USO DA MINERAÇÃO DE TEXTO EM FONTES INFORMACIONAIS PARA GRUPOS DE PRODUTORES RURAIS

Jacquelin Teresa Camperos-Reyes✉
Universidade Federal do Pará
Ricardo Cesar Gonçalves Sant'Ana✉✉
Universidade Estadual Paulista

Resumo: Observando a necessidade de ferramentas tecnológicas para analisar elementos de campos informacionais e reconhecer ligações entre fontes de dados acessadas por determinados sujeitos, esta pesquisa examina dados publicados para grupos de produtores rurais. Foca em comunicações para esses grupos e datasets do governo brasileiro. O objetivo é verificar a aderência lexical entre unidades extraídas de notícias para produtores no Brasil e datasets governamentais, analisando necessidades informacionais. Os métodos incluem revisão bibliográfica, técnica 5W1H, mineração de textos e cálculo de similaridade em R. Os resultados indicam comunicação insuficiente, com proximidade apenas na categoria de Crédito. A similaridade entre as duas fontes é favorável em um nível inicial, mas há necessidade de melhor contextualização de algumas palavras. Estudos futuros visam contrastar este procedimento com outras medidas de similaridade e aplicá-lo em diferentes fontes e contextos socioeconômicos.
Palavras-chave: Aderência lexical; mineração de texto; acesso a dados; dados de governo; economia solidária.

Title: USE OF TEXT MINING IN INFORMATIONAL SOURCES FOR GROUPS OF RURAL PRODUCERS.
Abstract: Observing the need for technological tools to analyze elements of informational fields and recognize links between data sources accessed by certain subjects, this research examines data published for groups of rural producers. It focuses on communications for these groups and government datasets from Brazil. The objective is to verify lexical adherence between units extracted from news for producers in Brazil and government datasets, analyzing informational needs. The methods include a literature review, the 5W1H technique, text mining techniques, and similarity calculation in R. The results indicate insufficient communication, with proximity only in the Credit category. The similarity between the two sources is favorable at an initial level, but there is a need for better contextualization of some words. Future studies aim to contrast this procedure with other similarity measures and apply it to different sources and socioeconomic contexts.
Keywords: Lexical adherence; Text mining; Data access; Government data; Solidarity economy.

Título: USO DE LA MINERÍA DE TEXTOS EN FUENTES DE INFORMACIÓN PARA GRUPOS DE PRODUCTORES RURALES.
Resumen: Observando la necesidad de herramientas tecnológicas para analizar elementos de campos informacionales y reconocer enlaces entre fuentes de datos a las que acceden ciertos sujetos, esta investigación examina datos publicados para grupos de productores rurales. Se enfoca en comunicaciones para estos grupos y conjuntos de datos del gobierno de Brasil. El objetivo es verificar la adherencia léxica entre unidades extraídas de noticias para productores en Brasil y conjuntos de datos gubernamentales, analizando necesidades informacionales. Los métodos incluyen una revisión bibliográfica, la técnica 5W1H, técnicas de minería de textos y el cálculo de similitud en R. Los resultados indican una comunicación insuficiente, con proximidad solo en la categoría de Crédito. La similitud entre las dos fuentes es favorable a un nivel inicial, pero se necesita una mejor contextualización de algunas palabras. Estudios futuros tienen como objetivo contrastar este procedimiento con otras medidas de similitud y aplicarlo a diferentes fuentes y contextos socioeconómicos.
Palabras clave: Adherencia léxica; Minería de textos; Acceso a datos; Datos de la administración; Economía solidaria.

1. INTRODUÇÃO

O campo informacional se constitui em uma alternativa para perceber caraterísticas de relações entre distintos ciclos de vida de dados que interagem, e é particularmente útil no estudo do acesso a dados, pois possibilita a construção de pontos de vista e dimensões de posicionamento para os pesquisadores (Sant'Ana, 2019). Essa alternativa surge diante das permanentes transformações nos fluxos de dados e informações, e da riqueza dos atributos

✉✉✉ jtcamperos@hotmail.com
✉✉✉✉✉✉ ricardo.santana@unesp.br

Recibido: 11-04-2024; 2ª versión: 25-05-2024; 3ª versión: 29-05-2024; aceptado: 09-07-2024.

CAMPEROS-REYES, J.T. y GONÇALVES SANT'ANA, R.C. Uso de la minería de textos en fuentes de información para grupos de productores rurales. *Anales de Documentación*, 2024, vol. 27. Disponible en: http://dx.doi.org/10.6018/analesdoc.611661.

que descrevem o acesso a dados, mantendo ativa a demanda por pesquisas sobre detalhes de um cenário, onde, como expressado por Barreto (2002), copiosas transformações ainda nem estão sendo percebidas.

Neste enquadramento, entende-se a necessidade do uso de ferramentas para analisar elementos de campos informacionais, coletando indícios sobre a ligação entre uma e outra fonte de dados às que tem acesso determinado sujeito informacional.

Esta pesquisa se situa em dados publicados para grupos de produtores rurais, indagando em duas fontes, em comunicações para produtores vinculados ao setor associativo brasileiro, e em *datasets* disponibilizados pelo governo do Brasil para o mesmo setor.

O objetivo do estudo é constatar a Aderência Lexical a dados publicados pelo governo do Brasil, usando como marco de análise necessidades informacionais de pequenos produtores.

Para atingir o propósito foram utilizadas revisão bibliográfica, técnica 5W1H, técnicas de mineração de textos e cálculo de similaridade entre textos mediante linguagem R, e software para planilhas eletrônicas.

Pretende-se formular argumentos que expliquem condições de acesso baseadas em eventos regulares nas fontes de dados escolhidas, visando, mediante o uso de categorias de necessidades informacionais, uma manifestação das proximidades lexicais entre dados de duas fontes, observada como elemento de análise dos fluxos de informação (Borko, 1968), e propendendo pelo aprimoramento da experiência humana ao participar de ambientes informacionais (Camperos-Reyes *et al.*, 2020).

2. PERCURSO METODOLÓGICO

Como caminho para entender os elementos circunstanciais que a perfilam a coleta dos dados, por um lado, grupos de produtores rurais, e de outro o detentor governo, foi usada a técnica 5W1H. Ela corresponde à sigla de seis questões, *What, Where, Who, When, Why, How* (O que, Onde, Quem, Quando, Porque, Como), ordenadas da menor à maior importância, proposta por Aristóteles, como um esquema retórico que permite descrever algo mediante enunciados coerentes (Sloan, 2010).

O uso de 5W1H se adota como plano de ação que de forma sistemática permite considerar tarefas necessárias para a compreensão do entorno das fontes de dados, e para atingir o fim da fase de coleta em elas, de forma que os resultados da sua aplicação entreguem elementos estruturais para a fundamentação do estudo.

Uma vez definidas as fontes dos dados, foram tratadas nas fases de coleta e análise dos dados conforme as suas particularidades. A mineração de texto foi uma das ferramentas usadas em ambas as fases, ela é um conjunto de técnicas que permite o descobrimento e extração de inferências de relevância a partir de textos escritos na linguagem natural e que não se encontram estruturados. Diversas ações podem ser realizadas com a mineração de textos, recuperação de textos na fonte escolhida, identificação de unidades de análise, classificação e agrupação de textos, bem como identificação de elementos de natureza conceitual a partir dos conteúdos coletados (Kao; Pottet, 2007).

A coleta de dados em comunicações publicadas para grupos de produtores rurais foi realizada mediante mineração de texto com a linguagem R[1], atividade executada no mês de março de 2021. Os dados do detentor governo foram coletados no site dados.gov.br mediante o uso dos descritores “pequeno produtor”, “desenvolvimento rural”, “associacao agricultura”, e “cooperativa agricultura”. A coleta realizada no dia 21 de abril de 2021 focou-se nos rótulos dos recursos em conjuntos de dados e nas descrições registradas pelos publicadores. Realizou-se extração manual desses objetos, justificando-se ao ser necessário discriminar a disponibilidade de cada recurso recuperado.

Para a análise dos dados foi também aproveitada a utilidade da Linguagem R. Consideraram-se as etapas funcionais da mineração de textos ou mineração textual, que segue as etapas da mineração de dados, onde a diferença fundamental com a primeira rege-se pela qualidade dos dados atingidos, pois, a mineração de dados trata unicamente com dados estruturados. Assim, as fases funcionais da mineração textual são pré-processamento dos dados, mineração central, camada de apresentação, e refinamento (Feldman; Sanger, 2006), onde as fases de pré-processamento e mineração central são as mais críticas do processo. Pré-processamento dos dados, fase também conhecida como de preparação dos dados (Castro; Ferrari, 2006), institui-se com técnicas que objetivam dispor os dados brutos, neste caso dados não estruturados, para assim depois serem analisados.

Uma vez pré-processados os dados obtidos das duas fontes, foram analisados na mineração central. Ambas as fases provêm as funcionalidades necessárias para atingir o objetivo desta pesquisa.

As tarefas realizadas na fase de pré-processamento, que alcançaram fins tanto da coleta como parte da análise dos dados, foram análise da estrutura dos sites, pastas e subpastas com os objetos a serem coletados, limpeza dos dados, agrupação e tokenização. Sendo que textos na linguagem natural se caracterizam por um fluxo contínuo de texto, para proceder com análises profundas é necessário que esse texto seja dividido em componentes significativos. Esses componentes significativos podem ser caracteres, fonemas, palavras, sintagmas, orações, parágrafos, seções ou capítulos; a segmentação depende do tipo de análise que o estudo demande e é conhecida como tokenização, ou divisão do texto em tokens (Feldman e Sanger, 2006).

