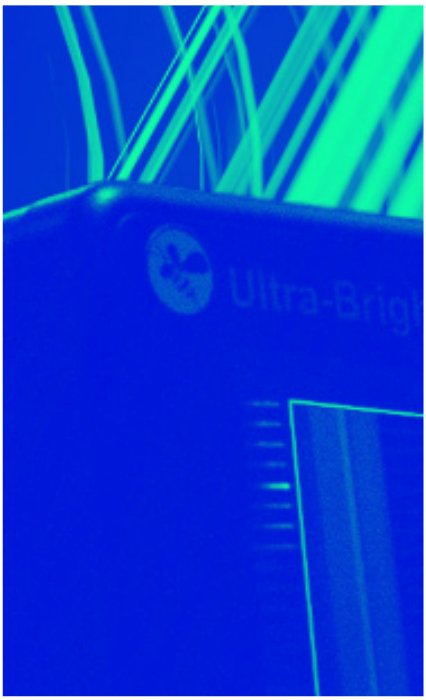
BOLETIM DE PESQUISA E DESENVOLVIMENTO
54

ISSN 2764-2623
Novembro / 2023

Mineração de textos para síntese de metodologias da Embrapa aplicadas à recuperação de pastagens degradadas



Embrapa

ISSN 2764-2623
Novembro/2023

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Agricultura Digital
Ministério da Agricultura e Pecuária

BOLETIM DE PESQUISA E DESENVOLVIMENTO 54

Mineração de textos para síntese de metodologias da Embrapa aplicadas à recuperação de pastagens degradadas

Maria Fernanda Moura
Clara Mattos Medeiros
Luis Eduardo Gonzales
Patrícia Menezes Santos

Embrapa Agricultura Digital
Campinas, SP
2023

Embrapa Agricultura Digital

Av. Dr. André Tosello, 209 - Cidade Universitária
Campinas, SP, Brasil
CEP. 13083-886
Fone: (19) 3211-5700
www.embrapa.br
www.embrapa.br/fale-conosco/sac

Revisão de texto
Adriana Farah Gonzalez

Normalização bibliográfica
Carla Cristiane Osawa

Projeto gráfico da coleção
Carlos Eduardo Felice Barbeiro

Editoração eletrônica
Magda Cruciol

Imagem da capa
Magda Cruciol

Publicação digital: PDF (2023)

Dados Internacionais de Catalogação na Publicação (CIP)
Nome da unidade catalogadora

Mineração de textos para síntese de metodologias da Embrapa aplicadas à recuperaçao de pastagens degradadas / Maria Fernanda Moura … [et al.]. – Campinas : Embrapa Agricultura Digital, 2023.
PDF (37 p.) : il. color. - (Boletim de pesquisa e desenvolvimento / Embrapa Agricultura Digital, ISSN 2764-2623 ; 54).

1. Mineração de texto. 2. Processamento de língua natural. 3. Recuperação de informação. 4. Recuperação de pastagem. I. Moura, Maria Fernanda. II. Embrapa Agricultura Digital. III. Série.

CDD (21. ed.) 006.33

Carla Cristiane Osawa (CRB-8/10421)

Sumário

Mineração de textos para síntese de metodologias da Embrapa aplicadas à recuperação de pastagens degradadas

Maria Fernanda Moura[1]
Clara Mattos Medeiros[2]
Luis Eduardo Gonzales[3]
Patrícia Menezes Santos[4]

Resumo – A Empresa Brasileira de Pesquisa Agropecuária (Embrapa) possui um rico acervo de dados, informações e conhecimentos agropecuários resultantes de suas pesquisas, tecnologias e publicações desenvolvidas desde a década de 70, incluindo aqueles sobre o processo de degradação de pastagens, suas causas e estratégias de recuperação. A análise desse material pode contribuir para acelerar o processo de desenvolvimento tecnológico e transferência de tecnologia e, consequentemente, para a recuperação de pastagens degradadas no campo. Neste trabalho foi adaptado um processo de mineração de textos com o objetivo de produzir uma síntese do mapeamento de metodologias, tecnologias, produtos, serviços e possíveis recomendações da Embrapa sobre recuperação de pastagens degradadas. O detalhamento do processo, desde a seleção da fonte de dados e delimitação do tema em publicações disponibilizadas pela Embrapa até a síntese em subtópicos de interesse, é apresentado, bem como critérios de escolha de ferramentas de software, como parametrizá-las e utilizá-las. O processo desenvolvido por especialistas em mineração de dados e no tema pastagens mostrou-se eficiente para sintetizar o material disponível em subtemas, tais como recomendações em recuperação de pastagens aplicadas a diferentes biomas (Amazônia, Cerrado, Caatinga) e em combinação com diferentes sistemas de produção.

[1] Estatística, doutora em Ciências da Computação, pesquisadora da Embrapa Agricultura Digital, Campinas, SP

[2] Estudante de Engenharia da Computação, bolsista da Embrapa Agricultura Digital, Campinas, SP

[3] Engenheiro da Computação, analista da Embrapa Agricultura Digital, Campinas, SP

[4] Engenheira-agrônoma, doutora em Agronomia, pesquisadora da Embrapa Pecuária Sudeste, São Carlos, SP

A síntese em subtemas facilita a recuperação das informações e a identificação das estratégias de recuperação de pastagens mais adequadas para um determinado contexto (local, solo, clima, sistema de produção, tipo de capim, etc), facilitando a adoção das tecnologias pelo setor produtivo.

Termos para indexação: extração de informação em textos, processamento de língua natural, agrupamentos textuais.

Text mining to synthesize Embrapa methodologies applied to the recovery of degraded pastures

Abstract – The Brazilian Agricultural Research Corporation (Embrapa) has a rich collection of data, information and agricultural knowledge resulting from its research, technologies and publications developed since the 1970s, including those on the pasture degradation process, its causes and recovery strategies. The analysis of this material can contribute to accelerate the process of technological development and technology transfer and, consequently, to the recovery of degraded pastures in the field. In this work, a text mining process was adapted with the objective of producing a synthesis of the mapping of methodologies, technologies, products, services and possible recommendations from Embrapa on recovery of degraded pastures. The detailing of the process, from the selection of the data source and delimitation of the theme in publications made available by Embrapa to the synthesis in sub-topics of interest, is presented, as well as criteria for choosing software tools, how to parameterize them and use them. The process developed by specialists in data mining and in the field of pastures proved to be efficient in synthesizing the material available in sub-topics, such as recommendations for recovering pastures applied to different biomes (Amazon, Cerrado, Caatinga) and in combination with different production systems. The synthesis in sub-themes facilitates the retrieval of information and the identification of the most appropriate pasture recovery strategies for a given context (location, soil, climate, production system, type of grass, etc.), facilitating the adoption of technologies by the productive sector.

Index terms: extraction of information in texts, natural language processing, textual clusters.

Introdução

A Embrapa possui um valioso acervo de dados, informações e conhecimentos agropecuários provenientes de suas pesquisas, tecnologias e publicações desenvolvidas desde os anos 70. Essas informações podem ser acessadas por meio de ferramentas de busca, como a Base de Dados da Pesquisa Agropecuária (BDPA) (Vacari et al., 2007), que permite recuperar obras cadastradas disponíveis nas bibliotecas da Embrapa, assim como aquelas produzidas exclusivamente pela instituição.

O projeto "Gestão da informação e do conhecimento como suporte à gestão estratégica do Portfólio de Pastagens" teve como um de seus objetivos mapear o conhecimento, as informações e os dados gerados sobre tecnologias e metodologias aplicadas ao tema "pastagens degradadas" utilizando a informação existente na BDPA. Esse mapeamento visava identificar tendências em relação a metodologias e tecnologias desenvolvidas pela Embrapa, considerando aspectos como regiões geográficas de aplicação, temas e subtemas de cobertura e época de desenvolvimento das tecnologias e metodologias. Para realizar essa análise de tendências, foram utilizados filtros de busca e extração de informação, com processamento de linguagem natural, sobre os resultados fornecidos pela BDPA e aprendizado de máquina para auxiliar a sintetização desses resultados.

A informação de cobertura geográfica, ou seja, locais onde se possa aplicar uma tecnologia, não é claramente representada nos metadados na BDPA. Muitas vezes, ela se encontra apenas no contexto da obra, quando se descreve o uso ou desenvolvimento da tecnologia ou metodologia desenvolvida. Para obter essa informação foi necessário identificá-la no contexto das publicações técnico-científicas, em língua portuguesa, por meio de um processo de reconhecimento de entidades nomeadas e sua tipificação – local, instituição, nome próprio, etc. No trabalho de Moura e Medeiros (2022), após vários testes de ferramentas desse tipo para a língua portuguesa, os autores propuseram um classificador utilizando o reconhecedor de entidades nomeadas da biblioteca SpaCy (Vasiliev, 2020), com conjuntos de treinamento próprios, que se mostrou bastante eficaz para identificar especificamente localizações geográficas brasileiras, tendo sido o escolhido para uso neste trabalho.

Outro ponto de interesse é a compreensão das temáticas dessas tecnologias ou metodologias e suas correlações, que poderiam ser derivadas das palavras-chaves que qualificam as obras. Porém, observou-se em uma primeira disponibilização de resultados por meio do mapeamento dos metadados correspondentes às palavras-chaves que, apesar de fornecerem uma visão geral muito boa, sua análise poderia ser facilitada se permitisse observar agrupamentos em temas mais concisos. Para obter essa concisão, decidiu-se utilizar um processo de aprendizado de máquina sobre informações de metadados e contexto das obras, a fim de obter os tópicos mais significativos aos quais os textos mutuamente pertencessem. Para o reconhecimento de tópicos em grandes coleções textuais ou mesmo a partir de seus metadados básicos, tais como título, redes de autorias, palavras-chaves, descrições simplificadas ou resumos, os processos mais utilizados são os probabilísticos (Blei, 2012; Churchill; Singh, 2022). No trabalho aqui apresentado, optou-se pela utilização do método Latent Dirichlet Allocation (LDA) (Blei et al., 2003), cujos resultados têm sido praticamente imbatíveis nos últimos anos, e que possui implementações de fácil integração a ferramentas de software em geral – em particular a implementação do ambiente Machine Learning for Language Toolkit (MALLET), como o *wrapper* disponível no GitHub[5].

