Ciência de Dados como Ferramenta de Apoio à Tomada de Decisão: Classificação Hierárquica Descendente de Pedidos de Acesso à Informação na Prefeitura de São Paulo[1]

La Ciencia de Datos como Herramienta de Apoyo a la Toma de Decisiones: Clasificación Jerárquica Descendente de Solicitudes de Acceso a la Información en la Alcaldía de São Paulo[2]

Data Science as a Tool to Support Decision-Making: Descending Hierarchical Classification of Access to Information Requests in the Municipality of São Paulo

Claudio Henrique Fontenelle Santos e Ana Lúcia da Silva Romão

https://doi.org/10.36428/revistadacgu.v14i26.544

Resumo: Buscou-se compreender de que forma a ciência de dados e as tecnologias de mineração e classificação de textos podem contribuir para a tomada de decisões a partir de uma melhor compreensão agregada dos pedidos de acesso à informação. A pesquisa utilizou dados dos pedidos de acesso à informação feitos à Prefeitura Municipal de São Paulo (PMSP), de 2012 a 2019, disponíveis no Portal de Dados Abertos da municipalidade, propondo a identificação e classificação das principais questões apresentadas. Os 39.369 textos dos pedidos de acesso submetidos à PMSP foram reunidos em um corpus e submetidos a análise por meio de Classificação Hierárquica Descendente (CHD). Ao propor uma classificação de textos como uma metodologia para análise de dados textuais, reforçou-se um paradigma de que dados textuais não pertencem apenas ao campo qualitativo. Além disso, a consideração de apenas substantivos, excluídos verbos e advérbios; e os adjetivos mais ocorrentes serem usados como parte de expressões, permitiu uma otimização do contexto dos pedidos, proporcionando classificar os dados textuais de maneira mais objetiva, mitigando o viés dos investigadores. O artigo apresenta também outros estudos de caso relevantes para a pesquisa, com referências encontradas na análise de pedidos de acesso à informação, contribuindo para a compreensão de pedidos dos cidadãos de modo aglutinado e permitindo aos tomadores de decisões um melhor entendimento das demandas da sociedade, podendo resultar em políticas públicas mais focadas. Conclui-se que a análise dos dados através da CHD permite obter informações relevantes para a tomada de decisão baseada em dados e evidências e que a abordagem favorece a concretização de decisões fundamentadas e mais próximas das necessidades dos cidadãos.

Palavras-chave: tomada de decisão, pedidos de acesso à informação, mineração de texto, classificação hierárquica descendente.

Abstract: This article sought to understand how data science, text mining and text classification technologies can contribute to improved decision-making based on an overall understanding of access to information requests. Researchers used data from access to information requests made to the Municipality of São Paulo (PMSP), from

1. Artigo submetido em 18/07/2022 e aceito em 01/12/2022.

2. Artigo recebido em 18/06/2022 e aprovado em 01/12/2022.

2012 to 2019, available on the municipality's Open Data Portal, proposing presented the main issues identification and classification. 39,369 requests submitted to PMSP were gathered into a corpus and submitted for analysis through a Descending Hierarchical Classification (CHD). In proposing a classification of texts as a methodology for analyzing textual data, a paradigm has been reinforced: textual data does not belong only to the qualitative domain. In addition, considering only nouns, excluding verbs and adverbs, and the most frequent adjectives being used as part of vocabulary expressions, allowed a request context optimization, allowing to classify the textual data in a more objective way, mitigating researchers bias. We present main references found in access to information requests analysis, cases in Mexico, national studies, and in China, in the city of Beijing, and contributes to the understanding citizens requests in an aggregated way, allowing decision makers to a better understanding of society's demands, which may result in more focused public policies. It is concluded that data analysis through the CHD allows obtaining relevant information for decision-making based on data and evidence and that the approach favors the implementation of reasoned decisions that are closer to the citizens' needs.

Keywords: decision making, access to information requests, text mining, descending hierarchical classification.