Um dos aspectos tido em consideração durante a fase de pré-processamento foi a preparação dos dados para o estudo da aderência lexical que usou como marco de análise categorias de necessidades informacionais; isso direcionou para que a tokenização dos textos fosse realizada no nível de palavras.

Para a obtenção da similaridade como índice da aderência lexical entre dados publicados pelo governo e dados das comunicações para produtores, usando como marco de análise categorias de necessidades informacionais, foram usadas as bibliotecas da linguagem R '*tidytext'*, '*tm*' e '*philentropy'*; as funções *'findAssocs'* e *'jaccard'* fizeram trabalho conjunto para a classificação das notícias e conjuntos de dados conforme as necessidades informacionais, e para calcular o índice de similaridade entre grupos de palavras usou-se a *'jaccard'*. Este índice entrega um valor decimal que está entre 0 e 1, sendo que um valor próximo de zero indica baixa similaridade entre os textos comparados.

Conforme a estrutura adotada para os dados durante o pré-processamento, *Jaccard* foi observada como a função mais pertinente para calcular a similitude entre os conjuntos de palavras que foram conformados, conjuntos de tipo assimétrico: conjuntos de palavras advindos das comunicações para produtores, conjuntos de palavras advindos de rótulos e descrições de *datasets*, e conjuntos de palavras que descrevem necessidades informacionais de produtores.

Na mineração textual existem diversas funções que calculam similaridade entre conjuntos de dados; a escolha da função depende tanto dos fins da análise quanto da estrutura em que se encontram os dados. Há funções direcionadas a dados simétricos, assimétricos, segundo a natureza, dados binários, categóricos, numéricos, ou para combinações de tipos de dados, ainda podendo ser orientadas para análises de dissimilaridade em dados de essas e outras naturezas. Han et al. (2012) indicam *jaccard* como uma opção apropriada para analisar conjuntos de dados como os resultantes da fase de pré-processamento deste estudo. Ainda, considerando o fim dos estudos que usam mineração textual, Huang (2008), indica que, não objetivando análises de significado nos dados coletados, *Jaccard* é uma das opções adequadas.

2.1 Uso da técnica 5W1H

No elemento **"O que"**, optou-se por coletar textos de comunicações do tipo notícias realizadas em sites de grupos de produtores rurais no Brasil. O motivo dessa escolha é observar nas comunicações alguns dos seus rasgos comuns, identificando mediante técnicas de mineração de texto, indícios da aderência lexical a elementos de dados publicados pelo detentor governo. A linguagem usada nas comunicações, por ser geradas no entorno de agrupações de produtores rurais, é uma linguagem direcionada para alcançar esse público-alvo, portanto, presumem-se escritas com um teor adequado às características de grupos de produtores.

Do lado dados do detentor governo, optou-se pelas descrições e rótulos de *datasets* e os seus recursos, publicados dentro de categorias relacionadas com agricultura e desenvolvimento rural no site de dados abertos do Brasil.

Abordando os elementos **"Onde"** e **"Quem"** do lado do produtor rural, foi necessário avaliar alternativas, no cenário brasileiro, de organizações que tenham como usuários predominantes a grupos de produtores, e ainda, que contem com presença na internet mediante sites que publicam notícias para membros associados, associações e/ou cooperativas. Assim sendo, observando organizações de interesse nacional, foram explorados sites do Brasil e identificadas as organizações consignadas no Apêndice B.

Na construção da lista, iniciou-se por considerar instituições da ordem federal na procura de maior abrangência, motivo que permitiu deparar com a Confederação da Agricultura e Pecuária do Brasil (CNA).

A CNA congrega associações e lideranças políticas e rurais em todo o país. Entre suas funções está promover a geração de novas tecnologias para auxiliar ao produtor no plantio e manejo de lavouras, bem como ao fortalecimento das agroindústrias. A CNA congrega federações filiadas no país, constituindo-se como ponte entre as necessidades dessas federações e o Governo Federal, Congresso Nacional e Tribunais Superiores do Poder Judiciário. A intenção é que produtores rurais congregados em federações vejam na CNA um ator determinante para seus interesses, e como uma ponte com estamentos da esfera federal (Confederação da Agricultura e Pecuária do Brasil, 2022a).

A seguir, para os elementos **"Onde"** e **"Quem"** para dados do detentor governo, foi determinado recuperar os rótulos e as descrições de *datasets* no site de dados abertos do governo brasileiro, www.dados.gov.br, em conjuntos de dados recuperados mediante os termos "pequeno produtor", "desenvolvimento rural", "associacao agricultura" e "cooperativa agricultura".

Ao abordar o elemento **"Quando"**, para a coleta automática dos dados no site escolhido, o da CNA, determinou-se coletar notícias que foram publicadas nos anos de 2019 a 2020.

Do lado do detentor do governo, uma vez explorados os conjuntos de dados, optou-se por coletar todos os recursos publicados até a data da coleta, 28 de abril de 2021, sem descartar algum conjunto de dados. Tomou-se esta determinação devido a que a análise não precisa delimitar a temporalidade dos dados disponíveis desde o lado governamental pois, além de não apresentar um alto volume de recursos, é necessário esgotar as possibilidades de unidades de análises nos conjuntos de dados do lado governo.

"Porque" coletar comunicações do tipo notícias nos sites escolhidos? A coleta dos dados contidos em comunicações, publicadas em sites que divulgam informações para grupos de produtores rurais, se sustenta ao entender-se como meios de manifestação de assuntos e necessidades informacionais dos produtores e para o fortalecimento das suas empresas; considera-se que são objetos determinantes para o sucesso e viabilidade das produções.

Em relação ao **"Porque"** do uso dos rótulos e descrições dos conjuntos de dados, explica-se em razão de que eles são referências tangíveis ao conteúdo que têm os usuários dos sites de dados abertos. Esses rótulos manifestam as características dos dados outorgando aos usuários uma aproximação às possibilidades de uso de aqueles recursos.

O elemento **"Como",** do lado do produtor rural, realizou-se mediante mineração textual. Dada a quantidade de informação que repousa nas notícias divulgadas, disponíveis como dados não estruturados, é possível desvendar inferências sobre a aderência lexical a dados publicados pelo governo ao observar esse texto em palavras como unidades significativas.

Para realizar a coleta automatizada dos dados no site CNA, foi necessário analisar a estrutura geral do site, www.cnabrasil.org.br, a arquitetura da informação do site, nos sistemas de organização e navegação (Rosenfeld et al., 2015). Era necessário descobrir os locais onde são publicadas as notícias, e na sequência, observar a estrutura seguida pelo site em relação à designação das URL das notícias e os elementos CSS contidos.

No site da CNA foi identificado o subdiretório disposto para a publicação das notícias, https://www.cnabrasil.org.br/noticias. A exemplo, nessa URL de notícia, https://www.cnabrasil.org.br/noticias/queijos-puxam-queda-do-mercado-de-lacteos, é possível observar que além do domínio aparecem os subdiretórios "notícias" e o próprio da notícia publicada.

Para coletar especificamente as notícias no período da pesquisa, atentou-se às possibilidades do site, sendo observado um sistema de busca que permite a filtragem das notícias. Possui filtros por Instituição, Área de atuação, Tipo de conteúdo, Data de início e Data fim. O único filtro aplicado foi o período 01/jan/2019 até 31/dez/2020.

Assim, em um primeiro momento, para a codificação do algoritmo da coleta dos dados, foi observada a composição das URL das notícias resultantes à aplicação da filtragem. A primeira página dos resultados apresenta o endereço https://www.cnabrasil.org.br/noticias?instituicao=cna&termo=undefined&firstdate=01-01-2019&enddate=31-12-2020, entregando detalhes de como o site organiza na URL os parâmetros de busca e uma possível numeração de todas as páginas recuperadas após a filtragem.

Conforme observação das URL após a filtragem, identificou-se ainda que o próprio site acrescenta o elemento "p1" à URL desde o primeiro resultado da busca, agindo como controlador das páginas no formato "p*n*"[2], sendo esse um contador que possibilitou coletar de forma automática cada página dos resultados mediante programação em R.

Assim sendo, as URL consideradas são as compreendidas entre a https://www.cnabrasil.org.br/noticias/**p1**?instituicao=cna&termo=undefined&firstdate=01-01-2019&enddate=31-12-2020, primeira página do resultado da filtragem, e, https://www.cnabrasil.org.br/noticias/p174?instituicao=cna&termo=undefined&firstdate=01-01-2019&enddate=31-12-2020, última página do mesmo resultado.

O texto contido nas notícias para grupos de produtores rurais, apresenta informações que se encontram marcadas apenas em atributos que o site decide usar durante o seu desenvolvimento, de modo que um uso automatizado de elementos textuais torna-se desafiante.

Diante disso, para identificar os nós que seriam coletados, foi usada a ferramenta SelectorGadget para reconhecer os elementos CSS[3] nas páginas de notícias, os quais constituíram os atributos a serem coletados. Com o foco da pesquisa no conteúdo das comunicações, foram identificados três elementos para a análise: título, data e corpo da notícia, os quais têm por nome dos elementos em notação CSS, .entry-title, small e .content-body, respectivamente.

A Figura 1 apresenta a sequência de passos realizados durante a coleta e análise dos dados do lado dos produtores rurais, já descritos até aqui, e os passos realizados para a coleta e análise dos dados do lado governo.