Assim, neste trabalho foi adotada uma abordagem completa de mineração de texto adaptando uma metodologia proposta em trabalhos anteriores (Moura et al., 2017). O processo todo é descrito nas próximas seções, desde a recuperação dos dados via BDPA, método de pré-processamento de dados utilizado, sintetização em tópicos, análise de resultados e a disponibilização de uma visualização dos resultados obtidos, a partir do uso do Google Looker Studio, para fins de avaliação e validação.

O processo, desenvolvido por especialistas em mineração de dados e no tema de pastagens, mostrou-se eficiente para mapear de forma sintetizada o material disponível em subtemas de interesse, tais como recomendações para recuperação de pastagens em diferentes biomas e sistemas de pro-

[5] Módulo little-mallet-wrapper, disponível em: https://github.com/maria-antoniak/little-mallet-wrapper.

dução. Isso facilita a recomendação e a adoção das tecnologias pelo setor produtivo. Acredita-se que esse processo possa ser aplicado a outros temas, com ajustes e a participação de especialistas no assunto e em mineração de dados.

Material e métodos

A análise das obras da Embrapa referentes a pastagens degradadas foi realizada a partir de um processo de mineração de textos com uso de aprendizado não supervisionado, ilustrado na Figura 1. Nas próximas seções são descritas as etapas: 1) escolha das fontes de dados: como elaborar as expressões de busca para coleta dos dados; 2) pré-processamento: o que significa e como realizá-lo neste caso; 3) extração de padrões: escolha de métodos e suas configurações; e 4) avaliação: uma estratégia para a visualização dos resultados obtidos para validação junto aos especialistas.

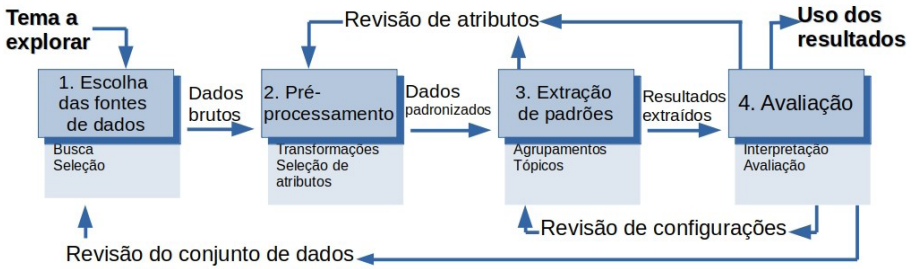


Figura 1. Processo de mineração de textos utilizado.

Escolha das fontes de dados

Os dados utilizados no problema abordado são de domínio público, dado que a grande maioria está disponível via BDPA. Essa base de dados permite a elaboração de consultas complexas, facilitando muito a filtragem de resultados. Por exemplo, para buscar informações sobre pastagens degradadas,

recuperação ou reforma destas incluindo forrageiras e sistemas silvopastoris, pode-se especificar expressão de busca tal como a ilustrada na Figura 2:

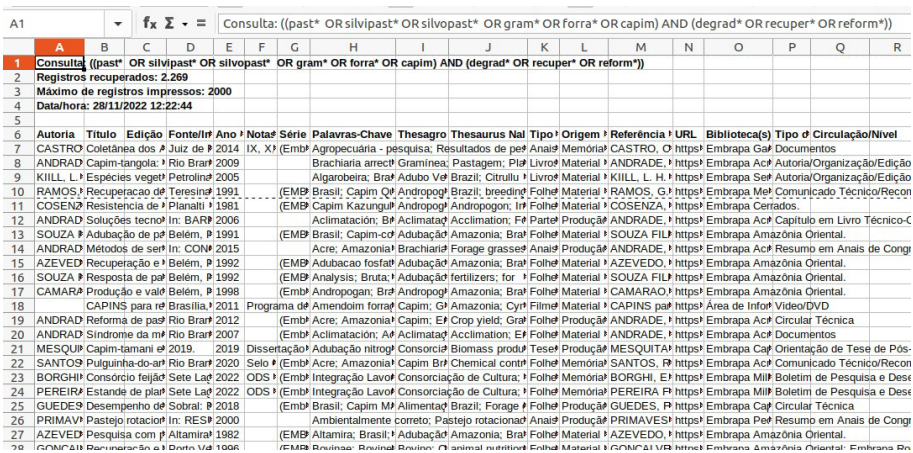
```
((past*  OR silvipast* OR silvopast*  OR gram* OR forra* OR capim)
                AND (degrad* OR recuper* OR reform*))
```

Figura 2. Exemplo de expressão de busca.

A consulta, cuja expressão é ilustrada na Figura 2, em novembro de 2022, resultava em 2.684 obras da Produção Científica da Embrapa, conforme ilustração na Figura 3, na parte esquerda. Aplicando-se o filtro para obras em língua portuguesa, reduz-se a 2.269 obras, conforme parte direita da Figura 2 indicada pela seta. Esse filtro é importante, pois mesmo os textos em português fazem referências a resumos em inglês e outras línguas, o que pode gerar ruídos no restante do processo.

Figura 3. Busca na Base de Dados da Pesquisa Agropecuária seguida de filtro.

Para armazenar os resultados, pode-se gerar uma planilha, ilustrada na Figura 4, selecionando-se a opção de imprimir registros no formato resumido, no canto direito do alto da tela de resultados de busca via BDPA. Note, na Figura 4, que, além dos resumos dos registros resultantes da busca, também são registrados a expressão de busca, o total de registros recuperados, o total de registros gravados e a data e horário da busca. Há dois problemas relativos a essa planilha: 1) o máximo armazenado é de 2.000 resultados e há 2.269 registros resultantes; e 2) embora mais metadados que os apresentados na busca sejam armazenados na planilha, ela não contém os resumos das obras, que seria uma informação necessária para melhorar o processo de mineração de textos. Dessa forma, para o projeto especificamente, os resultados com maior completude foram gerados pela equipe responsável pela base de dados – em lugar do uso direto da planilha resultante. De qualquer forma, para uso fora do domínio Embrapa, os metadados fornecidos e o fato de a planilha armazenar apenas 2.000 resultados não representam uma grande limitação – considerando-se uma simples replicação do processo aqui demonstrado.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Consulta: ((past* OR silvipast* OR silvopast* OR gram* OR forra* OR capim) AND (degrad* OR recuper* OR reform*))																	
2	Registros recuperados: 2.269																	
3	Máximo de registros impressos: 2000																	
4	Data/hora: 28/11/2022 12:22:44																	
5																		
6	Autoria	Título	Edição	Fonte/In	Ano	Notas	Série	Palavras-Chave	Thesagro	Thesaurus Nal	Tipo	Origem	Referência	URL	Biblioteca(s)	Tipo d	Circulação/Nivel	
7	CASTRO	Coletânea dos A	Juiz de F		2014	IX, XI	(Emb	Agropecuária - pesquisa; Resultados de pes		Anais	Memória	CASTRO, O	https	Embrapa Ga	Documentos			
8	ANDRAD	Capim-tangola:	Rio Bran		2009			Brachiaria arrecti	Gramínea;	Pastagem; Pla	Livro	Material	ANDRADE,	https	Embrapa Ac	Autoria/Organização/Edição		
9	KIILL, L.	Espécies veget	Petrolina		2005			Algarobeira; Bra	Adubo Ve	Brazil; Citrullu	Livro	Material	KIILL, L. H.	https	Embrapa Se	Autoria/Organização/Edição		
10	RAMOS,	Recuperacao de	Teresina		1991		(EMB	Brasil; Capim Qu	Andropog	Brazil; breeding	Folhe	Material	RAMOS, G	https	Embrapa Me	Comunicado Técnico/Recon		
11	COSENZ	Resistencia de	Planalti		1981		(EMB	Capim Kazungul	Andropog	Andropogon; In	Folhe	Material	COSENZA,	https	Embrapa	Cerrados.		
12	ANDRAD	Soluções tecno	In: BARB		2006			Aclimatação; Bi	Aclimata	Acclimation; Fe	Parte	Produção	ANDRADE,	https	Embrapa Ac	Capítulo em Livro Técnico-C		
13	SOUZA,	Adubação de pa	Belém, P		1991		(EMB	Brasil; Capim-co	Adubação	Amazonia; Bra	Folhe	Material	SOUZA FIL	https	Embrapa	Amazônia Oriental.		
14	ANDRAD	Métodos de ser	In: CONG		2015			Acre; Amazonia	Brachiaria	Forage grasses	Anais	Produção	ANDRADE,	https	Embrapa Ac	Resumo em Anais de Congr		
15	AZEVED	Recuperação e	Belém, P		1992		(EMB	Adubacao fosfat	Adubação	Amazonia; Bra	Folhe	Material	AZEVEDO,	https	Embrapa	Amazônia Oriental.		
16	SOUZA,	Resposta de pa	Belém, P		1992		(EMB	Analysis; Bruta;	Adubação	fertilizers; for	Folhe	Material	SOUZA FIL	https	Embrapa	Amazônia Oriental.		
17	CAMARA	Produção e valo	Belém, P		1998		(EMB	Andropogan; Bra	Andropog	Amazonia; Bra	Folhe	Material	CAMARAO,	https	Embrapa	Amazônia Oriental.		
18		CAPINS para re	Brasília,		2011	Programa de		Amendoim forra	Capim; G	Amazonia; Cyr	Filme	Material	CAPINS pa	https	Área de Infor	Vídeo/DVD		
19	ANDRAD	Reforma de pas	Rio Bran		2012		(EMB	Acre; Amazonia	Capim; Er	Crop yield; Gra	Folhe	Produção	ANDRADE,	https	Embrapa Ac	Circular Técnica		
20	ANDRAD	Síndrome da m	Rio Bran		2007		(EMB	Aclimatação; Ac	Aclimata	Acclimation; Er	Folhe	Material	ANDRADE,	https	Embrapa Ac	Documentos		
21	MESQUI	Capim-tamani e	2019.		2019	Dissertação		Adubação nitrog	Consorcia	Biomass produ	Tese	Produção	MESQUITA	https	Embrapa Ca	Orientação de Tese de Pós-		
22	SANTOS,	Pulguinha-do-ar	Rio Bran		2020	Selo P	(EMB	Acre; Amazonia	Capim Br	Chemical contr	Folhe	Memória	SANTOS, R	https	Embrapa Ac	Comunicado Técnico/Recon		
23	BORGHI,	Consórcio feijão	Sete Lag		2022	ODS 9	(EMB	Integração Lavo	Consorciação de Cultura;		Folhe	Memória	BORGHI, E	https	Embrapa Mil	Boletim de Pesquisa e Dese		
24	PEREIRA	Estande de plan	Sete Lag		2022	ODS 9	(EMB	Integração Lavo	Consorciação de Cultura;		Folhe	Memória	PEREIRA F	https	Embrapa Mil	Boletim de Pesquisa e Dese		
25	GUEDES	Desempenho de	Sobral: E		2018		(EMB	Brasil; Capim Mi	Alimentaç	Brazil; Forage p	Folhe	Produção	GUEDES, F	https	Embrapa Ca	Circular Técnica		
26	PRIMAV	Pastejo rotacion	In: RESU		2000			Ambientalmente correto; Pastejo rotacionad			Anais	Produção	PRIMAVES	https	Embrapa Pe	Resumo em Anais de Congr		
27	AZEVED	Pesquisa com p	Altamira		1982		(EMB	Altamira; Brasil;	Adubação	Amazonia; Bra	Folhe	Material	AZEVEDO,	https	Embrapa	Amazônia Oriental.		
28	GONÇALV	Recuperação e	Porto Ve		1996		(EMB	Bovinae; Bovinos	Bovino; C	Animal nutrition	Folhe	Material	GONÇALVE	https	Embrapa Amazônia Oriental; Embrapa Ro			