Resumen: Este artículo buscó comprender cómo las tecnologías de ciencia de datos, minería de texto y clasificación de texto pueden contribuir a una mejor toma de decisiones a partir de la comprensión agregada de las solicitudes de acceso a la información. La investigación utilizó datos de solicitudes de acceso a la información realizadas al Municipio de São Paulo (PMSP), de 2012 a 2019, disponibles en el Portal de Datos Abiertos del municipio, proponiendo la identificación y clasificación de los principales problemas presentados. 39.369 textos de solicitudes enviados a PMSP fueron reunidos en un corpus y sometidos a análisis a través de una Clasificación Jerárquica Descendente (CHD). Al proponer una clasificación de textos como metodología para el análisis de datos textuales, se reforzó un paradigma de que los datos textuales no pertenecen únicamente al campo cualitativo. Además, considerar solo los sustantivos, excluyendo verbos y adverbios, y los adjetivos más frecuentes utilizados como parte de las expresiones del vocabulario, permitió una optimización del contexto de la solicitud, lo que permitió clasificar los datos textuales de forma más objetiva, mitigando el sesgo de los investigadores. Presentamos los principales referentes encontrados en el análisis de solicitudes de acceso a la información, casos en México, estudios nacionales y en China, en la ciudad de Beijing, y contribuye a la comprensión de las solicitudes ciudadanas de manera agregada, permitiendo a los tomadores de decisiones una mejor comprensión de las necesidades de la sociedad. demandas, lo que puede redundar en políticas públicas más focalizadas. Se concluye que el análisis de datos a través del CHD permite obtener información relevante para la toma de decisiones basadas en datos y evidencias y que el enfoque favorece la implementación de decisiones razonadas y más cercanas a las necesidades de los ciudadanos.

Palabras clave: toma de decisiones, solicitudes de acceso a la información, minería de textos, clasificación jerárquica descendente.

INTRODUÇÃO

A Lei de Acesso à Informação (LAI) completou, em 2022, 10 anos de existência no Brasil, e com ela surgiram, em todas as esferas de governo do Brasil, federal, estadual e municipal, portais de recebimento de pedidos de acesso à informação. A Escala Brasil Transparente 360 graus - 2ª edição identificou que todos os estados brasileiros e o Distrito Federal já haviam regulamentado a LAI até 2019 (Controladoria-Geral da União, 2019). Ao nível municipal, a Prefeitura de São Paulo, maior cidade do Brasil, publicou em agosto de 2012 o Decreto nº 53.623 regulamentando a LAI no âmbito do município.

Os governos, através do acesso à informação, colhem subsídios para aprimorar suas políticas públicas, pois cada pedido representa uma solicitação, pessoal ou coletiva, e a compreensão desses pedidos, de modo individual e agregado, é muito importante para a tomada de decisão pública. A Ciência de Dados facilita essa compreensão, através da extração de valor dos

dados com a utilização de técnicas de tratamento de dados, como mineração de textos e inteligência artificial (Kotu e Deshpande, 2018; Soares, 2020).

O objetivo deste artigo passa por compreender de que forma as tecnologias de mineração e classificação de textos contribuem para a tomada de decisão baseada em dados e evidências, através da análise agregada de pedidos de acesso à informação. A pesquisa utilizou os dados dos pedidos feitos à Prefeitura Municipal de São Paulo (PMSP), de 2012 até 2019, propondo a identificação e classificação dos principais temas apresentadas.

Existem diversas técnicas de análise, entre elas, a Classificação Hierárquica Descendente (CHD), que, em linguagem simplificada, analisa um grande volume de textos, e, usando um algoritmo previamente criado e validado, apresenta de forma agregada, os principais assuntos abordados nos textos investigados.

Este artigo traz contributo ao utilizar a técnica de CHD na análise e classificação de pedidos de acesso à informação. Inova-se também ao considerar na análise apenas substantivos, palavras compostas, como "conselho municipal", "rede municipal" e siglas, como "SME" e "SPTRANS", e ao desconsiderar verbos, adjetivos e advérbios. Martin e Johnson (2015) concluíram que usar apenas substantivos pode trazer melhores resultados.