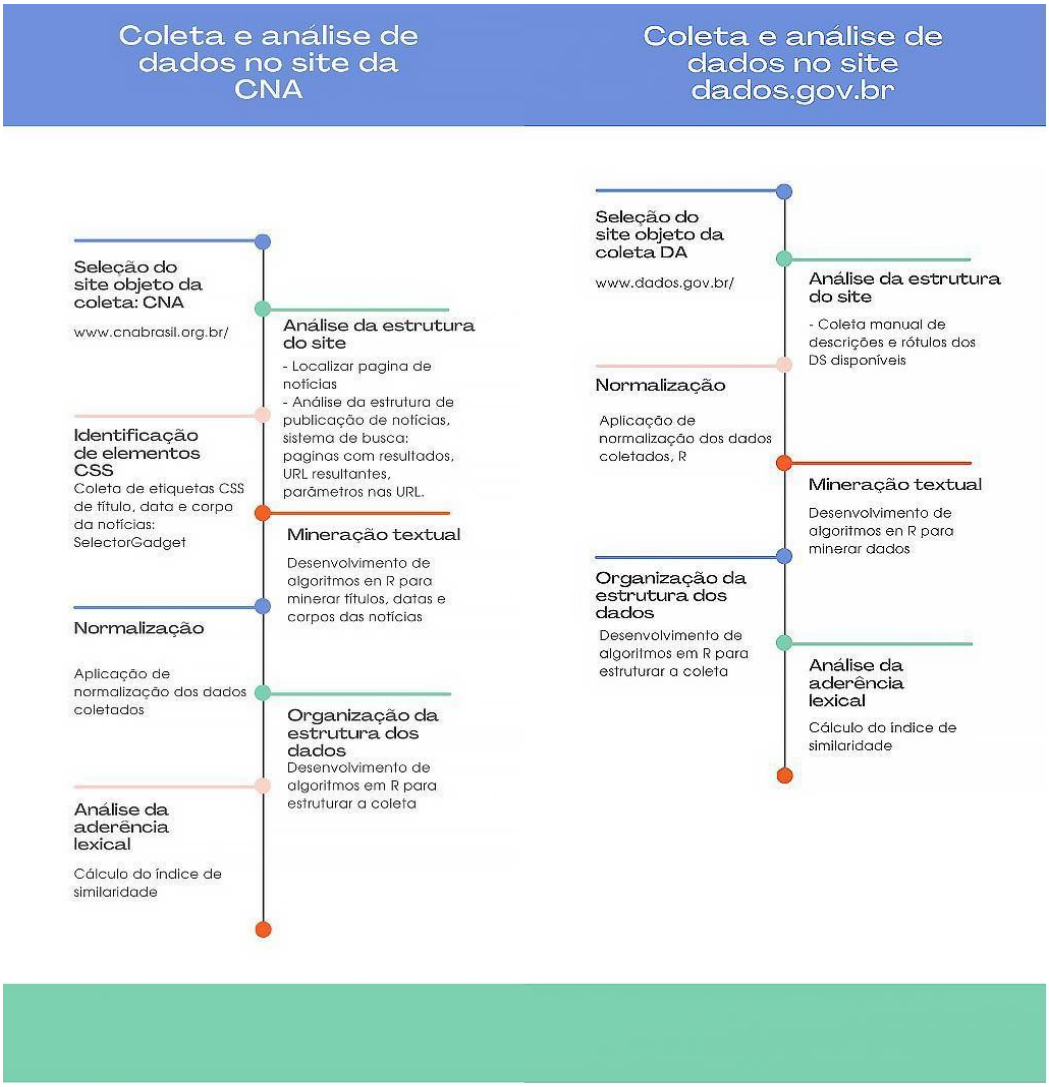

Figura 1. Diagrama coleta e análise de dados no site da CNA e no dados.gov.br. Fonte: elaborado pelos autores.

O elemento **"Como"** da coleta e análise do lado governo, dados de *datasets*, descrições e rótulos nos conjuntos de dados, foi abordado em um primeiro momento de forma manual, coletando e organizando em planilhas eletrônicas, que foram posteriormente processadas em R para estruturar os dados e realizar análise idêntica à primeira fonte indicada.

3. MINERAÇÃO TEXTUAL COMO TÉCNICA DE ANÁLISE

Abordar documentos escritos na linguagem natural com ferramentas tecnológicas propicia amplas possibilidades de análise. Em um primeiro momento, pelas capacidades *per se* da linguagem natural como sistema de signos expressivo e predileto na comunicação (Korn et al., 1988); de outro lado, técnicas informáticas permitem revelar informações em grandes corpos textuais, que aproveitando das capacidades da linguagem natural, viabilizam a obtenção de inferências conforme interesses de pesquisa em um contexto determinado.

Apelando ao princípio da causalidade, Korn et al. (1988) assinalam que embora todas as situações estejam sujeitas à lógica, ela não está sempre evidente de forma imediata. Formular argumentos explicativos é uma atividade que pode alicerçar-se em condições que ocorrem com alguma regularidade, possibilitando, por exemplo, mediante descrições qualitativas, classificações em categorias que descrevem uma situação. Isto tudo no entendimento de que a riqueza obtida pelo uso da linguagem natural, pela complexidade da estrutura sintática e pelas possibilidades do uso funcional do vocabulário, pode requerer o apoio de ferramentas tecnológicas ao abordar corpus volumosos, diante da intenção de explicitar a lógica de uma situação determinada.

Nesse quadro, manifesta-se a necessidade do apoio da abordagem científica, no que pode contribuir a Ciência de Dados ou *Data Science*, que corresponde à articulação interdisciplinar que procura a extração de insights a partir de dados (Dutra, 2021; Grus, 2015; Sant'ana e Rodrigues, 2020). Segundo a tipologia dos dados a serem abordados, podem ser usadas técnicas para mineração de dados, *Data Mining*, ou técnicas para mineração de textos, *Text Mining*. A mineração de dados usa como insumo dados estruturados, enquanto a mineração de textos tem como insumo dados semiestruturados ou não estruturados, ambos os conceitos, na procura de inferências e informações contidas em dados dentro do contexto digital (Feldman e Sanger, 2006; Castro e Ferrari, 2016; Silge e Robinson, 2017).

Da mesma forma que a mineração de dados, a mineração de textos busca extrair informação útil das fontes mediante a identificação de padrões, tendências, índices, informações relevantes, segundo o interesse de estudo que aborda as coleções de textos. Assim, os padrões não se extraem a partir de registros ou do cruzamento de entidades de alguma estrutura de dados senão do texto fluido que conforma os documentos a serem minerados (Feldman e Sanger, 2006).

A análise automatizada de textos permite descobrir o desconhecido, encontrando pedras preciosas em corpos textuais porque seus resultados revelam insumos inéditos, que, de outra forma não automatizada, dificilmente poderiam ser descobertos. A mineração de texto objetiva "descobrir ou derivar novas informações a partir de dados, encontrar padrões em conjuntos de dados e/ou separar um sinal do ruído" (Hearst, 1999, 3). Em consequência, o corpus textual que age como insumo, estará disposto para a geração de reportes estatísticos que possibilitem insights acerca do ambiente informacional abordado (Dutra, 2021; Hearst, 2003; Kao e Poteet, 2007).

Fontes de textos não estruturados vão desde livros, e-mails, documentos empresariais, notícias, web sites e post em redes sociais, que possibilitam aplicações nas ciências sociais aplicadas, administração, ciências políticas, ciências da saúde etc. Algumas práticas da mineração textual estão no monitoramento de opiniões, análise de sentimentos, estudos de psicologia, saúde pública, análises de discurso político, monitoramento de reputações, visões e atitudes, repercussão de eventos, e, como neste caso de estudo, como índice da proximidade entre dados publicados por governo e um grupo de usuários escolhido.

As técnicas usadas são amplas, podendo incluir conceitos da estatística descritiva, frequências de unidades significativas, coocurrência de palavras, análise de colocação, similaridade entre textos, classificação de textos, agrupamentos temáticos, sumarização etc. (Kao e Poteet, 2007; Silge e Robinson, 2017).

4. RESULTADOS E DISCUSSÕES

Efetuadas as fases de pré-processamento dos dados e a mineração central, foram obtidos os seguintes resultados do site da CNA: nas 174 páginas após filtragem permitiram recuperar 2073 notícias; a seguir, normalizando os dados e tokenizando em palavras cada notícia, resultaram 546.571 palavras.

É necessário indicar que essa quantidade de palavras é posterior à eliminação de *stopwords*, palavras que apareceram com frequências altas, mas que carecem de relevância ao serem palavras gramaticais, artigos, preposições, conjunções, bem como algumas outras de pouca significância, rotulagem institucional, páginas das redes sociais da CNA, números etc. Remover palavras pouco significativas potencializa o resultado de análises com mineração textual, sendo uma das práticas recomendadas e indicadas, que na maioria das vezes, implica criar listas customizadas acrescentando palavras que surgem conforme a fonte de dados escolhida (Silge e Robinson, 2017).

Em seguida se apresentam os elementos recuperados na coleta de dados do lado governo estão (Quadro 1).

Descritor	*Datasets*	Disponíveis	Recursos relacionados	Disponíveis	% Disp. de Recursos
Desenvolvimento Rural	15	6	293	182	62%
Cooperativa Agricultura	5	3	54	41	76%
Associacao Agricultura	1	1	20	20	100%
Pequeno Produtor	11	8	32	8	25%
	32	**18**	**399**	**251**	**63%**

Quadro 1. *Datasets* recuperados. Fonte: elaborado pelos autores a partir dos dados coletados.

Foram processados 251 recursos publicados em 18 conjuntos de dados disponíveis, e como resultado do pré-processamento e mineração central dos dados, foram identificadas 4.159 palavras.