Figura 4. Parte da planilha resultante da gravação do carrinho de resultados da Base de Dados da Pesquisa Agropecuária (BDPA).

Pré-processamento

Para qualquer tarefa de análise de dados há sempre uma etapa de pré-processamento. Quando os dados são numéricos pode-se utilizar alguma transformação dos valores, seja para colocá-los em uma mesma escala ou outras padronizações necessárias. No caso de dados textuais há alguns cuidados básicos exigidos pelas ferramentas de mineração de textos e de busca, e outros mais específicos quando procuramos por informações ou inferências. Como os dados são textuais e os resultados gerados levam a combinações de vocábulos, deve-se procurar uma forma legível (por máquina e humano) de padronizar esses vocábulos, ou expressões compostas por eles. Outra preocupação é com o tipo de modelo de aprendizado de máquina que será aplicado aos dados, pois cada modelo possui suas exigências de padronização de dados para entrada e saída.

Para obter tópicos a partir de uma coleção de textos, em geral, os modelos utilizados são probabilísticos e trabalham com os vocábulos de cada texto como uma *bag of words* (BoW), limitando-se apenas ao léxico e impossibilitando a distinção de documentos com vocabulário semelhante e ideias diferentes sobre um mesmo assunto; isto é, não há preocupação com a sequência em que os vocábulos são apresentados nos textos. Nesse tipo de representação de dados, a probabilidade dos vocábulos ocorrerem em determinada sequência não é levada em consideração. Por exemplo: “estado de Alagoas”, “estado Alagoas de”, “Alagoas de estado” e “de Alagoas estado” são interpretados como três vocábulos probabilisticamente independentes pelo conjunto de vocábulos [estado, de, Alagoas]. Uma coleção qualquer de textos será representada por uma matriz, na qual cada linha corresponde a um texto, cada coluna corresponde a um vocábulo e cada valor corresponde a uma medida – que pode ser frequência absoluta de ocorrência do vocábulo no texto, ou 0 (ausente) e 1 (presente), ou outras medidas de interesse, que dependem do algoritmo de aprendizado a ser utilizado. Dessa forma, a seleção de atributos, isto é, do conjunto de vocábulos (ou combinações deles) a ser utilizado consiste basicamente em reduzir o número de vocábulos utilizados, procurando pelos mais estatisticamente significativos e mais humanamente interpretáveis (escolha subjetiva), dado que o número de vocábulos é muito grande e a matriz pode, além de ter altíssima dimensionalidade, ser altamente esparsa - o que pode introduzir mais ruídos no modelo de aprendizado utilizado.

Dada a planilha de dados resultante da busca anteriormente realizada, o que se deve fazer é selecionar os metadados (colunas da planilha) de interesse e considerar cada obra (no caso, linha da planilha) como o texto de interesse. Essa primeira seleção de obras e seus descritores (metadados) resulta na coleção de documentos a ser utilizada no processo - *córpus*.

A seguir, aplica-se o primeiro processo de padronização e seleção de atributos, que são os vocábulos presentes em cada célula (linha e coluna) da planilha. Uma forma de selecionar os vocábulos (que serão os atributos do modelo de aprendizado) é utilizar uma ferramenta de processamento de língua natural que permita identificar e rotular cada vocábulo em elementos gramaticais, tais como nomes, verbos, adjetivos, advérbios, etc. Essas ferramentas normalmente utilizam processos denominados *part-of-speech tagging* (POS TAG), uma etiquetação morfossintática, que consiste na associação, a cada palavra do corpus, de uma etiqueta que indica a sua categoria sintática (Voutilainen, 2003). A biblioteca Spacy (Industrial-strengh, 2022), em Python, possui ferramental para identificar POS TAGs e também lematizar os vocábulos para vários idiomas, inclusive Português.

```
nlp=pt_core_news_md.load()
nlp=spacy.load('pt_core_news_md')       # corpus marcado em português
controle=0
with open('tabela.csv','r') as csvFile:  # uma tabela em csv com os dados de interesse
  reader = csv.reader(csvFile)
  for row in reader:
        if controle != 0:
            textao =  row[1] + ", " + row[3] + ", " + row[4] + ", " + row[5]
            document = nlp(textao)          # aplica lematização, POS tagging e outros
            texto = ""
           for token in document:
                if (token.pos_ == 'NOUN') | (token.pos_ == 'VERB') |  \
                        (token.pos_ == 'PUNCT') | (token.pos_ == 'PROPN') :
                   # ficaremos apenas com as formas lematizadas
                   texto = texto + token.lemma_ + " "
                 # como queremos um texto, formado pelos lemas,
                 #manteremos a pontuação
                  if (token.lemma_ == "."):
                        texto = texto + "\n"
            linhaSaida = texto
….
```

Figura 5. Código para a seleção de atributos - metadados e vocábulos em forma lematizada

Por exemplo, no trecho de código mostrado na Figura 5, são selecionados os metadados de interesse correspondentes às colunas 1 (título da obra), 3 (palavras-chaves), 4 (palavras-chaves padronizadas pelo Thesagro[6]) e 5 (conteúdo ou resumo da obra) da planilha de entrada como componentes do texto a ser pré-processado. Na sequência, processa-se o texto (utilizando-se o conjunto de bibliotecas nlp) e para cada *token* (vocábulo, pontuação, etc) verifica-se se é nome, nome próprio, verbo ou pontuação. Em caso afirmativo, a forma lematizada do *token* permanece como elemento do texto a ser gravado. Lematização é uma forma de normalização de palavras em que uma palavra flexionada é substituída pela sua forma canônica, eliminando-se gênero, número e demais flexões, por exemplo em: andarei → andar; andamos → andar; andarás → andar; andou → andar; garota → garoto. Essa forma de normalização permite diminuir o número de vocábulos utilizados como atributos num modelo de aprendizado.

Esses textos padronizados formam a base textual a ser utilizada nas próximas etapas, pois o modo de representação dos dados depende do modelo de aprendizado a ser utilizado. Na Figura 6 pode-se observar o exemplo de um registro da planilha e o seu respectivo resultado após a aplicação desta padronização:

```
"ANDRADE, C. M. S. de; ASSIS, G. M. L. de; FAZOLIN, M.; GONCALVES,
R. C.; SALES, M. F. L.; VALENTIM, J. F.; ESTRELA, J. L. V.",Capim-
-tangola: gramínea forrageira recomendada para solos de baixa permea-
bilidade do Acre.,2009,Brachiaria arrecta.,Gramínea; Pastagem; Planta
Forrageira.,Livros

A saída é

Capim-tangola : gramínea recomendar solo permeabilidade Acre .
, brachiaria .
, Gramínea ; pastagem ; Planta Forrageira .
```

Figura 6. Exemplo de registro e seu resultado após pré-processamento.

Note, na saída ilustrada na Figura 6, que os vocábulos utilizados como adjetivos não aparecem, é o caso de: forrageira, baixa e arrecta. No entanto,

6 O Thesaurus Agrícola Nacional (Thesagro) é um vocabulário especializado em agricultura e utilizado para o controle terminológico nos processos de indexação e recuperação dos documentos.

Planta Forrageira, devido à grafia em letras maiúsculas, foi considerada um nome próprio. Esses detalhes, com alguma perda de informação, podem ser considerados importantes em outras etapas e, nesse caso, pode-se voltar a esta etapa e selecionar também adjetivos ou locuções adjetivas.

Ainda é importante salientar que tanto a planilha convertida para o formato CSV quanto os arquivos resultantes deste pré-processamento devem ser gravados em formato *8-bit Unicode Transformation Format* (UTF-8), devido aos pré-requisitos de padronização de entrada e saída das implementações do ferramental de software utilizado.