Esse artigo apresenta, primeiramente, na seção dois, um referencial teórico sobre ciência de dados, textos como dados e acesso à informação, seguido por uma revisão da literatura dos artigos que já realizaram investigações semelhantes. Na seção quatro é apresentada a metodologia e na seção cinco são mostradas e discutidas as análises realizadas, incluindo breves descrições estatísticas dos dados, segmentação por palavras e refinamento por classificação de tópicos. A última seção apresenta a conclusão do estudo, limitações e sugestões de estudos futuros.

REFERENCIAL TEÓRICO

Ciência de Dados

Estamos vivendo uma era de evolução do "mundo de dados", continuamente conectados à nossa vida diária, ao nosso trabalho e à economia. Governos, instituições privadas, e acadêmicos estão diariamente empreendendo esforços para chegar a maneiras de converter dados em instrumentos de tomada de decisões, promover a pesquisa e o desenvolvimento da ciência (Cao, 2017). Os cientistas agora possuem, com mais facilidade, acesso a esse grande universo de dados, incluindo dados governamentais. Como resultado, pesquisas empíricas podem ser conduzidas em uma escala que não seria possível a uma ou duas gerações anteriores (Lane et al., 2022).

Para Kotu e Deshpande (2018), a ciência de dados envolve inferência e iteração de muitas hipóteses diferentes relacionadas aos dados disponíveis, sendo um dos seus aspectos-chave o processo de generalização de padrões a partir desses dados, generalização essa que deve ser válida, não só para a base de dados usada para observar o padrão, mas para novos dados ainda não acessados.

Em resumo, a Ciência de Dados busca aprimorar as tomadas de decisão baseando-as em "*insights*" extraídos das informações obtidas nos conjuntos de dados (Kelleher e Tierney, 2018).

Texto como Dados

O advento das novas tecnologias tem permitido a disponibilização de vasta quantidade de textos em formato digital por órgãos governamentais e formuladores de políticas públicas (Hollibaugh, 2019). Para os cientistas sociais, a informação codificada em texto é um rico complemento a suas pesquisas (Gentzkow et al., 2019).

As investigações usando textos como dados, auxiliadas por computador, estão se tornando relevantes em estudos nos campos da administração pública, políticas públicas e ciências políticas (Hollibaugh, 2019). Contudo, até 2022 ainda são escassos os estudos usando abordagens de análise de textos e modelagem de tópicos para extrair conhecimento de pedidos de acesso à informação. Em pesquisa realizada em junho de 2022 nas bases Scopus e Web of Science foram encontrados apenas quatro artigos que utilizaram metodologias do campo de ciências de dados para investigar pedidos de acesso à informação, conforme mostra o Quadro 1, a seguir:

QUADRO 1 – ESTADO DA PESQUISA

AUTORIA	PAÍS	ESFERA DE INVESTIGAÇÃO	PERGUNTAS DE PESQUISA	OBJETIVO	CONCEPÇÃO TEÓRICA	BASE DE DADOS	METODOLOGIA DE ANÁLISE DE DADOS	VARIÁVEIS/ CATEGORIAS ANALÍTICAS	PRINCIPAIS CONCLUSÕES E CONTRIBUIÇÕES
Berliner, D., Bagozzi, B. E., & Palmer-Rubin, B. (2018)	México	Nacional	Que informação os cidadãos procuram? Como essas informações se relacionam com questões de importância pública?	Identificar os temas relacionados aos pedidos de acesso à informação feitos e como eles se relacionam com as questões públicas.	Modelo da *accountability* pública (principal agente), modelo iceberg (interesse privado) e transparência.	Pedidos feitos ao governo federal mexicano entre 2003-2015.	Análise não supervisionada de texto (LDA) e modelação de tópicos.	(dependente) Potencial de *accountability* do pedido.	Vinte tópicos (ver Quadro 2).
Bagozzi, B. E., Berliner, D., & Almquist, Z. W. (2019)	México	Nacional	Quais são os temas de pedidos de acesso à informação feitos pelos cidadãos mexicanos que mais se associam com negação de respostas?	Identificar os temas dos pedidos mais associados a negações de respostas.	Motivação dos burocratas para responder ou negar respostas a pedidos de acesso à informação e transparência.	Pedidos feitos ao governo federal mexicano entre 2003-2015.	Análise supervisionada de texto (sLDA) e modelação de tópicos.	(dependente) Não responsividade.	Os cinco tópicos com maior probabilidade de receber uma negação apresentaram natureza investigativa, relacionados com atividades policiais, financeiras, políticas, compras governamentais.
Berliner, D., Bagozzi, B. E., Palmer-Rubin, B., & Erlich, A. (2021)	México	Nacional	Como o governo responde quando os cidadãos fazem questionamentos via pedidos de acesso à informação?	Compreender as motivações para responder ou não responder a pedidos de acesso à informação.	*Accountability* vertical (principal agente) e teoria da responsividade do governo.	Pedidos feitos ao governo federal mexicano entre 2003-2015.	Análise não supervisionada de texto (LDA) e modelação de tópicos.	(dependente) Não responsividade e (independente) alinhamento político ao nível municipal.	Pedidos feitos de locais onde o governo é bem votado recebem mais respostas.
Wang, Z., & Zhong, Y. (2020)	China	Local (Pequim)	Quais são os principais temas que os residentes de Pequim questionam ou sugerem ao governo?	Compreender os principais temas relacionados aos pedidos de acesso à informação feitos.	Governança	Solicitações feitas à Prefeitura de Pequim no ano de 2015.	Mineração de textos, análise de classes e modelação probabilística de tópicos.	(segmentação) Tópico do assunto.	Sete tópicos (ver Quadro 2).

Fonte: elaboração própria, com base nos estudos da pesquisa.

Berliner et al. (2018) investigaram, no México, no âmbito federal, que informações os cidadãos procuram, e como essas informações se relacionam com questões de importância pública. Os autores fizeram uso de um modelo generativo probabilístico de tópicos (*Latent Dirichlet Allocation* – LDA – ver Blei et al., 2003) para descobrir tópicos "latentes" no conjunto agregado de pedidos de acesso à informação recebidos entre 2003 e 2015. Concluíram que alguns tópicos e os pedidos que os compõem, como os relacionados a temas como "Meio Ambiente" ou "Gastos Públicos", contribuem mais para a "*accountability* pública" do que outros, como os que contêm "Demandas Individuais" ou "Informações Comerciais".

Bagozzi et al. (2019), dessa vez usando a técnica de (*Latent Dirichlet Allocation* supervisionada – sLDA – ver Mcauliffe e Blei, 2007) buscaram identificar os tópicos, e seus pedidos relacionados, mais associados a negações de respostas. Os tópicos com maior probabilidade de receberem uma "negação de resposta" foram de natureza investigativa, ligados a atividades policiais, financeiras, políticas e compras governamentais.

Berliner et al. (2021) investigaram como o governo responde às solicitações dos cidadãos. Segundo os autores, usando a responsividade como variável dependente, isto é, a motivação para responder os pedidos, recebem mais respostas favoráveis aqueles oriundos de localidades onde o partido do governo foi mais votado.

Um estudo realizado em Pequim por Wang e Zhong (2020), fazendo uso das técnicas de "mineração de textos", "análises de *clusters*" e "modelação de tópicos", investigou quais eram as principais questões apresentadas pelos cidadãos. Os autores concluíam serem mais frequentes os relacionados a registro de imóveis, construções ilegais, educação e moradia (Cf. Quadro 1).

Acesso à Informação

O direito ao acesso à informação surgiu em 1766, com a lei de liberdade de imprensa da Suécia, ao adotar que dados governamentais são, em princípio, abertos ao público (Rydholm, 2013). Em 1789, a Declaração dos Direitos do Homem e do Cidadão, aprovada na França, preconizou o direito dos cidadãos a acessar o orçamento (art. 14.º) e a pedir contas à Administração Pública (art. 15.º). Direito similar foi outorgado na Holanda, no ano de 1795, ao ser declarado que todos têm o direito de exigir, de cada funcionário público, prestação de contas e justificativas de suas condutas (Banisar, 2006).