Como marco de análise na busca por identificar a aderência lexical entre as duas fontes de dados, foram consideradas categorias de necessidades informacionais no acesso a dados por produtores (Camperos-Reyes, 2023), a saber, Mercado, Tratos culturais, Crédito, Direitos e Oportunidades.

As categorias de necessidades informacionais foram usadas com o fim de identificar se as unidades coletadas, tanto as notícias como os *datasets*, têm relação com assuntos de relevância demandados por produtores rurais, elas constituem um marco de análise intermediário entre as duas fontes de dados escolhidas para identificar a proximidade lexical entre uma e outra fonte.

Para isto foi usada a função da R '*findAssocs*', que permitiu descobrir quais palavras aparecem com maior frequência quando encontrado o nome de categoria. Para manter a análise no nível lexical, determinou-se usar as seguintes palavras para o cálculo da função: "mercado", "cultura", "crédito", "direitos" e "oportunidades[4]". Dessa forma, nas notícias e *datasets* ao aparecer, por exemplo, a palavra "mercado", foram elencadas as palavras com que existe maior coocorrência, isto em cada categoria de necessidade informacional.

Ao aplicar a função, a lista de palavras resultantes foi analisada na ordem decrescente do índice de coocorrência, que na função apresenta-se com valores entre 0 e 1, sendo que valores próximos a 1 indicam maior coocorrência. Como critério de escolha das palavras foram consideradas apenas palavras do tipo substantivo, sendo assim escolhidas as cinco mais coocorrentes em cada categoria[5]:

- Mercado: preço, oferta, demanda, exportações, importações;
- Cultura: beneficiadora, declividade, descaroçado, rotação, fiação;
- Crédito: financiamento, juros, financieras, custeio, taxas;
- Direitos: creditórios, deveres, consif[6], oit[7], violação;
- Oportunidades: negócios, mpog[8], agroexportador, carreira, exportação.

Algumas palavras identificadas não foram consideradas devido a que manifestam relação estreita com atividades ou atores específicos. É o caso de esalq, acopar, cotonicultores, trigo, algodão, triticultores etc., observadas nas coocorrências de 'mercado' e 'cultura'. Outras, identificadas no cálculo com a palavra 'oportunidades', indicam casos específicos de relações com outros países, é o caso das palavras botsuana, coreanas, egípcios, marroquinos, namíbia, suazilândia, china, islâmica, asiático, halal, mexicanos e britânicos. Ainda não foram considerados nomes de cidades, países, instituições, gentílicos e sobrenomes.

O resultado da função possibilitou conformar um vetor de comparação geral para cada categoria de necessidade informacional, atentando a que a unidade de comparação são unidades lexicais e adicionando, outros descritores (Camperos-Reyes, 2023) por cada categoria.

Os vectores de comparação para identificar com qual categoria de necessidade informacional se relaciona cada notícia e cada dataset, foram estabelecidos como segue:

- Mercado: ("mercado", "preços", "oferta", "demanda", "exportações", "importações", "publicidade", "consumo", "comercialização");
- Tratos culturais: ("cultura", "beneficiadora", "declividade", "descaroçado", "rotação", "fiação", "irrigação", "fertilização", "pragas");
- Crédito: ("crédito", "financiamento", "juros", "financeiras", "custeio", "taxas", "microcrédito", "empréstimo", "subsídios");
- Direitos: ("direitos", "creditórios", "deveres", "consif", "oit", "violação", "legais", "autor", "registro");
- Oportunidades: ("oportunidades", "negócios", "mpog", "agroexportador", "carreira", "exportação", "associativismo", "cooperativismo", "empreendedorismo").

Em seguida foi utilizada a função '*jaccard*' para obter a similaridade entre cada notícia e os vetores. Assim, obtiveram-se os seguintes resultados: 1170 notícias têm similaridade com a categoria Mercado, 289 com Tratos culturais, 501 com Crédito, 78 com Direitos, e 670 com Oportunidades. O cálculo de similaridade entre cada conjunto de dados de governo e os vetores de comparação, identificou que unicamente há conjuntos de dados relacionados com a categoria Crédito, 8 no total, sendo que com as outras 4 categorias de necessidades informacionais o índice de similaridade usando a função '*jaccard'* foi de zero, não houve similaridade.

Até este ponto, a função '*jaccard*' (Equação 1), foi usada para determinar a proximidade entre cada unidade de análise, bem notícias, bem conjuntos de dados, e as categorias de necessidades informacionais. Logo, conforme esses resultados, foi calculada a similaridade entre as palavras usadas nos dados de governo e nas notícias, apontando que somente foi calculado onde houve relação de forma concomitante na categoria de necessidade, neste caso, com a categoria 'Crédito', única onde houve proximidade.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Equação (1)

O Quadro 2 resume as quantidades de palavras extraídas e os índices *Jaccard* para a categoria 'Crédito':

Unidades extraídas			**Índice *Jaccard* calculado**		
Categoria Crédito			**Categoria Crédito**	**Máximo possível**	**Relação obtido/máximo**
	Notícias	***Datasets***			
Total palavras	175.818	800	0,54%	0,667098%	80,5%
Palavras únicas	15.440	103			

Quadro 2. Palavras extraídas e índices *Jaccard* da categoria 'Crédito'. Fontes: os autores com base na coleta dos dados.

Devido a que as quantidades de palavras são expressivamente diferentes, determinou-se obter o *Jaccard* máximo possível entre as fontes dos dados, calculando as palavras únicas dos conjuntos de dados de governo por ser o vector menor, 103 palavras, sendo possível assumi-lo como o valor máximo da intersecção com as palavras das notícias. É com esse valor que foi calculado o *Jaccard* máximo, o índice 0,00667098 (0,667098%). Dessa forma, a relação entre os *Jaccard* da categoria 'Crédito" e o máximo possível, resulta em 0,804782, lido como que 80,5% das unidades lexicais usadas nos dados de governo, foram similarmente encontradas nas notícias mineradas do site da CNA.

Unidades extraídas			**Índice *Jaccard* calculado**		
	Notícias	***Datasets***	**Total**	**Máximo possível**	**Relação obtido/máximo**
Total palavras	546.571	4.159	0,854983%	1,03665%	82,48%
Palavras únicas	27.203	282			

Quadro 3. Palavras extraídas e índices *Jaccard* total. Fonte: elaborado pelos autores a partir dos dados coletados.

Ao observar o conjunto total de palavras tanto dos dados de governo como das notícias, foi possível observar que 82,48% das unidades lexicais usadas nos dados de governo, foram similarmente encontradas nas notícias mineradas do site da CNA. Algumas das palavras similares encontradas nas duas fontes tanto na categoria Crédito como no conjunto de unidades lexicais estão no Apêndice A.

Interessa destacar algumas palavras identificadas unicamente na coleta do lado governo, as quais manifestam um nível de especificidade que merece reflexão por parte do detentor e os seus publicadores. Um exemplo é o caso de siglas como BSM[9], SIATER[10], SIMOG[11], UPFS[12], que, correspondendo a atributos de programas de governo, manifestam a necessidade de aproximação por parte dos sujeitos alvo, ao menos de forma inicial, a conceitos que esclareçam essas palavras encontradas. Em alguns casos, nomes por extenso são indicados nas descrições dos *datasets*, porém, nem todos apresentam essa caraterística.

Outros exemplos de palavras resultantes do lado governo são as observadas ao retirar caracteres especiais como o '_', pois, na forma original, dito carácter permite detalhar atributos dos dados, a exemplo, os rótulos N_SOCIOS_COM_DAP_PF, N_SOCIOS_TOTAL_PJ, que uma vez limpos durante o processamento dos dados resultam em unidades lexicais como nsocioscomdappf e nsociostotalpj, onde, não havendo correspondência do lado das notícias para o produtor, também não manifestam inconveniente que leve a apontar necessidade de melhoria das estratégias de disponibilização, pelo menos dentro do escopo desta pesquisa. Uma alternativa para calcular o índice *Jaccard* foi a executada mediante a extração dos lemas das palavras identificadas na coleta. A lematização refere-se à eliminação das terminações flexivas das palavras para devolver a forma base delas. O processo é realizado baseado no conjunto de palavras de um dicionário de uma língua e na morfologia das palavras (Wachelke e Wolter, 2011). Assim, por exemplo, as palavras 'produtor' e 'produtores' estarão agrupadas em um mesmo lema. Dessa forma, analisando a coleta com as unidades na forma de lemas, obtiveram-se índices *Jaccard*, na categoria crédito de 0.00786304, e, no total dos dados coletados, 0.01282410. Esses valores, mais altos que os calculados nas palavras conforme coletadas nas fontes, permitem inferir que a lematização constitui uma opção conveniente para a análise de aderência entre duas fontes, conclusão equivalente ao estudo de Ross e Cruz (2021).

4.1 Relação de n-grams com as categorias de necessidades informacionais

Como uma análise derivada dos dados coletados, observou-se o corpus das notícias na forma de n-grams, iniciando com palavras únicas, totalizando o número de vezes que foi usada cada palavra, e em seguida, mediante uma análise de colocação, essas palavras tokens foram processadas junto às palavras consecutivas no corpus de notícias coletadas, sendo extraídas sequências de 2 e 3 palavras, bigramas e trigramas. As Figuras 2 e 3 apresentam um informe estatístico dos bigramas coletados no site do CNA.