Extração de padrões

Nesta etapa são utilizados modelos de aprendizado de máquina a fim de observar os padrões obtidos, que são os resultados dos modelos, para de avaliá-los quanto a sua utilidade. Para o caso específico de obtenção de tópicos, decidiu-se pelo uso do modelo LDA, implementado no ambiente Mallet[7]. Para essa implementação do modelo LDA há uma particular representação do conjunto de treinamento e então o modelo pode ser utilizado.

Montando o conjunto de treino para o modelo LDA/Mallet

Na etapa anterior foram padronizados e selecionados os vocábulos de interesse e a pontuação foi mantida devido ao uso de um *wrapper* específico para o modelo LDA implementado no Mallet – LDA/Mallet, que separa os atributos de interesse, isto é, os vocábulos, e precisa da pontuação para separá-los. Para criar a base de treino do *wrapper* utilizado, deve-se aplicar mais algumas padronizações aos dados gerados, como mostra o código ilustrado na Figura 7.

```
# Processar os textos com a função little_mallet_wrapper.process_string().
# E, adicionar os textos ao conjunto de treinamento
...
training_data = []     # conjunto de treinamento, no caso os arquivos padronizados
for file in files:
    text = open(file, encoding='utf-8').read()  # textos gravados no passo anterior
    processed_text = little_mallet_wrapper.process_string(text, numbers='remove')
    training_data.append(processed_text)         # alimentando a base de treino
...
```

Figura 7. Código para a montagem do conjunto de treinamento para o modelo LDA/Mallet.

7 Documentação disponível em: https://mimno.github.io/Mallet/ consultado em dezembro de 2023.

O pré-processamento disponível no *wrapper* utilizado deveria ser suficiente para a maior parte dos textos, pois ele permite ler individualmente cada texto, transformar a grafia de todos os vocábulos em letras minúsculas, remover *stopwords* (artigos, preposições, etc) e remover pontuação e números. Algumas dessas padronizações independem da língua na qual o texto está escrito, pois mesmo as *stopwords*, em geral, correspondem a uma lista de vocábulos. Por isso, quando se deseja realizar uma seleção de vocábulos específicos, tais como os nomes e/ou adjetivos, precisa-se de alguma ferramenta de processamento de língua natural, como a utilizada na etapa anterior.

Usando o pré-processamento do *wrapper*, o córpus para o treino do modelo (*training_data* no código ilustrado na Figura 7) é uma lista de textos em que cada elemento contém apenas os vocábulos já padronizados. Um exemplo de alguns elementos da lista encontram-se ilustrados na Figura 8.

```
['resposta morfofisiológico brachiaria spp resposta morfofisiológico morfo-
fisiológico brachiaria brachiaria spp resposta morfofisiológico brachiaria
morfofisiológico brachiaria spp alagamento solo síndrome morte capim
marandu alagamento solo solo síndrome síndrome morte morte capim marandu
alagamento solo síndrome solo síndrome morte síndrome morte capim marandu
açucares solúvel açucares solúvel pastagem degradar pastagem degradar pa-
tógeno anoxia brachiaria brizantha brachiaria brizantha', 'principal foco
fonte queimada brasil principal foco foco fonte fonte queimada queimada
brasil principal foco fonte foco fonte queimada fonte queimada brasil
causa aceiros alternativas capoeira manejo carrapato controle cigarrinha
pastagem cigarrinha pastagem controle cobertura orgânico cobertura orgâni-
co fertilização manejo florestal manejo florestal … ']
```

Figura 8. Exemplo de alguns elementos da lista do conjunto de treino (*training_data*).

O modelo LDA implementado no ambiente Mallet interpreta essa lista de textos, verificando as frequências de cada vocábulo em todo o conjunto de treino e inferindo a distribuição de probabilidades dos vocábulos para este conjunto de treino. Não existe nenhuma consideração a priori sobre a classe gramatical do vocábulo e cada um é considerado estatisticamente independente, pois trata-se de uma *bag-of-words*.

Modelo LDA/Mallet

A alocação latente de Dirichlet (*Latent Dirichlet Allocation* (LDA)) é um modelo estatístico generativo, pois procura explicar semelhanças entre partes de dados em um conjunto de observações a partir de variáveis latentes. Os tópicos são as variáveis latentes, isto é, variáveis não diretamente observadas, mas que podem ser inferidas, com um bom grau de confiança, a partir das variáveis que foram observadas. Neste caso, as variáveis observadas são palavras em documentos, logo, cada documento é uma mistura probabilística de tópicos (variáveis latentes), inferidos a partir das palavras presentes em todo o conjunto de documentos. O conjunto de documentos precisa fazer sentido e pertencer a algum domínio (recorte) pré-estabelecido, o que é garantido pela expressão de busca utilizada na BDPA na etapa 1, relacionada à escolha das fontes de dados. O modelo em si exige um bom conhecimento da teoria de probabilidades e pode ser encontrado em Blei et al. (2003).

O seu uso, via o *wrapper* escolhido, é bem simples, como ilustrado no trecho de código da Figura 9. Os parâmetros para a configuração do *wrapper* são:

```
little_mallet_wrapper.quick_train_topic_model(
                path_to_mallet,
            output_directory_path,
                 num_topics,
                training_data)
```

Figura 9. Parametrização e execução do modelo LDA.

- path_to_mallet: diretório onde foi instalado o pacote do Mallet.
- output_directory_path: especifica um diretório para as saídas do modelo. Será gravado um arquivo com os textos usados no treino, outro para o modelo treinado, outro para a descrição dos tópicos e outro para as distribuições dos tópicos, com as respectivas distribuições dos documentos em cada tópico.

- num_topics: o número de tópicos deve ser pré-fixado. Pode-se ir variando o número a cada experimento, até chegar a tópicos mais subjetivamente compreensíveis.
- training_data: corresponde à lista de documentos com seus vocábulos, criada com o código ilustrado na Figura 5.

O critério de escolha para o número de tópicos é bastante subjetivo neste tipo de aplicação, pois a leitura humana dos resultados é o melhor direcionamento. Para a escolha, pode-se utilizar uma visualização dos resultados de cada modelo aprendido.

Escolha do número de tópicos

Espera-se que o usuário deste processo seja capaz de avaliar subjetivamente os tópicos encontrados, a fim de decidir o número de tópicos que o modelo deve buscar.

Os resultados podem ser visualizados de diversas formas. O gráfico, por se tratar da estimativa da probabilidade do tópico pertencer ao documento, pode ser um mapa de calor, uma representação de dados em que os valores (neste caso, as probabilidades) são representados em gradações de cores. Por exemplo, no código ilustrado na Figura 10, são considerados os seguintes parâmetros:

```
# com o número de tópicos:  num_topicos = 14
import random
target_labels = random.sample(nossos_titles, 21)
little_mallet_wrapper.plot_categories_by_topics_heatmap(nossos_titles,
                                    topic_distributions,
                                    topics,
                                    output_directory_path + '/categories_by_topics.pdf',
                                    target_labels=target_labels,
                                    dim= (13, 9)
                                   )
```

Figura 10. Parametrização e execução do gráfico do Mapa de Calor.

1. num_topics = 14
2. nossos_titles: lista com os títulos de cada obra utilizada, das quais escolhemos aleatoriamente 21. Essa lista contém todos os títulos das obras (textos) utilizadas no processo.

3. topic_distributions: calculado no código ilustrado na Figura 9, um dos resultados é o arquivo com a distribuição dos textos nos tópicos.
4. topics: são os tópicos enumerados, calculados com a execução código ilustrado na Figura 9.
5. categories_by_topic.pdf: arquivo com o mapa de calor gerado.
6. target_labels: os tópicos selecionados aleatoriamente.
7. 13,9: dimensão do mapa – são 14 tópicos (de 0 a 13), para os quais são mostrados os 10 (0 a 9) primeiros vocábulos.

Na Figura 11 é mostrado o mapa de calor no qual os títulos dos 21 docu-mentos aleatoriamente escolhidos aparecem à esquerda e os 14 tópicos na parte superior, onde podem ser observados os cinco primeiros vocábulos de cada. Os tópicos correspondem a conjuntos de palavras estatisticamente mais significantes para representá-los, logo, cada tópico precisa ser subjeti-vamente inferido a partir do conjunto de palavras. A probabilidade do docu-mento pertencer a cada tópico é representada pela intensidade da cor.

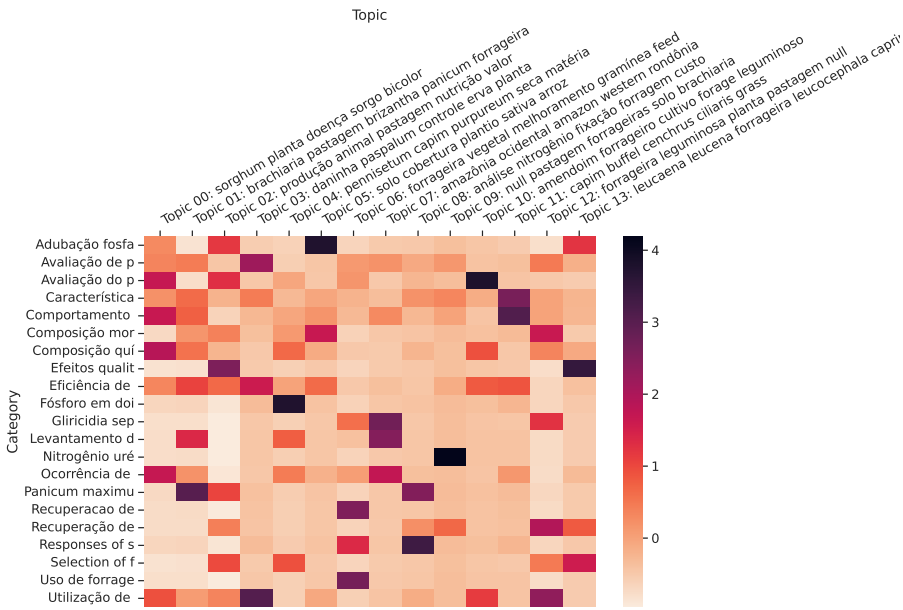


Figura 11. Mapa de calor para representar os tópicos e documentos associados.