Após a II Guerra Mundial, com a criação da Organização das Nações Unidas (ONU), padrões internacionais de direitos humanos, entre eles o direito ao acesso à informação, começaram a serem propagados, inicialmente pela Declaração Universal dos Direitos Humanos (1948), que, em seu artigo 19º, diz que todos têm o direito de procurar, solicitar e receber informações (Banisar, 2006).

Para Walby e Larsen (2012), o acesso à informação e a liberdade de informação são relevantes mecanismos democráticos pelos quais a sociedade, organizações civis e jornalistas podem solicitar informações não publicadas pelos governos, em busca de mais detalhes sobre a atuação governamental. As leis de acesso à informação regulam a maneira como os órgãos do governo devem disponibilizar as informações ao público e o modo como devem lidar com os pedidos de informação (Angélico e Teixeira, 2012).

Savage e Hyde (2014) indicaram que leis de acesso à informação são uma poderosa ferramenta para os cientistas sociais, embora, àquela época ainda não haviam sido aproveitado todo esse potencial. Walby e Luscombe (2017) ratificaram, afirmando que pedidos de acesso à informação eram pouco utilizados como meios de produção de dados.

Os pedidos feitos pelos cidadãos aos governos, no âmbito das suas respetivas leis de acesso à informação possuem grande potencial, tanto ao nível prático, quanto ao nível teórico. Na prática, eles permitem acesso a dados, que não se encaixam facilmente como primários ou secundários, nem qualitativos ou quantitativos, sem compromissos epistêmicos ou antológicos, porém, susceptíveis de diversas análises com o uso das mais variadas lentes teóricas, contribuindo assim para o avanço teórico (Savage e Hyde, 2014). Walby e Luscombe (2017) argumentaram que, quando sistematicamente desenhadas e conduzidas, pesquisas com pedidos de acesso à informação terão credibilidade, ética, coerência e contribuição significativas.

Serviços de acesso à informação pública apresentam um número de oportunidades de pesquisa no desenvolvimento de modelos computacionais que puderam ajudar a compreender os cidadãos e suas necessidades (Flores et al., 2022). No entanto, a maioria dos pesquisadores parece não ter familiaridade com esses dados ou não lhe dá a devida importância, escrevendo

sobre eles com uma abordagem "jornalística", em vez de produzir de forma sistemática dados qualitativos longitudinais sobre as práticas governamentais (Walby e Larsen, 2012).

O direito ao acesso à informação no Brasil foi outorgado com a promulgação da Lei n. 12.527 (2011), conhecida também como LAI, sendo uma resposta do legislador aos comandos constantes no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição da República Federativa do Brasil (1988/2022). Considerada um marco no processo de abertura das informações da administração pública brasileira, a LAI se tornou um dos principais instrumentos de promoção da transparência no âmbito das três esferas públicas, bem como no Distrito Federal (Angélico e Teixeira, 2012).

METODOLOGIA

Desenho do Estudo

Esta investigação, de natureza exploratória e descritiva, utiliza uma abordagem mista, qualitativa e quantitativa, com recurso à técnica de CHD, para categorizar os pedidos de informação feitos à Prefeitura Municipal de São Paulo. Este método permite a caracterização, de modo individual e agregado, do conteúdo de textos, reunindo-os em conjuntos de classes, apresentando os vocábulos mais associados a cada uma das classes, bem como às variáveis demográficas disponíveis.

A CHD, segundo Mendes et al. (2019), é uma técnica que categoriza as palavras em classes lexicais considerando a frequência e a posição relativa com que elas aparecem nos textos. Vários testes de qui quadrado são realizados, até que são esgotadas as possibilidades de diferenciação entre as classes. O número inicial de classes é uma variável subjetiva a ser inserida pelo investigador, que pode fazer vários testes, e, a partir dos resultados, encontra a mais adequada para apresentar na análise. A força de associação entre as palavras e sua respectiva classe é representada pelo índice "qui-quadrado de Pearson", o qual, quanto maior, mais provável é a hipótese de pertencimento da palavra na classe. A classificação final é representada por uma figura denominada "dendrograma", contendo as classes e palavras mais representativas.