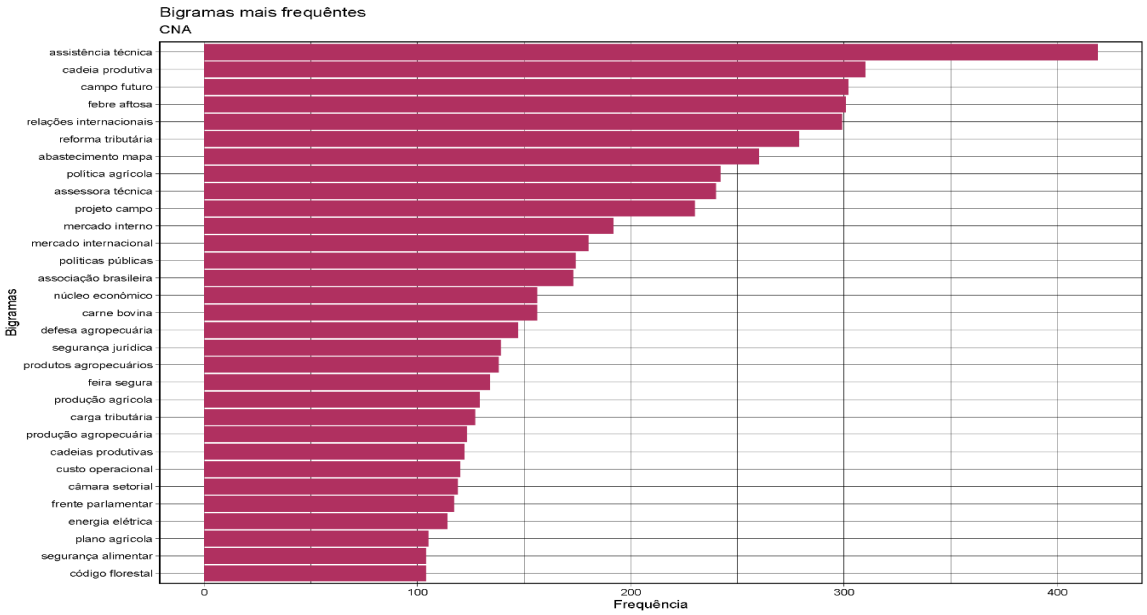

Figura 2. Histograma com bigramas mais frequentes no CNA. Fonte: elaborado pelos autores com base na coleta dos dados.

O bigrama mais frequente foi "assistência técnica" com mais de 400 ocorrências, seguida pelas duplas "cadeia produtiva", "campo futuro", "febre aftosa", e "relacoes internacionais".

Associando esses bigramas a temáticas de interesse nas comunicações da CNA, pode inferir-se a dedicação com aspetos práticos associados às unidades produtivas, assinalada na ação de assistência ligada à cadeia produtiva, portanto, uma forte relação com a categoria de necessidade informacional 'Tratos culturais'.

O "campo futuro" corresponde a um projeto da CNA junto ao SENAR, cuja finalidade é "aliar a capacitação do produtor à geração de informações estratégicas do setor rural, contribuindo para as tomadas de decisão no campo" (Confederação da Agricultura e Pecuária do Brasil, 2022b, *[s.p.]*). Mediante ações realizadas em parceria com universidades e centros de pesquisa, o Campo Futuro realiza acompanhamento à evolução de custos, análises de rentabilidade, gerenciamento de preços e de comportamento da produção (Confederação da Agricultura e Pecuária do Brasil, 2022b). Observa-se relação desse bigrama com a categoria 'Oportunidades'.

Em relação ao bigrama "febre aftosa", sabe-se que o Brasil é o segundo maior produtor mundial de gado bovino (Portal DBO, 2021), com 218,2 milhões de cabeças em 2020 (Instituto Brasileiro de Geografia e Estatística, 2020). Ela é uma doença de notificação obrigatória e atinge "animais de produção como bovinos, suínos, caprinos, ovinos e outros animais, em especial os de cascos bipartidos (cascos fendidos)" (Brasil, 2022). Esse bigrama se relaciona com a categoria 'Tratos culturas'.

Várias estratégias nacionais, e por estados do Brasil, são implementas, e conformam elementos de atenção tanto do governo como dos produtores junto aos coletivos que os conglomeram, considerando sobretudo pela representatividade da renta que o setor produz, e pode, inclusive enlaça-se com o quinto bigrama, "relações internacionais" pois o setor de gado de corte, particularmente o bovino, representa o 5 lugar no ranking das exportações totais e o 2 lugar nas exportações da indústria de transformação, tendo como destinos, na ordem de importância, China, Chile, Estados Unidos e Egito (Agência Brasil, 2021; Ministério da Indústria, Comércio Exterior e Serviços, 2022), observando assim relação com a categoria 'Oportunidades'.

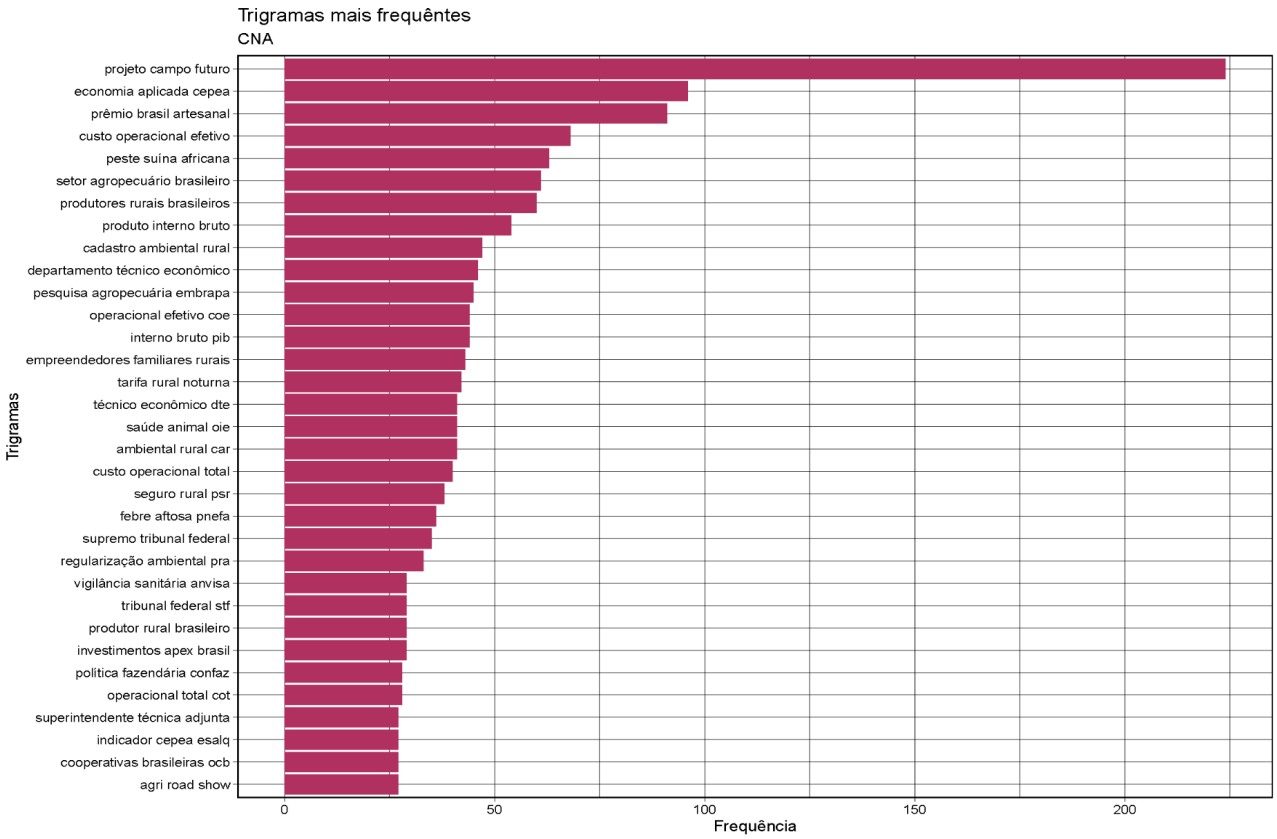

Figura 3. Histograma com trigramas mais frequentes no CNA. Fonte: elaborado pelos autores a partir dos dados coletados.

Já havendo abordado o "projeto campo futuro", tratar-se-á o concernente ao "economia aplicada cepea". O Centro de Estudos Avançados em Economia Aplicada (CEPEA), da Escola Superior de Agricultura Luiz de Queiroz (ESALQ), Universidade de São Paulo (USP), é um grupo de pesquisa com foco em temas como meio rural e setores

econômicos relacionados ao agronegócio, fundamentalmente em instrumentos de índole econômico-administrativo. Alguns dos eixos de trabalho estão no contexto das cadeias produtivas, questões sanitárias, políticas comerciais, novas tecnologias, e desempenho macroeconômico do setor (Centro de Estudos Avançados em Economia Aplicada, 2022a).

O centro fornece indicadores, índices, listas de insumos pecuários, acompanhamento ao mercado de produtos como grãos, gado de corte, ovos, produtos florestais e hortifrutis (Centro de Estudos Avançados em Economia Aplicada, 2022b), elementos que constituem bens informacionais para os produtores, o que pode sustentar o porquê esse trigrama, relacionado com 'Tratos culturais', está alocado na segunda posição.

Dentro do "Programa de alimentos artesanais e tradicionais", programa dos CNA/SENAR, são promovidas melhorias para produtores apoiando-os nos eixos Regulamentação, Capacitação e assistência técnica e gerencial, Comercialização e marketing, Organização coletiva, e, Tributação e crédito; no eixo Comercialização e marketing, foi implementado o Prêmio Brasil Artesanal, destacando até a data, desenvolvimentos em produtos de chocolate e charcutaria (CNA/SENAR, 2022). Isto pode explicar a aparição do trigrama "premio brasil artesanal", pois no site do CNA foram divulgadas informações sobre as fases da execução da estratégia pelo CNA/SENAR. A relação do trigrama está com a categoria 'Mercado'.

