/abstract=1496195

Penrose, E. T. (2009). *The Theory of the Growth of the Firm*. Oxford university press.

Quadros, R. (2008). *Aprendendo a Inovar: Padrões de Gestão da Inovação Tecnológica em Empresas Industriais Brasileiras*. https://silo.tips/download/campinas-agosto-de-2008

Ratinaud, P. (2009). *Iramuteq—Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires [Computer software]*. http://www.iramuteq.org/

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach* (First edition). O'Reilly.

Teece, D. J. (2006). Reflections on "Profiting from Innovation". *Research Policy*, *35*(8), Artigo 8. https://doi.org/10.1016/j.respol.2006.09.009

Teece, D. J. (2007). Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance. *Strategic Management Journal*, *28*(13), Artigo 13. https://doi.org/10.1002/smj.640

The Economist. (2016, junho 11). *The future of agriculture | Jun 11th 2016*. The Economist. https://www.economist.com/technology-quarterly/2016-06-11

Tidd, J., Bessant, J., Nonnenmacher, F., & Matte, G. A. (2015). *Gestão da Inovação* (5ª edição). Bookman.

United Nations. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development*.

Wickham, H. (2022). *rvest: Easily Harvest (Scrape) Web Pages* (1.0.3) [Software]. https://CRAN.R-project.org/package=rvest

latmétricas

Tabela 1. Relação dos grupos de tópicos, categorizados, com os objetivos finalísticos do VII PDE.

Tópico	Classificação	Termos/Temas	Qtd Soluções	Unidades	Objetivo Finalístico
1	Meio Ambiente e Preservação do Solo	solo, ambient, efeito, área, manejo, anális, praga, carbono, campo, sistema	320	Embrapa Cerrados, Embrapa Solos, Embrapa Florestas	7. Enfrentamento de mudança do clima na agropecuária; 5. Biomassa, resíduos, bioinsumos e energia renovável
2	Capacitações e manejo de sistemas produtivos leiteiro	produção, sistema, curso, leit, manejo, técnica, tecnologia, agricultura, instituiçõ, anális	296	Embrapa Cerrados, Embrapa Tabuleiros, Embrapa Gado de Leite	3. Novas tendências de consumo e agregação de valor
3	Sistemas integrados de produção, controle de doenças e pragas de espécies nativas do Cerrado	sistema, produção, espéci, control, cerrado, muda, cultivo, praga, área, sement	322	Embrapa Cerrados, Embrapa Agropecuária Oeste, Embrapa Amazônia Ocidental	1. Sustentabilidade e competitividade
4	Cultivares de grãos (arroz, feijão, milho) de ciclo curto e resistentes	cultivar, brs, grãos, ciclo, resistência, produção, dia, média, qualidad, arroz	369	Embrapa Arroz e Feijão, Embrapa Cerradas, Embrapa Amazônia Ocidental	1. Sustentabilidade e competitividade
5	Produção, manejo e beneficiamento de frutos	produção, fruto, manejo, qualidad, sistema, planta, pode, prática, área, renda	211	Embrapa Cerrados, Embrapa Agroindústria Tropical, Embrapa Recursos Genéticos	1. Sustentabilidade e competitividade
6	Controle de doenças do maracujá e produção de mudas	cultivar, doença, planta, produção, muda, sement, maracujazeiro, flore, antracnos, cerrado	214	Embrapa Cerrados, Embrapa Amazônia Oriental e Embrapa Hortaliças	4. Segurança e defesa zoofitossanitária
7	Produção de sementes e cultivares	sement, cultivar, apó, dia, raíz, polpa, germinação, plantio, devem, teor	315	Embrapa Cerrados, Embrapa Amazônia Oriental, Embrapa Agroindústria de Alimentos	1. Sustentabilidade e competitividade
8	Produção de Mandioca	produção, mandioca, modelo, área, cultura, genético, utilizado, fase, part, método	292	Embrapa Cerrados, Embrapa Recursos Genéticos, Embrapa Meio Ambiente	1. Sustentabilidade e competitividade
9	Produção agrícola no Cerrado	espéci, produção, cerrado, crescimento, solo, planta, cultivo, doi, sistema, área	196	Embrapa Cerrados, Embrapa Tabuleiros Costeiros, Embrapa Recursos Genéticos	1. Sustentabilidade e competitividade
10	Preservação da vegetação e agricultura nos Bioma (Amazônia e Cerrado)	área, vegetação, solo, amazônia, bioma, município, cerrado, região, mapeamento, terra	175	Embrapa Cerrados, Embrapa Semiárido, Embrapa Solos	6. Desenvolvimento regional sustentável e inclusão produtiva
11	Produção e alimentação animal	produção, animai, forragem, pastagen, forrageira, anim, rebanho, pastejo, sistema, pasto	397	Embrapa Cerrados, Embrapa Pecuária Sul, Embrapa Caprinos e Ovinos	1. Sustentabilidade e competitividade
12	Fixação biológica do Nitrogênio	nitrogênio, fixação, biológica, ambient, planta, ambientai, serviço, desenvolvimento, inoculant, solo	244	Embrapa Cerrados, Embrapa Meio Ambiente, Embrapa Caprinos e Ovinos	1. Sustentabilidade e competitividade
13	Sistema plantio direto (soja e milho) e preservação do solo e da água no Cerrado	solo, sistema, plantio, cultura, agrícola, milho, direto, soja, cerrado, água	266	Embrapa Cerrados, Embrapa Clima Temperado, Embrapa Arroz e Feijão	1. Sustentabilidade e competitividade
14	Sistema produtivo, cultivares de soja e controle do nematóide	rec, soja, resistência, cultivar, nematoid, brs, região, regiõ, crescimento, grupo	298	Embrapa Cerrados, Embrapa Soja, Embrapa Recursos Genéticos	1. Sustentabilidade e competitividade
15	Produção agrícola e preservação dos recursos naturais	água, solo, irrigação, cultura, planta, método, cultivo, chuva, base, área	316	Embrapa Cerrados, Embrapa Semiárido, Embrapa Tabuleiros Costeiros	2. Dados e informações dos recursos naturais; 9 - Racionalização de Recursos e Diversificação de Fontes