A amostra constou de 39.369 pedidos de acesso à informação realizados à Prefeitura de São Paulo (PMSP), a partir da instituição do acesso à informação no município, em agosto de 2012, até o final do ano de 2019. Os pedidos recebidos nos anos de 2020 a 2022 não estavam disponíveis à época da coleta de dados. A cidade de São Paulo foi escolhida por dois motivos: primeiro, por ser a maior cidade do Brasil, e o segundo, por ter a base de dados do dos pedidos disponibilizada em dados abertos.

Além do conteúdo dos pedidos, foram também utilizadas a seguintes variáveis, todas oriundas dos conjuntos de dados abertos da PMSP: órgão ao qual o pedido foi endereçado, ano do pedido, tipo do solicitante (pessoa física ou jurídica), sexo (masculino ou feminino), e distrito de residência (pessoa física) ou localização (pessoa jurídica).

Procedimentos de Análise de Dados

Os dados foram analisados em duas etapas. Primeiramente, usando o programa SPSS, versão 27, foram analisados os dados demográficos, para identificação dos órgãos mais demandados e tipos de solicitantes, sexo e distritos mais frequentes. Esses dados serão apresentados mais adiante.

Na segunda etapa, os 39.382 pedidos foram reunidos em um arquivo texto e transformado em um "*corpus*", termo utilizado na academia para identificar um conjunto de textos destinados à análise textual. O programa Iramuteq foi o escolhido para a análise dessa investigação. O programa usa os ambientes estatísticos das linguagens de programação R e Python, e oferece um conjunto de análises estatísticas de dados qualitativos, entre as quais a CHD.

O corpus textual

O arquivo contendo os pedidos e as respectivas variáveis demográficas dos solicitantes compreendeu o corpus textual que foi utilizado na análise[3]. Cada pedido foi precedido, em uma linha única, com as respectivas variáveis demográficas indicadas com "*", como no exemplo a seguir. O exemplo inicia-se com quatro asteriscos, exigência do programa Iramuteq para indicar o início de um segmento de texto, no caso, um pedido de acesso à informação, seguido das variáveis. Foram elas: "o" – órgão, "a" – ano, "p" – pessoal solicitante, "s" – sexo, e "d" – distrito.

3. O arquivo está disponível em: https://github.com/chfsantos/RevistaCGU_Ciencia_de_Dados_PMSP/blob/main/corpus_couleur.docx.zip

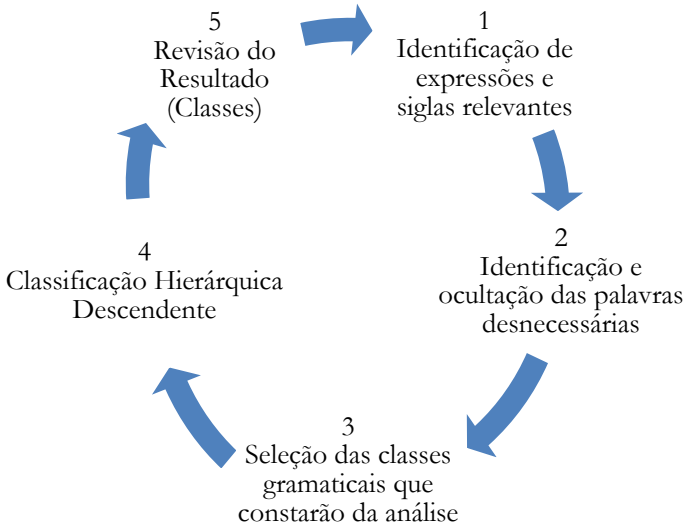
**** *o_19 *a_2013 *p_PF *s_M *d_cidadedutra

"gostaria de saber se existe alguma intervenção de canalização de córrego e ou remoções de moradias na rua samuel scott programada cep 04857060 grato"

Preparação do corpus textual

A preparação do *corpus* foi operacionalizada pelos pesquisadores em um ciclo de cinco etapas, conforme a Figura 1, a seguir:

FIGURA 1 – CICLO DE PREPARAÇÃO DO CORPUS



Fonte: elaboração própria.