De outro lado, na gestão de empreendimentos rurais alguns indicadores econômicos fundamentais são usados na tomada de decisão dos produtores e em relação ao processo produtivo e comercial das suas lavouras. Um desses indicadores é o que foi encontrado no trigrama "custo operacional efetivo", portanto, relacionando-se com divulgações dirigidas aos fluxos informacionais dos produtores em um aspecto determinante como é a gestão de custos.

Interessante apontar que o trigrama "economia aplicada cepea" embora aponte a duas categorias 'Mercado' e 'Tratos culturais', como no caso de "custo operacional efetivo", estão também assinalando aspectos sobre instrumentos de gestão económica e administrativa, os quais não estão incluídos nas categorias que esse estudo usa como marco de análise. Em relação ao último trigrama a considerar, "peste suína africana", trata-se de uma doença que, embora não ofereça risco para os humanos, é altamente contagiosa no gado suíno, e já causou enorme prejuízo em países como China, Polônia, Romênia e Filipinas. No Brasil, tendo aparecido na década dos anos 70, foi controlada em 1984. O Brasil é o quarto maior produtor de carne suína no nível global (Empresa Brasileira de Pesquisa Agropecuária, 2022; UOL Economia, 2022). Configura-se relação desse trigrama com a categoria 'Tratos culturais'. Do lado da coleta no site de dados do Brasil, a análise de colocação permitiu gerar os seguintes elementos gráficos (Figuras 4 e 5).

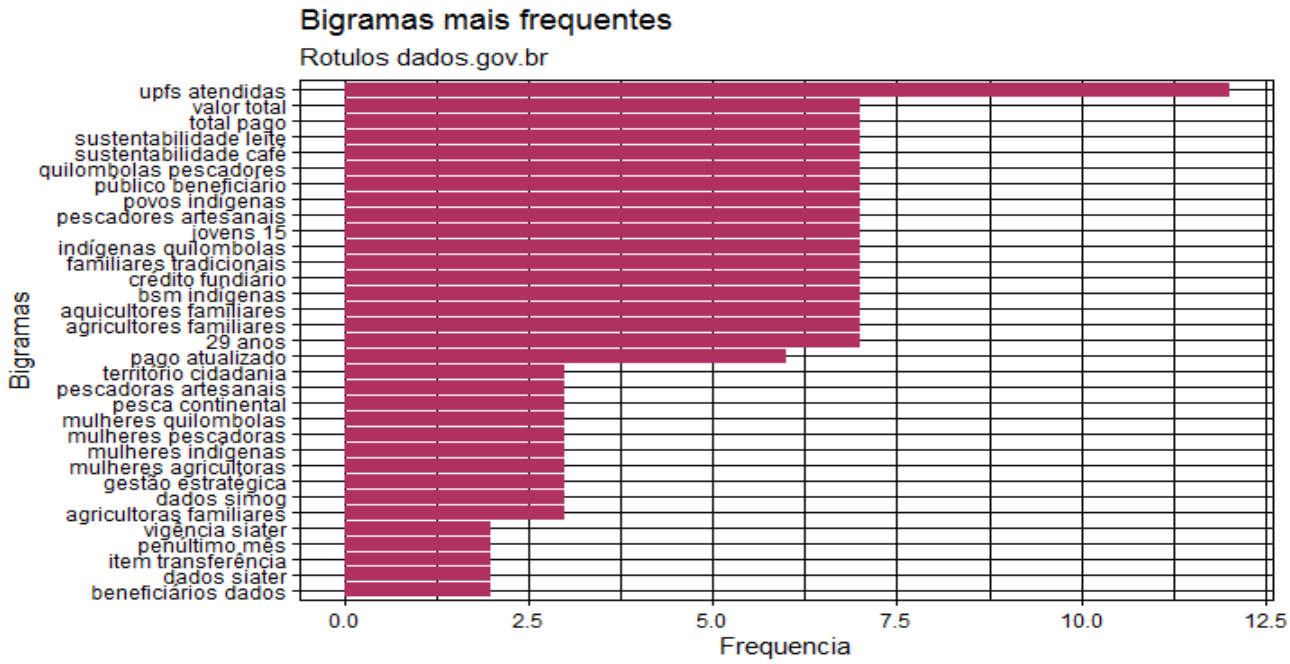

Figura 4. Histograma com bigramas mais frequentes em dados.gov.br.
Fonte: elaborado pelos autores a partir dos dados coletados.

O bigrama “upfs atendidas” está registrado devido à rotulagem relacionada com Unidades Produtivas Familiares (UPFS), atendidas nas chamadas do serviço de Assistência Técnica e Extensão Rural (ATER).

Nos casos de “sustentabilidade leite”, “sustentabilidade café” e “quilombolas pescadores”, correspondem a atributos que caracterizam também chamadas aos serviços do ATER. Ele classifica as chamadas segundo o tipo de contrato dos beneficiários dos serviços. Existem categorias para os contratos segundo a sua natureza, sendo que três delas são Sustentabilidade Leite, Sustentabilidade Café, e BSM - Indígenas/Quilombolas/Pescadores[13].

Por fim, a categoria “público beneficiário”, indica o rótulo das formas de classificar os beneficiários dos serviços ATER, tendo como exemplos desses tipos de públicos beneficiários a Agricultores Familiares Tradicionais, Aquicultores Familiares, Pescadores Artesanais, Povos Indígenas e Quilombolas.

Assim sendo, frisa-se que os cinco bigramas mais frequentes se encontram dentro de *datasets* que apresentam resumos de chamadas aos serviços do ATER, intitulados “ATER - Chamadas em Atendimento”. O ATER foi criado dentro da Política Nacional de Assistência Técnica e Extensão Rural para a Agricultura Familiar e Reforma Agrária (PNATER), e está sob a gestão do Ministério da Agricultura, Pecuária e Abastecimento do Brasil, logo, pode se inferir que todos os bigramas do top 5 estão relacionados com a categoria ‘Tratos culturais’.

De outro lado, o processamento na forma de trigramas (Figura 5) mostra que os cinco trigramas mais representativos foram “indígenas quilombolas pescadores”, “bsm indígenas quilombolas”, “agricultores familiares tradicionais”, “mulheres pescadoras artesanais” e “mulheres agricultoras familiares.

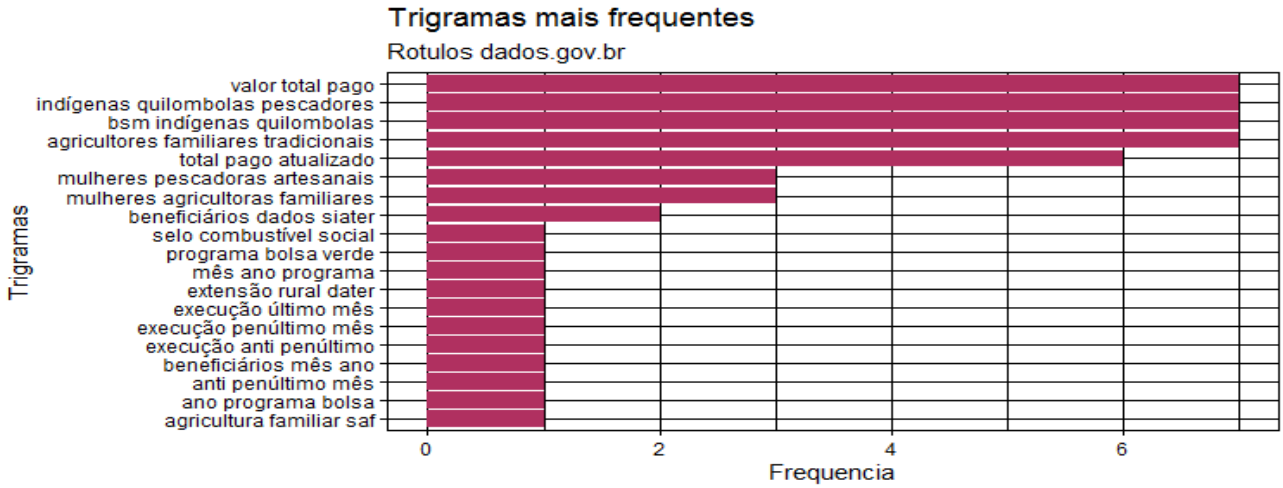

Figura 5. Histograma com trigramas mais frequentes em dados.gov.br. Fonte: elaborado pelos autores a partir dos dados coletados.

Os trigramas “indígenas quilombolas pescadores” e “bsm indígenas quilombolas”, correspondem a rótulos que fazem referência à categoria “BSM - Indígenas/Quilombolas/Pescadores”, a qual é um tipo de contrato dos beneficiários do ATER. No caso dos trigramas “agricultores familiares tradicionais”, “mulheres pescadoras artesanais” e “mulheres agricultoras familiares”, correspondem a rótulos que representam formas de classificar aos beneficiários dos serviços ATER.

Em suma, todos os trigramas do top 5 foram extraídos de *datasets* onde se publicam dados sobre chamadas em atendimento do ATER, e, portanto, também relacionados com a categoria ‘Tratos culturais’.

O exemplo de identificação de unidades *n-grams* foi usado aqui como proposta para a descrição dos assuntos tratados nas notícias e *datasets* coletados, e, portanto, como uma alternativa para a valorização dos interesses nas comunicações e dados disponibilizados para produtores rurais.