O especialista que ajusta o modelo de aprendizado deve variar a escolha dos documentos e o número de tópicos, a fim de melhor conseguir inferir o significado de cada tópico de maneira subjetiva. Também pode ocorrer alguma dificuldade devido à perda de informação com o primeiro pré-processamento, como a ausência de “verbos” ou “adjetivos”, o que pode levar a um retorno à etapa 2 (pré-processamento).

Ao final de algumas repetições dessas visualizações, pode-se decidir tanto o número de tópicos, como a qualidade de pertencimento dos documentos a eles. Assim, pode-se estabelecer a probabilidade mínima de um documento pertencer a um desses tópicos, desde que faça sentido para o especialista.

Avaliação

O mapa gerado na etapa anterior é estático e não é possível conseguir uma boa visualização de todos os documentos - por isso, essa escolha deve ser aleatória. A representação é muito boa, pois permite ver quais documentos são mais próximos na divisão em tópicos, pela gradação das cores. Porém, para apresentar a análise desses dados ao especialista de domínio, que não ajusta os modelos de aprendizado, é necessário escolher uma forma mais simples de se filtrar a apresentação gráfica tópicos ou documentos. É possível, por exemplo, separar os documentos por tópicos pelos intervalos de distribuição como [0,0.25], [0.26,0.50], [0.51,0.75] e [0.76,1.0], ou escolher uma probabilidade mínima dos documentos pertencerem a algum tópico, ou ainda a visualização em uma página web.

O Google Looker Studio oferece a opção de mapas de árvores para mostrar os dados organizados em hierarquias de dimensão. O problema é a impossibilidade de usar a gradação da probabilidade de pertencimento do documento ao tópico; logo, é necessário escolher um intervalo de probabilidades de pertencimento de um documento a um tópico, que seja subjetivamente julgado suficiente. Então, considerando-se os tópicos como cada dimensão dos dados, uma possibilidade é mostrá-los como ilustrado na Figura 12. Ainda, essa visualização permite a utilização de filtros. Logo, se filtrarmos os resultados, na Figura 12, pelo tópico “degradabilidade capim…”, pode-se navegar apenas pelos documentos desse tópico específico, como ilustrado na Figura 13.

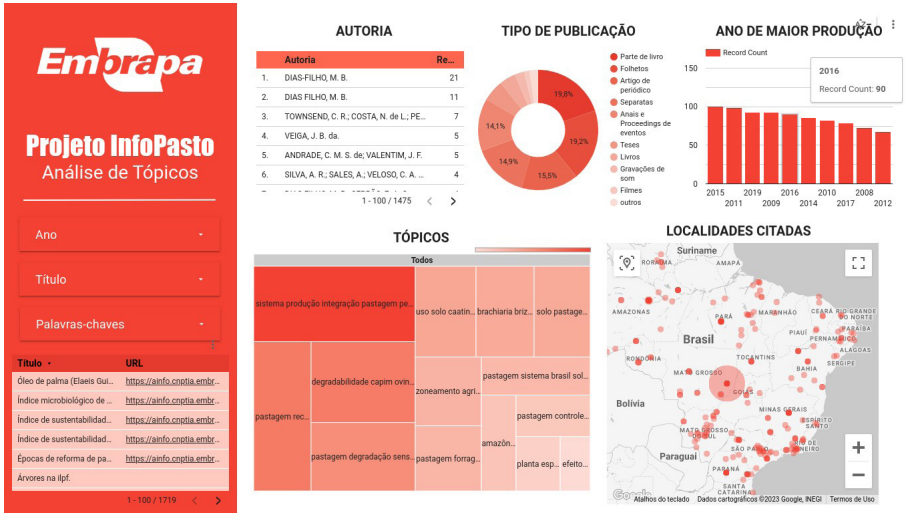


Figura 12. Representação dos documentos em tópicos utilizando árvore no Google Looker Studio.

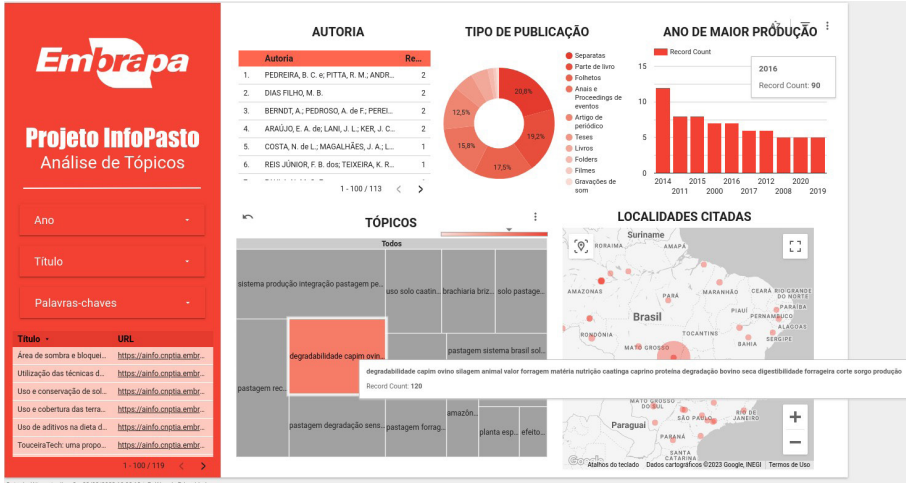


Figura 13. Filtrando os resultados por um tópico específico.

Neste tipo de visualização, o usuário navega com mais facilidade e pode melhor avaliar esses resultados, chegando à conclusão de que: 1) os tópicos obtidos não fazem sentido, o que pode resultar em um retorno à etapa 3

(extração de padrões); 2) os vocábulos que aparecem nos tópicos poderiam ser ampliados (retorno à etapa 2, pré-processamento); ou 3) os tópicos e documentos não o ajudam a tomar alguma decisão, o que pode provocar um retorno à etapa 1 (escolha das fontes de dados).

Experimentos e resultados

Nesta seção são abordados os experimentos realizados, a avaliação desses experimentos e a comparação dos seus resultados.

Experimento base

A primeira separação de publicações da Embrapa na área de pastagens foi realizada de modo a cobrir o tema de forma ampla, pois o objetivo da construção dessa base de dados é mapear esse conhecimento na Embrapa. A expressão de busca utilizada na BDPA foi obtida após discussões exaustivas da equipe, sendo a sua versão final a apresentada na Figura 14.

```
(((past* OR graz* OR silvipast* OR silvopast* OR grass* OR rangeland
OR gram* OR forra* OR capim) AND (abandon* OR compact* OR conserva* OR
manejo OR degrada* OR desertifica* OR ecologic* OR restorat* OR restaura*
OR erosa* OR erosi* OR invas* OR nativ* OR plant* OR specie* OR espécie
OR degrada* OR leaching OR lixivia* OR recov* OR recuper* OR reforest* OR
reflorest* OR regenera* OR runoff OR escoa* OR biomass* OR loss* OR perda*
OR ecoss* OR servi* OR noxious OR nociva* OR weed* OR erva OR "water re-
pellency" OR "repelência à água" OR daninha*) or (Panicum* OR Brachiaria*
OR Andropogon* OR Pennisetum* OR Cynodon* OR Cenchrus* OR Digitaria* OR
"Setaria italica*" OR "setaria sphacelata*" OR Hemarthria* OR Chloris* OR
Paspalum* OR Stylosanthes* OR Arachis* OR Cajanus* OR Leucaena* OR calo-
pogonium* OR lespedeza* OR neonotonia* OR macroptilium* OR macrotyloma* OR
desmanthus* OR Medicago* OR trifolium* OR ryegrass* OR lolium** OR Lotus*
OR Avena* OR Sorghum Sudanense*  OR Centrosema* OR Desmodium* OR Pueraria*
OR Echinochloa* OR Melinis*)) AND (ano-publicacao:[* TO 2010]))
```

Figura 14. Expressão de busca utilizada no experimento base.

Para essa expressão foram recuperadas 10.848 obras, em setembro de 2021 em toda a BDPA, de maneira que inclui a produção científica da Embrapa, além de livros, artigos e outros documentos catalogados nas bibliotecas da Embrapa, de acordo com

as palavras-chaves e suas combinações explicitadas na expressão de busca até o ano de 2010 – segundo escolhas da equipe à época. Foi produzido um relatório com o uso das palavras-chaves catalogadas junto às obras como temas. Na Figura 15 é ilustrado o resultado após o uso do filtro de palavras-chaves com "ILPF" e "integração lavoura pecuária e floresta". Esse relatório, embora bibliometricamente bastante completo, restringia o entendimento dos temas aos quais as obras se referem, dificultando a navegação em busca, por exemplo, de áreas para as quais as soluções mostradas nas obras pudessem ser aplicadas. Devido à dificuldade em representar temas comuns e agrupar conhecimentos específicos, foi configurado o Experimento 1.

Figura 15. Obras resultantes da busca abrangente sobre o tema Pastagens na produção científica da Embrapa, filtradas pelo tema ILPF.

Experimento 1

A relação de obras obtidas no experimento base consistia apenas nos metadados de interesse bibliométricos das obras, a saber: "Autoria", "Título", "Edição", "Fonte/Imprenta", "Ano de publicação", "Notas", "Série", "Palavras-Chave", "Thesagro", "Thesaurus Nal", "Tipo da Publicação", "Origem Material", "Referência bibliográfica", "URL" e "Biblioteca(s)". Para ter mais informações sobre os conteúdos das obras foi realizada uma nova busca diretamente sobre a base de dados, o que não é disponibilizado pela ferramenta de acesso disponível ao público.