Identificação de expressões e siglas relevantes

Na primeira etapa, o *corpus* foi analisado com apoio do programa RStudio e a função "textstat.collocations" do pacote R "Quanteda". Nessa etapa foi possível identificar as "expressões compostas" mais frequentes, como "São Paulo" (12.048 ocorrências), "prefeitura regional" (2.433 ocorrências), e "servidores comissionados" (1.569 ocorrências). As expressões consideradas relevantes pelos pesquisadores foram inseridas no arquivo "expressões_pt.txt", posteriormente usado na análise. A lista de expressões selecionadas está em arquivo suplementar a esse artigo[4]. Ainda nessa etapa, foi executada uma análise CHD no corpus, para identificar as palavras ou expressões mais ocorrentes. Dessa lista foram identificadas as siglas mais mencionadas, para serem inseridas no arquivo lexique_pt.txt, dicionário usado para auxílio na análise CHD.

Identificação e ocultação das palavras consideradas desnecessárias para a análise

Usando o resultado da análise CHD operacionalizada na primeira etapa, foram identificadas palavras e expressões muito usadas nos pedidos, porém consideradas desnecessárias para a análise. Por exemplo, palavras como. "informação" (5.765 ocorrências), "boa tarde" (3.782 ocorrências), e "não" (17.714 ocorrências), caso fossem consideradas na análise, não trariam significância suficiente para indicar o tema do pedido, afinal, todos eram pedidos de "informação". Usando o

4. O arquivo está disponível em: https://github.com/chfsantos/RevistaCGU_Ciencia_de_Dados_PMSP/blob/main/expressoes_adicionadas.txt

editor de texto *Visual Studio Code*, com funções avançadas de busca e substituição, uma lista de palavras foi "ocultada" nos pedidos. Essa ocultação é feita sendo inserido o caractere "_" (*underline*) imediatamente à frente e após a palavra (exemplo: "_informação_"). A decisão de considerar uma palavra desnecessária é subjetiva aos pesquisadores, e depende do contexto da investigação. Na análise textual, alguns autores defendem a exclusão de palavras específicas em cada análise, como Schofield et al. (2017) e Sarica e Luo (2021). A lista de palavras e expressões ocultadas está disponível em arquivo suplementar a esse artigo[5].

Seleção das classes gramaticais que constaram da análise

Com base em Martin e Johnson (2015) optou-se pela inclusão somente das palavras classificadas gramaticalmente como "substantivos", excluídos os verbos, adjetivos, e advérbios, comumente considerados também das análises CHD. Segundo os autores, em comparação entre três análises: original, com lematização (ver Balakrishnan e Lloyd-Yemoh, 2014) e somente com substantivos, essa última se mostrou mais coerente pelos resultados apresentados.

A seleção das classes gramaticais foi feita no início da execução do programa Iramuteq, em um menu onde são selecionadas como "ativas", "suplementares" ou "excluídas". Conforme dito, somente substantivos, expressões e siglas foram marcados como ativas.

Realização da CHD

Passadas as etapas iniciais, operacionalizou-se várias CHD, com uso de lematização, também por opção dos autores. Na lematização, a partir de um arquivo "txt" contendo a lista de palavras, palavras com o mesmo radical gramatical, e significado associado, por exemplo, "valor" e "valores", são consideradas uma só palavra, no caso, "valor", com o objetivo de reduzir a variabilidade do vocabulário, permitindo uma maior homogeneidade a ser submetida ao tratamento lexicométrico (Sousa, 2021). Essas CHD operacionalizadas serviram para que fossem identificadas outras palavras desnecessárias, que também foram excluídas (ver nota de rodapé 6).

ANÁLISE E DISCUSSÃO DOS RESULTADOS

Estatística Descritiva

Os 39.364 pedidos de acesso à informação submetidos à Prefeitura Municipal de São Paulo entre os anos de 2012 e 2019 foram enviados, predominantemente, pela internet (96,9%) e tiveram o e-mail como canal prioritário para recebimento da resposta (97,4%).