5. CONCLUSÕES

As fontes de dados escolhidas para a coleta entregaram insumos suficientes para a análise proposta, cada uma em seu contexto e com as particularidades técnicas na disponibilização dos dados, permitiram gerar as unidades lexicais que indicariam um caminho para a observação da aderência entre dados de uma e outra fonte.

Os instrumentos técnicos usados para o cálculo da aderência lexical entre dados obtidos por mineração textual se manifestaram propícios para identificar indícios de interpretação por parte de grupos de produtores rurais.

Ao implementar a análise usando como marco estruturante categorias de necessidades informacionais é importante atentar que, do lado governo, unicamente houve proximidade com uma das categorias utilizadas, Crédito, o que certamente aponta para uma insuficiência de elementos de análise no caminho delineado por este estudo. Ainda mais quando as duas categorias amplamente presentes nas comunicações da CNA foram Mercado e Oportunidades, demandas que, na amostra do estudo, não foram tratadas nos dados do governo. Considera-se determinante para o aproveitamento de dados publicados pelo detentor estejam estritamente relacionados com as necessidades informacionais dos sujeitos alvo.

Os índices resultantes, tanto na única categoria de necessidade informacional que foi atingida de forma concomitante nas duas fontes, 'Crédito', como no total de unidades analisadas, sustentam a potencial aderência lexical como indício da probabilidade de interpretação dos dados que o governo está publicando.

A análise de colocação apresentada, identificação de n-grams, se perfila como alternativa para a descrição de assuntos tratados em corpos de origem textual, podendo ser explorada buscando relações com elementos estruturantes de análise ou como proposta de entendimento de atributos do campo informacional do contexto abordado.

Limitações do estudo são observadas nas origens dos dados. Frisa-se a parcialidade das comunicações publicadas pela CNA, pelo direcionamento para atividades econômicas de interesse particular; outro elemento limitador corresponde com aspectos da qualidade e assertividade da rotulagem dos dados do detentor governo. Frické (2013) manifesta que diante da possibilidade de referir-se a um conceito de distintas formas, situação que o autor extrapola a ações de rotulagem, a assertividade na designação de rótulos com a responsabilidade de indicar um conceito com uma ou outra palavra, tem implicações determinantes nos resultados de análises deste tipo.

Outra limitação, dá-se pela possibilidade de que o procedimento técnico que permitiu formar os vetores de comparação, não considere palavras do tipo sigla que estejam no corpo das comunicações usadas.

Estudos futuros apontam à diversificação das fontes dos dados, tanto do lado detentor governo como na aproximação aos sujeitos alvo. Na continuidade ao nível lexical, propõe-se desdobrar este estudo utilizando instrumentos metodológicos que permitam obter indícios do nível semântico, apontando então à diversificação das análises junto aos instrumentos de coleta, sobretudo do lado do pequeno e médio produtor, pois, seguindo a premissa dos Santos e Sant'Ana (2019), na camada de localização dos dados, o acesso a eles está atrelado ao ponto de vista das necessidades dos usuários.

Por outro lado, a Mineração Textual oferece técnicas de extração não supervisionada como o *Topic Modelling*, que poderia ser usada de forma alternativa para modelagem temático de corpus.

Espera-se ter contribuído na busca da resposta ao interrogante sobre o aproveitamento, desde o ponto de vista do sujeito alvo dessa pesquisa, de dados publicados pelo detentor governo no contexto de grupos de produtores rurais, sobre se o governo traça de forma intrínseca um uso potencial dos dados por parte dos produtores, considerando fatores de acesso a dados do lado de quem precisa deles (Santos; Sant'ana, 2019), tais como conhecimento prévio, domínio, interesses e valores.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - Código de Financiamento 001.

6. REFERÊNCIAS

Agência Brasil. (2021). *Brasil exporta menos carne, mas registra aumento de faturamento: De janeiro a maio foram exportadas 710.093 toneladas*. https://agenciabrasil.ebc.com.br/economia/noticia/2021-06/brasil-exporta-menos-carne-mas-registra-aumento-de-faturamento

Barreto, A. de A. (2002). O tempo e o espaço da ciência da informação. *Transinformação*, v. 14, n. 1, 17-24, jan/jun. 2002. https://www.scielo.br/j/tinf/a/H3pxvkm6ZjBKNfMLsp7Gfrt/?format=pdf&lang=pt

Borko, H. (1968). Information science: what is it? *American Documentation*, v. 19, n. 1, 3-5,. https://doi.org/10.1002/asi.5090190103

Brasil. Ministério da Agricultura, Pecuária e Abastecimento. (2023). *Febre aftosa.* https://www.gov.br/agricultura/pt-br/assuntos/sanidade-animal-e-vegetal/saude-animal/programas-de-saude-animal/febre-aftosa/programa-nacional-de-erradicacao-de-febre-aftosa-pnefa

Camperos-Reyes, J.T. et al. (2020). Elementos de modelado para intercambio de información en ciencia de la información e ingeniería de sistemas. *Ciência da Informação*, v. 49, n. 1. https://doi.org/10.18225/ci.inf.v49i1.4801

Camperos-Reyes, J.T. (2023). *Aderência Lexical a dados publicados para produtores rurais*. http://hdl.handle.net/11449/242749

Castro, L.N.; Ferrari, D.G. (2016). *Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações*. São Paulo: Saraiva.

CNA/SENAR. *Programa Alimentos Artesanais e Tradicionais*. (2022). https://www.cnabrasil.org.br/projetos-e-programas/alimentos-artesanais-e-tradicionais

Centro de Estudos Avançados em Economia Aplicada. (2022a). *Sobre o Cepea*. https://www.cepea.esalq.usp.br/br/sobre-o-cepea.aspx

Centro de Estudos Avançados em Economia Aplicada (2022b). *CepeaEsalqUSP*. https://www.youtube.com/channel/UCxWkJKksyxJD3ccmEZxBIqw

Confederação da Agricultura e Pecuaria do Brasil. (2022a). *Destaques*. https://www.cnabrasil.org.br/

Confederação da Agricultura e Pecuaria do Brasil. CNA/SENAR. (2022b). *Projeto Campo Futuro*. https://www.cnabrasil.org.br/projetos-e-programas/campo-futuro#:~:text=O%20Campo%20Futuro%20%C3%A9%20um,se%20destina%20aos%20produtores%20rurais

Dutra, M.L. *Mineração Textual: entrevista com o prof. Moisé Dutra*. (2021). https://youtu.be/WOLU_67MmEA

Empresa Brasileira de Pesquisa Agropecuária. (2020). *Produção dos Cafés do Brasil atinge 61,62 milhões de sacas de 60kg em 2020, volume 25% maior que 2019*. https://www.embrapa.br/busca-de-noticias/-/noticia/56084554/producao-dos-cafes-do-brasil-atinge-6162-milhoes-de-sacas-de-60kg-em-2020-volume-25-maior-que-2019

Empresa Brasileira de Pesquisa Agropecuária. (2022). *Embrapa Suínos e Aves. Especial: Sanidade Animal Peste Suína Africana*. https://www.embrapa.br/suinos-e-aves/psa

Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina. (2021). *Conceitos e métodos aplicados à gestão de empreendimentos rurais e custos de produção nos programas da Epagri 2021.* https://docweb.epagri.sc.gov.br/website_cepa/publicacoes/Conceitos_Metodos_Gestao_Custo_producao_programas.pdf

Federação da Agricultura e Pecuária do Estado do Espírito Santo. (2022). *Presentaçao*. https://faes.org.br/apresentacao_faes

Federação da Agricultura e Pecuária do Estado de Minas Gerais. (2022). Sistema FAEMG. http://www.faemg.org.br/faemg/

Federação da Agricultura e Pecuária do Pará. (2022). *A FAEPA: o qué é?* http://sistemafaepa.com.br/faepa/a-faepa/

Feldman, R.; Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press*. https://dl.icdst.org/pdfs/files/25a6d982ee80e1db7a4ebf7eeca4e0ec.pdf

Frické, M. (2013). Logic and the Organization of Information: An Introduction. *North American Symposium on Knowledge Organization (NASKO)*, v.4, n.1, 70-75. https://journals.lib.washington.edu/index.php/nasko/article/view/14646/12290

Hearst, M.A. (1999). Untangling text data mining. In: *Proceedings of the 37th Annual meeting of the Association for Computational Linguistics,* 3-10. https://aclanthology.org/P99-1001.pdf

Hearst, M.A. (2003). *What is text mining*. https://www.jaist.ac.jp/~bao/MOT-Ishikawa/FurtherReadingNo1.pdf

Huang, A. (2008). Similarity measures for text document clustering. In: *Proceedings of the sixth New Zealand Computer Science Research student conference (NZCSRSC2008)*, 6, 9-56.

Instituto Brasileiro de Geografia e Estatística. (2020). *Missão Institucional*. https://www.ibge.gov.br/acesso-informacao/institucional/o-ibge.html

Instituto de Pesquisa Econômica Aplicada. (2011). Políticas Sociais - Um plano para acabar com a miséria. *Desafios do desenvolvimento* 8, 67. https://www.ipea.gov.br/desafios/index.php?option=com_content&id=2579:catid=28

Kao, A. and Poteet, S.R. (Eds.). (2007). *Natural language processing and text mining*. Springer.

Korn, J.; Huss, F. and Cumbers, J.D. (1998). Natural language for modelling situations. In: *IEE Colloquium on Natural Language Understanding*. London: IET.