A nova busca considerou o mesmo conjunto de vocábulos e expressão de busca do experimento base, porém, restringiu-se à produção científica da Embrapa, isto é, a obras produzidas por pesquisadores da Embrapa, e incluiu nos resultados obtidos

o Resumo (ou Conteúdo) de cada obra. Ainda, em termos de tipos de publicações foram considerados: partes de livros, artigos, separatas, anais de congressos e eventos, folhetos e teses. As seguintes informações foram providas para cada obra: Autoria, Afiliação(ões), Título, Resumo, Edição, Fonte/Imprenta, Ano de Publicação, Notas, Série, Palavras-Chaves, Thesagro, Thesaurus NAL, Tipo da Publicação, Origem Material, Referência Bibliográfica, Bibliotecas, URI. Essa busca resultou em 1.867 obras, em meados de 2021, que são as utilizadas neste experimento.

Foram realizados diferentes pré-processamentos, resultando em dois grupos de coleções textuais:

1. **Coleção 1**: para este grupo foram utilizados apenas os metadados de palavras-chaves (termos livres) e o de Thesagro (palavras-chaves no padrão Thesagro). Pré-processando os textos com as bibliotecas de processamento de língua natural (PLN), foram considerados apenas os substantivos para os dados padronizados. Na etapa de extrações de padrões foram especificados 7, 14 e 21 tópicos.
2. **Coleção 2**: para este grupo de resultados foram utilizados os metadados título, palavras-chaves (termos livres), thesagro (palavras-chaves padronizadas pelo Thesagro) e conteúdo (resumo da obra). No pré processamento com as ferramentas de PLN foram pegos apenas os verbos e os substantivos, formando os dados padronizados. Na etapa de extração de padrões foram gerados 7, 14 e 21 tópicos.

A avaliação de resultados pelo especialista em mineração de textos foi ocorrendo em paralelo às escolhas dos parâmetros de cada coleção textual como colocado e foram gerados os gráficos de calor com amostras aleatórias dos textos de cada coleção. Ao especialista do domínio foram enviados: 1) os mapas de calor, com as devidas explicações dos dois conjuntos de dados; 2) relação de todas as palavras de cada tópico, dado que no mapa de calor aparecem apenas as 6 primeiras palavras descritivas de cada tópico; e 3) relação de obras e tópicos associados. As avaliações foram realizadas lado a lado para cada conjunto de tópicos gerado a partir das duas coleções de dados padronizados. Na Figura 16 são mostrados exemplos desses mapas de calor. Porém, a melhor comparação foi via valores tabelados, considerando-se a descrição completa (todas as palavras) de cada tópico e a relação de publicações associadas aos tópicos.

O melhor resultado para todo o conjunto de tópicos das duas coleções, subjetivamente interpretado pelo especialista, foi o de 14 tópicos para a Coleção 2. A partir desses tópicos, foi realizado um trabalho de inferência

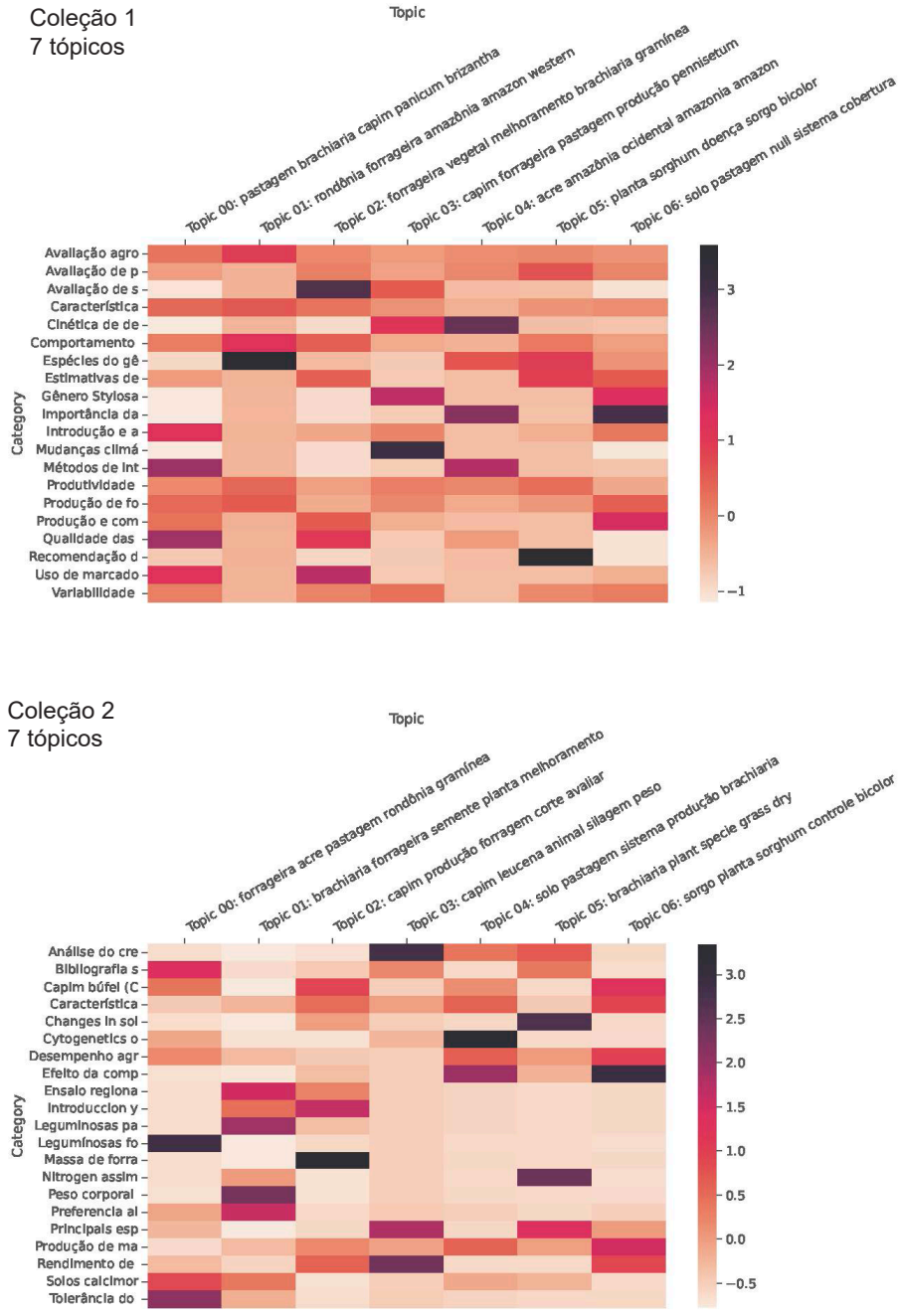


Figura 16. Tópicos para as Coleções 1 e 2, para comparar.

subjetiva sobre os possíveis relacionamentos entre o conjunto de palavras representativo de cada tópico (primeira coluna da Tabela 1) e alguns documentos a ele associados, resultando em um possível tema (segunda coluna da Tabela 1). Esses tópicos são bastante genéricos e podem servir a um processo de classificação automática, se os considerarmos como uma possível rotulação das obras da Coleção 2; assim, após essa rotulação, outras obras poderiam ser recuperadas ou classificadas de acordo com os conjuntos de palavras que identificam cada tópico. Note, na Tabela 1, que alguns dos tópicos identificados auxiliam reconhecer a quais regiões (ou biomas) do Brasil a obra se refere, por exemplo, Amazônia ou Semiárido.

Tabela 1. Coleção 2 – 14 Tópicos e inferências subjetivas sobre eles.

Tópicos: conjunto de palavras	Inferência subjetiva
0: adubação nitrogênio dose solo produção fósforo efeito teor nutriente aveia guandu cultivar planta aplicação fertilizante alfafa cajanus cajan forrageira avaliar	Fertilidade de solos, leguminosas
1: rondônia forrageira ciat gramínea andropogon pastagem rendimento leguminosa velho porto amazônia campo western amazon ocidental forragem humidicola capim embrapa paspalum	Amazônia
2: sorgo sorghum planta bicolor híbrido genótipo cmsxs antracnose colletotrichum genotipo cultivar avaliar resistência graminicola plantio doença moench apresentar graminum pulgão	Melhoramento genético e avaliação de cultivares
3: peso animal ovino pastagem dieta caprino consumo suplementação alimentar feno bovino dia efeito tratamento receber vaca nível leite suplemento raça	Produção animal
4: brachiaria specie hybrids vegetal plant brizantha forrageira meiotic accessiom genetic humidicola melhoramento breeding genético cultivars chromosome pastagem markers selection feed	Biotecnologia
5: leucena semi capim buffel leucaena forrageira nordeste leucocephala cenchrus especie planta árido producao ciliaris regiao apresentar area uso periodo região	Semiárido

6: pastagem acre produção capim forrageira gramínea amazônia amendoim sistema estado leguminosa produtividade região brachiaria condição brasil manejo embrapa arachis solo	Amazônia oriental
7: solo pastagem sistema brachiaria área brizantha espécie trabalho produção ano manejo marandu avaliar objetivo região degradar plantio cerrado cobertura densidade	Pasto degradado; Cerrado
8: dry matter grass plant levels grazing leaf pasture legume study rate yield growth evaluate area used pastorar soil abstract four	Produção (inglês)
9: melhoramento forrageira planta brachiaria acesso vegetal genético seleção gramínea cultivar grande marcador panicum genótipo brasil maximum sul utilizar feed programa	Melhoramento genético e biotecnologia
10: capim silagem elefante pennisetum teor purpureum valor proteína digestibilidade matéria vitro fibra detergente corte fdn apresentar forragem fda gramínea dia	Valor nutritivo
11: capim produção forragem pastejo panicum maximum avaliar corte matéria pastagem dia folha taxa brachiaria tanzânia período altura tratamento cultivar brizantha	Produção (português)
12: semente brachiaria planta brizantha espécie forrageira germinação fungo gramínea marandu doença efeito pastagem controle campo stylosanthes gênero qualidade acesso bra	Tecnologia de sementes; plantio
13: milho cultura planta soja sistema cobertura plantio herbicida controle capim integração cultivo milheto sorgo arroz tratamento consórcio girassol brachiaria solo	ILPF; forrageiras anuais

Porém, ainda na Tabela 1, observa-se que os tópicos não auxiliam a identificação de métodos ou tecnologias diretamente associados à degradação ou reforma de pastagens. Por exemplo, embora o Tópico 7 pareça conter mais textos sobre Pasto Degradado, pode não ser o único. Um tópico como o de Nutrição de Animal ou Tecnologia de Sementes pode conter alguma técnica para reforma de pastagens, etc. Deve-se notar que: 1) a expressão de busca utilizada para recuperar essa coleção de documentos foi bastante ampla no tema pastagem, idêntica à do Experimento Base; 2) o modelo para

encontrar os tópicos é probabilístico, isto é, há diferentes probabilidades de um texto pertencer a um tópico; e 3) a divisão permite identificar assuntos mais genéricos sobre o tema pastagem cobertos pela coleção, que era o objetivo dessa primeira busca e, quiçá, separa de forma mais clara o tema degradação. Assim, um possível caminho, a fim de fechar o tema dos textos em recuperação e reforma de pastagens degradadas, seria melhor delimitar a expressão de busca e então analisar os tópicos obtidos.

Experimento 2

Como a amplitude da expressão de busca do Experimento 1, com os filtros relativos à produção científica da Embrapa e língua portuguesa, não permitiram identificar métodos ou técnicas de reforma ou recuperação de pastagens degradadas, decidiu-se utilizar uma expressão específica, que se encontra ilustrada na Figura 17.

```
((past* OR silvipast* OR silvopast* OR gram* OR forra* OR capim)
            AND (degrad* OR recuper* OR reform*))
```

Figura 17. Expressão de busca para delimitar o tema de reforma ou recuperação de pastagens degradas.

Com essa expressão de busca delimitam-se os resultados às pastagens degradadas ou em degradação, recuperadas ou em recuperação e reformadas ou em reforma. Assim, essa expressão foi utilizada na Etapa 1 (Escolha das Fontes de Dados) sobre a BDPA, com os seguintes filtros: produção científica da Embrapa, obras em português e tipos documentos, vídeos ou áudios. Foram utilizados apenas os metadados normalmente exportados como resultados de busca na BDPA. Foram apresentados 1725 resultados de busca (em 04/11/2022).

Para o pré-processamento foram selecionados: 1) os metadados de Título, Palavras-chaves e Thesagro; e 2) apenas substantivos, nomes próprios e verbos, lematizados e as pontuações, conforme o código ilustrado na Figura 18.

```
textao =  row[1] + ", " + row[3] + ", " + row[4] # Título, Palavras-chaves e Thesagro
document = nlp(textao)
texto = ""
for token in document:
    if (token.pos_ == 'NOUN') | (token.pos_ == 'VERB') |  \
            (token.pos_ == 'PUNCT') | (token.pos_ == 'PROPN') :
        texto = texto + token.lemma_ + " "
        if (token.lemma_ == "."):
            texto = texto + "\n"
linhaSaida = texto
```

Figura 18. Seleção de atributos do Experimento 2.

Construiu-se então o conjunto de treino para o LDA/Mallet, utilizando-se o código ilustrado na Figura 7.

Foram gerados 7, 14 e 21 tópicos, utilizando-se o código ilustrado na Figura 9 em cada caso. Para cada um dos conjuntos de tópicos procedeu-se às comparações e às análises subjetivas. O conjunto de 14 tópicos foi avaliado como o mais interessante pelo especialista em pastagens, pois, subjetivamente, parecia melhor sintetizar os dados recuperados. Para testar o potencial das palavras que descrevem o tópico, procedeu-se também uma busca no Google com expressões do tipo: "Embrapa AND (palavras descritoras do tópico)". Observe os primeiros resultados, mais aderentes à expressão de busca com o Google, na segunda coluna da Tabela 2. A partir desses resultados, procedeu-se à busca de um termo que resumisse o tópico a partir dessas palavras-chaves. Houve uma primeira sugestão de um especialista em mineração de textos (veja coluna "Esp MT" na Tabela 2) sobre o tema do tópico e, a seguir, a sugestão de um especialista em pastagens (veja coluna "Esp Pastagens" na Tabela 2) sem que este tenha conhecido as sugestões do Esp MT. Dado que os termos ficaram muito próximos e correspondentes aos grupos de textos, segundo o Esp MT, eles passaram a ser o identificador de cada qual dos tópicos encontrados.

Tabela 2. Análise dos 14 tópicos gerados para a terceira coleção de documentos (Coleção 3).

Tópico	Resultado Google	Esp MT	Esp Pastagens
0: brachiaria brizantha capim pastagem marandu nitrogênio recuperação gramínea milho adubação solo consórcio maximum panicum zea produção mays degradar fertilizante forrageira	Portal Embrapa Brachiaria brizantha cv. Marandu - Fonte: Portal Embrapa	Braquiária	Braquiária
1: sistema produção integração pastagem pecuária lavoura recuperação ilpf corte cultura agricultura floresta leite gado sistemas sustentabilidade bovino tecnologia integrar produtor	ILPF é a sigla de integração--lavoura-pecuária-floresta. ... Da mesma forma, o componente pecuário pode ser feito com bovinos de corte ou leite, bubalinos, … Fonte: Portal Embrapa		
2: efeito emissão carbono metano gases impacto marcador abc plano agricultura avaliação estufa gás mitigação protocolo modelo produção sequestro óxido método	O Plano ABC, oficialmente denominado “Plano Setorial de Mitigação e de ... do Estado Brasileiro na mitigação da emissão de Gases de Efeito Estufa (GEE) e no … Fonte: Portal Embrapa		
3: planta espécie cerrado área mineração semente forrageira recuperação revegetação fungo estabelecimento população efeito florestal fruta restauração doença reserva seleção raiz	Guia de restauração do Cerrado - Fonte: Agropedia brasilis - Beneficiamento de Sementes, Viveiros e Produção de Mudas Florestais Nativas e Capacitação Continuada em. Recuperação de Áreas Degradadas), Demarcação de Áreas		

4: pastagem degradação sensoriamento remoto imagem bacia nível avaliação índice análise classificação cobertura projeto meio satélite informação geotecnologia dado área erosão	Aplicação de sensoriamento remoto no estudo dos ... - Embrapa Resumo: A degradação de pastagens é um problema global, ... meio de técnicas advindas de Sensoriamento Remoto, com o uso de imagens de satélite Sentinel-2, … Fonte: Portal Embrapa		
5: uso solo caatinga terra rio região cobertura mata brasil agricultura área água ods desenvolvimento selo sistema gerais minas vegetação nordeste	Agricultura na Caatinga - Fonte: Portal Embrapa A agricultura da Caatinga ocorre em diferentes realidades: em perímetros irrigados ou em sistemas de sequeiro; em propriedades rurais de grande, médio e, …		
6: pastagem recuperação degradar amazônia área sistema brasil espécie florestal degradação pará código ambiente árvore floresta desenvolvimento biomassa plantio mogno amazonas	Recuperação de áreas degradadas ou alteradas na Amazônia Fonte: Infoteca Embrapa Estratégias de Recuperação de Pastagens Degradadas na Amazônia		
7: pastagem controle bovino planta manejo praga corte ovino peso pastejo campo brasil daninha capim erva animal gado ganho lotação doença	Fonte: Embrapa Gado de Corte ... Controle de plantas invasoras em pastagens cultivadas nos ... Erva daninha	plantas invasoras e pragas	Estresse Biótico
8: pastagem forrageira leguminosa adubação planta brasil capim recuperação pasture gramínea brachiaria melhoramento feed produção recuperacao degradar calagem fertilizante legume panicum	Estabelecimento da pastagem – Fonte: Embrapa Efeito da época de semeadura no número de plantas/m² de Brachiaria decumbens cv. ... leguminosas, ou seja, a quantidade de sementes das gramíneas foi …	Estabelecimento da Pastagem	Estabelecimento da Pastagem

9: amazônia acre ocidental pastagem amazon western forrageira amazonia brachiaria manejo solo occidental gramínea rondônia pasto amendoim del capim suelo forrageiro	Manejo dos solos e a sustentabilidade da produção agrícola Fonte: Embrapa Introdução ... É notório que a evolução das áreas de pastagens e agricultura na Amazônia. Ocidental …	Amazônia Ocidental	Amazônia Ocidental
10: degradabilidade capim ovino silagem animal valor forragem matéria nutrição caatinga caprino proteína degradação bovino seca digestibilidade forrageira corte sorgo produção	Fonte: Embrapa Valor nutritivo. Avaliou-se o consumo e valor nutritivo de silagens de maniçoba contendo níveis de O, 8, 16 e 24% de resíduo vitivinícola. Foram utilizados quatro ovinos machos, ...qualidade e valor nutritivo de forragem	Valor Nutritivo	Valor Nutritivo
11: zoneamento agricultura pesquisa agrícola embrapa brasil uso terra planejamento recurso desenvolvimento sustentabilidade município natural meio tecnologia produção agroecológico açúcar ambiente	Fonte: Portal Embrapa Na safra 1996/1997 o Ministério da Agricultura, Pecuária e Abastecimento (Mapa) passou a operacionalizar o zoneamento agrícola do Brasil e publicar anualmente o …	Zoneamento Agrícola	Zoneamento Agrícola
12: pastagem sistema brasil solo pasture recuperação arroz tropical cultura cerrado cultivo análise barreirão floresta system úmida agrofloresta reclamation amazonas deterioração	Fonte: Embrapa Renovação de pastagens de cerrado com arroz O Sistema Barreirão revela que, com a associação arroz x pasto na recuperação de pastagens degradadas, é possível melhorar-se substancialmente a produção ...	Barreirão Arroz Cerrado	Barreirão Cerrado
	Não encontrados: tropical floresta úmida agrofloresta reclamation amazonas deterioração		
13: solo pastagem manejo fertilidade degradação sistema degradar latossolo cerrado conservação recuperação carbono matéria deterioração uso soil atributo área cultivo água	Fertilidade do solo em pastagem da fertilidade do solo nesse ecossistema é essencial para a implementação de boas práticas de manejo que promovam o seu uso de forma eficiente. Fonte: Infoteca/Embrapa	Fertilidade dos Solos	Fertilidade dos Solos

Discussão dos resultados

Inicialmente dispunha-se dos resultados apresentados para uma busca sobre a BDPA, no tema mais amplo de pastagens, considerando todo o material disponível nessa base de dados, relatado no Experimento Base. Porém, não havia sido realizada uma boa síntese desse material, o que dificultava avaliar quais temas e subtemas poderiam resumir o conhecimento produzido pela Embrapa na área de pastagens.