Ministério da Industria, Comércio Exterior e Serviços. (2022). *Comex Stat. 2022*. https://comexstat.mdic.gov.br/

Nadkarni, P.M., Brandt, C. and Frawley, S. (1998). Managing attribute-value clinical trials data using the ACT/DB client-server database system. *Journal of the American Medical Informatics Association*, v. 5, n. 2, 139-151. https://doi.org/10.1136/jamia.1998.0050139

Portal DBO. (2012). *Brasil, maior exportador global de carne bovina, importou 50,8 mil toneladas premium em 2020*. https://www.portaldbo.com.br/brasil-maior-exportador-global-de-carne-bovina-importou-508-mil-toneladas-premium-em-2020/

Ross, S.D.; Cruz, B. de P. A. (2021). Análise Quantitativa de Textos: Apresentação e Operacionalização da Técnica via Twitter. *Administração: Ensino e Pesquisa*, v. 22, n. 1. https://doi.org/10.13058/raep.2021.v22n1.1859

Rosenfeld, L.; Morville, P. and Arango, J. (2015). *Information Architecture: For the Web and Beyond*. Sebastopol/CA: O'Reilly.

Sant'Ana, R.C.G. (2019). Campo informacional resultante da interação de ciclos de vida dos dados. In: Dias, G.A.; Oliveira, B.M.J.F. de. *Dados científicos: perspectivas e desafios. João Pessoa*: Editora UFPB, 13-31.

Santos, P.L.V. da C.; Sant'Ana, R.C.G. (2019). Camadas de representação de dados e suas especificidades no cenário científico. In: Dias, G.A.; Oliveira, B.M.J.F. de. *Dados científicos: perspectivas e desafios*. João Pessoa: Editora UFPB, 53-66.

Serviço Nacional de Aprendizagem Rural. (2022). *SENAR*. https://www.cnabrasil.org.br/senar

Silge, J. & Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly. https://www.tidytextmining.com/index.html

Sloan, M.C. (2010). Aristotle's nicomachean ethics as the original locus for the septem circumstantiae. *Classical Philology*, v. 105, 3, 236-251. https://doi.org/10.1086/656196

UOL Economia. (2022). Peste suína nas Américas. https://economia.uol.com.br/reportagens-especiais/agronegocio-peste-suina-nas-americas/#cover

Walcheke, J. & Wolter, R.P. (2011). Critérios de construção e relato da análise pro-totípica para representações sociais. *Psicologia: Teoria e Pesquisa*, 27(4), 521-526.

APÊNDICE A

Unidades lexicais similares na categoria da Necessidade Informacional Crédito			
Acordo	cidadania	indígenas	pré
agrícolas	concedidos	investimento	produtores
agricultoras	conceito	investimentos	público
agricultores	continental	jovens	quilombolas
agroecologia	contrato	jurídica	recursos
anos	cooperação	juros	resumo
aquicultores	crédito	leite	rurais
aquicultura	dados	livres	rural
artesanais	dap	mês	saldo
assentados	direcionados	modalidade	sustentabilidade
assistência	empreendimento	mulheres	tabaco
atendidas	enquadramento	nº	taxa
atendimento	estratégica	operações	taxas
atendimentos	extensão	origem	técnica
ater	familiares	pago	território
atualizado	financiamento	pecuários	tipo
beneficiário	financiamentos	pequeno	total
beneficiários	fixadas	pesca	tradicionais
café	fundiário	pescadores	uf
categoria	gestão	pessoa	valor
chamadas	inadimplência	porte	
Unidades lexicais similares em todo conjunto de análise			
abastecimento	cumprimento	início	produtores
acompanhamento	dados	investimento	produtos
acordo	dap	investimentos	profissionais
agrária	data	item	programa
agrícolas	declaração	itr	promover
agricultoras	declarações	janeiro	pronaf
agricultores	decreto	jovens	propriedade
agricultura	dentre	jurídica	proveniente
agroecologia	departamento	jurídicas	público
ambiente	desenvolvimento	juros	quantitativo
âmbito	desigualdades	legalmente	quilombolas
amenizar	desses	leite	receita
anos	destacam	livres	receitas
aptidão	df	lote	recursos
aquicultores	direcionados	mail	referência
aquicultura	dispositivos	manutenção	referente
arranjos	distrito	mapa	regionais
arrecadação	econômico	matéria	relação
artesanais	educação	mecanismo	repassada
assentados	efetuar	meio	repetição
assistência	empreendimento	mensal	representa
associação	enquadramento	mês	resumo
atendidas	entes	ministério	rurais
atendimento	equilíbrio	modalidade	rural
atendimentos	estabelecidos	mulheres	saf
ater	estados	município	saldo
ativas	estratégica	municípios	secretaria
ativos	estrutura	nacional	selo
atualizado	etapa	nº	singular
básica	evolução	nome	sobre

beneficiário	execução	nominal	social
beneficiários	exportação	nota	sócio
bolsa	extensão	número	sustentabilidade
busca	extrativismo	operações	sustentável
cabe	familiar	origem	tabaco
café	familiares	pago	taxa
categoria	federais	parcela	taxas
central	federal	participação	técnica
chamada	fim	partir	técnicos
chamadas	financiamento	pecuária	tema
cidadania	financiamentos	pecuários	territorial
combustível	fiscal	pequeno	território
comercializam	físicas	percentual	tesouro
compensação	fixadas	pesca	tipo
concedidos	folha	pescadores	todas
conceito	fornecedoras	pessoa	total
conjunto	fundamental	planilha	tradicionais
constitucionais	fundiário	pnpb	transferência
constituição	fundo	porte	transferida
continental	gestão	povos	uf
contratada	ibge	prazos	último
contrato	imposto	pré	união
contratos	impostos	previstas	unidades
controle	inadimplência	prima	valor
cooperação	incessante	principais	valorização
cooperativa	indígenas	processo	verde
cooperativas	industrializados	produção	versão
credenciamento	informa		vigência
crédito			

Fonte: dados da pesquisa.

APÊNDICE B - Instituições gremiais identificadas

INSTITUIÇÃO	DESCRIÇÃO	UNIFORM RESOURCE LOCATOR
Confederação da Agricultura e Pecuária do Brasil (CNA)	Entidade sindical patronal que representa 5 milhões de produtores rurais comerciais brasileiros, de pequeno, médio e grande porte (Confederação da Agricultura e Pecuária do Brasil, 2022a).	https://www.cnabrasil.org.br/
Federação da Agricultura e Pecuária do Pará (FAEPA)	Federação que tem como missão representar e defender a classe produtora rural, promovendo ações para a sustentabilidade do agronegócio paraense (Federação da Agricultura e Pecuária do Pará, 2022).	http://sistemafaepa.com.br/faepa/a-faepa/
Federação da Agricultura e Pecuária do estado da Bahia (FAEB)	Instituição criada para representar, organizar e fortalecer o produtor rural baiano, velando pelos seus direitos e interesses, promovendo o desenvolvimento econômico, social e ambiental do setor agropecuário (Federação da Agricultura e Pecuária do Estado da Bahia, 2022).	http://www.sistemafaeb.org.br/faeb/
Federação da Agricultura e Pecuária do estado de Minas Gerais (FAEMG)	Representa os interesses dos produtores rurais mineiros de maneira individual ou coletiva. Realiza assessoramento em aspectos ambientais, contábeis, técnicos, marketing, sindicais, jurídicos e de informática (Federação da Agricultura e Pecuária do estado de Minas Gerais, 2022).	http://www.faemg.org.br/faemg/
Federação da Agricultura e Pecuária do estado do Espírito Santo (FAES)	Instituição criada como entidade sindical visando a defesa dos produtores do estado; brinda assessoria jurídica, sindical, econômica, contábil e ambiental, bem como estratégias de formação profissional. Ainda age como órgão consultor em decisões da esfera executiva do estado e os seus municípios (Federação da Agricultura e Pecuária do estado do Espírito Santo, 2022).	https://faes.org.br/apresentacao_faes

Fonte: dados da pesquisa.

1 NOTAS

A extensão SelectorGadget permite identificar seletores CSS em websites; foi usada durante o processo de identificação dos blocos de texto que foram coletados. Cf. https://selectorgadget.com/.

2 Sendo *n* o número que se incrementa segundo a página de notícias criada na filtragem.

3 *Cascading Style Sheets.*

4 Estimou-se conveniente usar as duas palavras no plural pois durante os testes do código foi visto que dessa forma o algoritmo recuperou maior quantidade de palavras e de maior heterogeneidade temática.

5 As palavras resultantes da aplicação da 'findAssocs' provêm do corpo das notícias e por isso elas foram mantidas na forma em que foram publicadas, portanto, algumas encontram-se no plural.

6 Corresponde à Confederação Nacional do Sistema Financeiro (CONSIF).

7 Verificou-se nas notícias coletadas que a palavra oit provém de conteúdo onde se observa a expressão Organização Internacional do Trabalho (OIT).

8 Corresponde ao Ministério de Planejamento, Orçamento e Gestão (MPOG).

9 Brasil sem miséria (BSM).

10 Sistema informatizado de ATER. Cf. https://sistemas.agricultura.gov.br/siater/sys/siater/login

11 Sistema de monitoramento e Gestão da Secretaria Especial de Agricultura Familiar e do Desenvolvimento Agrário (SEAD). Cf. http://nead.mda.gov.br/login?si=simog

12 Unidades produtivas familiares UPFS.

13 A sigla BSM corresponde ao plano Brasil Sem Miséria. Ele tem como "objetivo erradicar a pobreza extrema no país, por meio de ações de transferência de renda, acesso a serviços públicos e inclusão produtiva" (Instituto de Pesquisa Econômica Aplicada, 2011).