Assim, utilizando-se a mesma expressão de busca, procurou-se filtrar apenas as publicações produzidas pela Embrapa, a fim de procurar sintetizá-las de acordo com seus temas, subtemas, regiões geográficas e época de disponibilização. Para isso, foi conduzido o Experimento 1, com diferentes escolhas de atributos na etapa de pré-processamento, de modo que se pudessem avaliar subjetivamente o que mais valeria a pena utilizar e qual o número de tópicos. Após essas avaliações subjetivas, conseguiu-se chegar a 14 tópicos: “FERTILIDADE DE SOLOS, LEGUMINOSAS”, “AMAZÔNIA”, “MELHORAMENTO GENÉTICO E AVALIAÇÃO DE CULTIVARES”, “PRODUÇÃO ANIMAL”, “BIOTECNOLOGIA”, “SEMIÁRIDO”, “AMAZÔNIA ORIENTAL”, “PASTO DEGRADADO”; “CERRADO”, “PRODUÇÃO (INGLÊS)”, “MELHORAMENTO GENÉTICO E BIOTECNOLOGIA”, “VALOR NUTRITIVO”, “PRODUÇÃO (PORTUGUÊS)”, “TECNOLOGIA DE SEMENTES e PLANTIO” e “ILPF e FORRAGEIRAS ANUAIS”. Embora ainda muito amplos, os tópicos permitiram uma boa sintetização das informações permitindo identificar, inclusive, biomas para os quais há mais pesquisa e desenvolvimento na Embrapa no tema pastagens.

Voltando a atenção para o tema de reforma e recuperação de pastagens degradadas, embora o tema Pasto Degradado esteja entre os tópicos obtidos no Experimento 1, como se trata de um modelo probabilístico de distribuição dos textos em tópicos, muitos dos textos dos demais tópicos poderiam conter informações de interesse sobre reforma e recuperação de pastagens degradadas. Dessa forma, para melhor delimitar esse tema foi realizada uma nova busca e mais uma vez aplicado o processo de mineração de textos, como descrito no Experimento 3.

Como as obras recuperadas, no Experimento 3, já se encontram com o tema fechado em reforma e recuperação de pastagens degradadas, os tópicos encontrados foram úteis para separar métodos e tecnologias sobre

pastagens degradadas ou em reforma ou em recuperação de acordo com biomas (Amazônia, Cerrado), modos de produção (ILPF, zoneamentos, estabelecimento de pastagens), características de solo (fertilidade) ou qualidade pastagem (valor nutritivo) e também estresse biótico (plantas invasoras, etc).

Para a apresentação do resultado final da síntese de metodologias foi disponibilizado um relatório interativo, ilustrado na Figura 19, que contém todas as informações de interesse da análise realizada: localidades citadas nas obras, autoria, link para as obras, autorias, tipos de publicação e distribuição temporal das obras. Esse tipo de relatório facilita a visualização dos dados, por meio de filtros sobre cada uma das categorias das informações apresentadas na síntese.

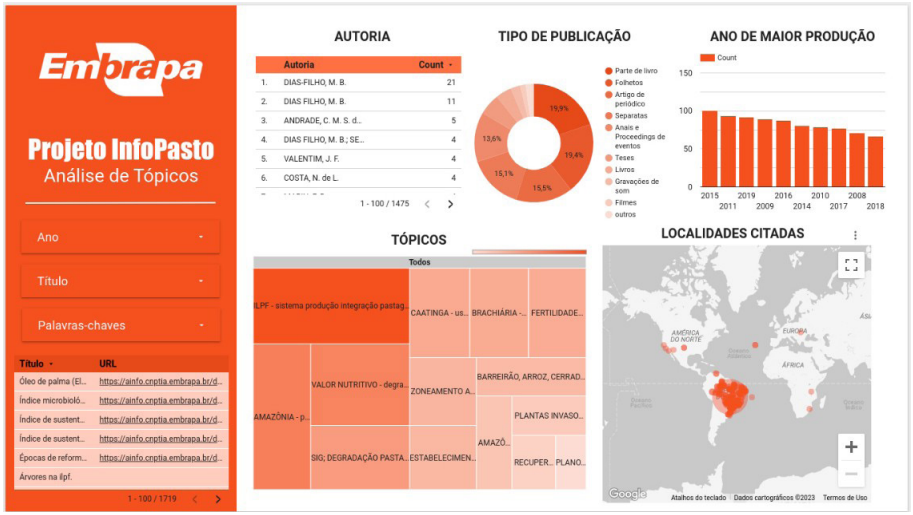


Figura 19. Relatório de tópicos sobre a busca específica realizada neste trabalho.

Considerações finais

O objetivo deste trabalho foi produzir uma síntese do mapeamento de metodologias, tecnologias disponíveis, produtos, serviços e possíveis recomendações da Embrapa no tema de recuperação de pastagens degradadas, de um modo semi-automático. A síntese deveria permitir identificar tendências

em relação a metodologias e tecnologias desenvolvidas pela Embrapa, considerando aspectos como regiões geográficas de aplicação, temas e subtemas de cobertura e época de desenvolvimento das tecnologias e metodologias.

Para obter a síntese foram utilizados resultados recuperados da BDPA submetidos a um processo de mineração de textos, desenvolvido por especialistas em mineração de dados e no tema pastagens. Os experimentos realizados, com textos recuperados no tema pastagem mais genérico e depois melhor delimitando o domínio de reforma e recuperação de pastagens degradadas, mostraram que o processo aplicado é eficaz para mapear de maneira sintetizada os textos recuperados. No tema específico de reforma e recuperação de pastagens degradadas foram sintetizadas recomendações e tecnologias aplicadas a diferentes biomas (Amazônia, Cerrado, Caatinga), em combinação com diferentes sistemas de produção e suas adequações a diferentes contextos (local, solo, clima, sistema de produção, tipo de capim, etc), facilitando a adoção das tecnologias pelo setor produtivo.

Desta forma, acredita-se que o processo descrito neste trabalho possa ser aplicado a outros temas, com as devidas parametrizações e a participação de especialista em mineração de dados e no domínio de conhecimento de interesse.

Referências

BLEI, D. M. Probabilistic topic models. **Communications of the ACM**, v. 55, n. 4, p. 77-84, 2012. DOI: 10.1145/2133806.2133826.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet allocation **Journal of Machine Learning Research**, v. 3, n. 4-5, p. 993-1022, 2003. DOI: 10.1162/jmlr.2003.3.4-5.993.

CHURCHILL, R.; SINGH, L. The evolution of topic modeling. **ACM Computing Surveys**, v. 54, n. 10s, 215, Jan. 2022. DOI: 10.1145/3507900.

INDUSTRIAL-STRENGTH natural language processing in Python. Disponível em: https://spacy.io/. Acesso em: 2 dez 2022.

MOURA, M. F.; MEDEIROS, C. M. **Uma estratégia para a identificação de citações geográficas em textos técnico-científicos da área agrícola na língua portuguesa.** Campinas: Embrapa Agricultura Digital, 2022. 19 p. il. color. (Embrapa Agricultura Digital. Boletim de pesquisa e desenvolvimento, 52). Disponível em: https://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1150456. Acesso em: 20 jun. 2022.

MOURA, M. F.; NOGUEIRA, B. M.; CONRADO, M. da S.; SANTOS, F. F. dos; REZENDE, S. O. Making good choices of non-redundant N-gram words. In: INTERNATIONAL CONFERENCE ON COMPUTER AND INFORMATION TECHNOLOGY, 11., 2008, Khulna **Proceedings** [...]**.** [Piscataway]: IEEE, 2008. p. 64-71. DOI: 10.1109/ICCITECHN.2008.4803111.

MOURA, M. F.; TAKEMURA, C. M.; SILVA, I. L. C.; TÁPIAS, L. M.; OLIVEIRA, C. T. de; BASSOI, L. H.; OLIVEIRA, S. R. de M. Metodologia para a construção de portfólios tecnológicos agrícolas a partir de publicações técnico-científicas. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 11., 2017, Campinas. **Ciência de dados na era da agricultura digital:** anais. Campinas: Editora da Unicamp: Embrapa Informática Agropecuária, 2017. p. 537-546. SBIAgro 2017. Disponível em: https://www.alice.cnptia.embrapa.br/alice/handle/doc/1085539. Acesso em: 13 jun. 2023.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. **Revista de Sistemas de Informacao da FSMA**, n. 7, p. 7-21, 2011.

VACARI, I.; LENK, L. M.; GONZALES, L. E.; VISOLI, M. C. Recuperação de informação: Bases de Dados da Pesquisa Agropecuária. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 6., 2007, São Pedro, SP. **Anais** [...]. Campinas: Embrapa Informática Agropecuária, 2007. p. 221-225. SBIAgro 2007. Disponível em: https://www.alice.cnptia.embrapa.br/alice/handle/doc/9582. Acesso em: 14 jun. 2023.

VASILIEV, Y. **Natural language processing with Python using spaCy:** a practical introduction. San Francisco: No Starch Press, 2020.

VOUTILAINEN, A. Part-of-speech tagging. In: MITKOV, R. (ed.). **The Oxford handbook of computational linguistics.** Oxford: University Press, 2003. chap. 11, p. 219-232. DOI: 10.1093/oxfordhb/9780199276349.013.0011.

Embrapa

Agricultura Digital

MINISTÉRIO DA AGRICULTURA E PECUÁRIA

GOVERNO FEDERAL

BRASIL

UNIÃO E RECONSTRUÇÃO

CGPE 018411