Com relação à quantidade de pedidos, percebeu-se que o número vem crescendo à medida que a população toma conhecimento do direito que possui (Tabela 1). Proporcionalmente à população, o número ainda é inexpressivo, se considerarmos que a cidade possui 11.253.503 habitantes, com dados do último censo (Instituto Brasileiro de Geografia e Estatística, 2010).

TABELA 1 – PEDIDOS DE ACESSO À INFORMAÇÃO – SÃO PAULO/SP, 2012-2019

ANO	PEDIDOS	PROPORÇÃO EM RELAÇÃO AO ANO ANTERIOR
2012	285	-
2013	2.521	784,56% (*)
2014	2.404	-4,64%
2015	4.217	75,42%
2016	5.212	23,59%
2017	7.860	50,81%
2018	8.093	2,96%
2019	8.772	8,39%
Total	39.364	

Fonte: elaboração própria. (*) O sistema começou a funcionar em 24/08/2012.

5. O arquivo está disponível em: https://github.com/chfsantos/RevistaCGU_Ciencia_de_Dados_PMSP/blob/main/termos_excluidos_do_corpus.csv

Quanto aos solicitantes, há um predomínio de pessoas físicas (36.723) sobre pessoas jurídicas (2.641), o que é um resultado comum quando se trata de pedidos de acesso à informação (Bagozzi et al., 2019; Wang e Zhong, 2020). Quanto ao sexo, o percentual de pedidos de mulheres variou entre 32,7% (2012) e 45,1% (2015), enquanto de homens oscilou entre 54,9% (2015) e 67,3% (2012). Com relação aos distritos, de residência (pessoa física) ou localização (pessoa jurídica) de quem fez o pedido, são os seguintes os dez de onde mais se originaram pedidos de acesso à informação (Tabela 2)[6]:

TABELA 2 – DEZ DISTRITOS DE ONDE MAIS SE ORIGINARAM OS PEDIDOS DE ACESSO À INFORMAÇÃO – SÃO PAULO/SP, 2012-2019

DISTRITO	PEDIDOS	%	% ACUMULADO
Vila Mariana	616	4,6	4,6
Pinheiros	570	4,2	8,8
Bela Vista	546	4,0	12,8
Barra Funda	489	3,6	16,4
Jardim Paulista	426	3,2	19,6
Perdizes	419	3,1	22,7
Jabaquara	412	3,0	25,7
Itaim Bibi	347	2,6	28,3
Artur Alvim	346	2,6	30,9
Santa Cecília	328	2,4	33,3

Fonte: elaboração própria.

Com relação à profissão, destacou-se a quantidade de pedidos feitos por jornalistas (1.134, 2,9%), considerados por alguns como "intermediários de transparência" (Lane et al., 2022; Porumbescu et al., 2022). As pessoas que se declararam pesquisadores (349) e professores (267) representaram menos de 2% do total.

Quanto aos órgãos recebedores dos pedidos de acesso à informação, conforme os dados, dez deles receberam cerca de 50% das solicitações (Tabela 3).

TABELA 3 – PEDIDOS DE ACESSO À INFORMAÇÃO DOS DEZ ÓRGÃOS MAIS DEMANDADOS – SÃO PAULO/SP, 2012-2019

ÓRGÃO	PEDIDOS	% INDIVIDUAL	% ACUMULADO
SMS – Secretaria Municipal da Saúde	3.160	8,0	8,0
SME – Secretaria Municipal de Educação	3.059	7,8	15,8
CET – Companhia de Engenharia de Tráfego	2.873	7,3	23,1
SF – Secretaria Municipal da Fazenda	2.363	6,0	29,1
SPTrans – São Paulo Transportes S/A	2.241	5,7	34,8
SMT – Secretaria Municipal de Mobilidade e Transportes	2.038	5,2	39,9
SMADS – Secretaria Municipal de Assistência e Desenv. Social	988	2,5	42,5
SMG – Secretaria Municipal de Gestão	911	2,3	44,8
SVMA – Secretaria Municipal do Verde e do Meio Ambiente	840	2,1	46,9
SMC – Secretaria Municipal de Cultura	825	2,1	49,0

Fonte: dados da pesquisa.

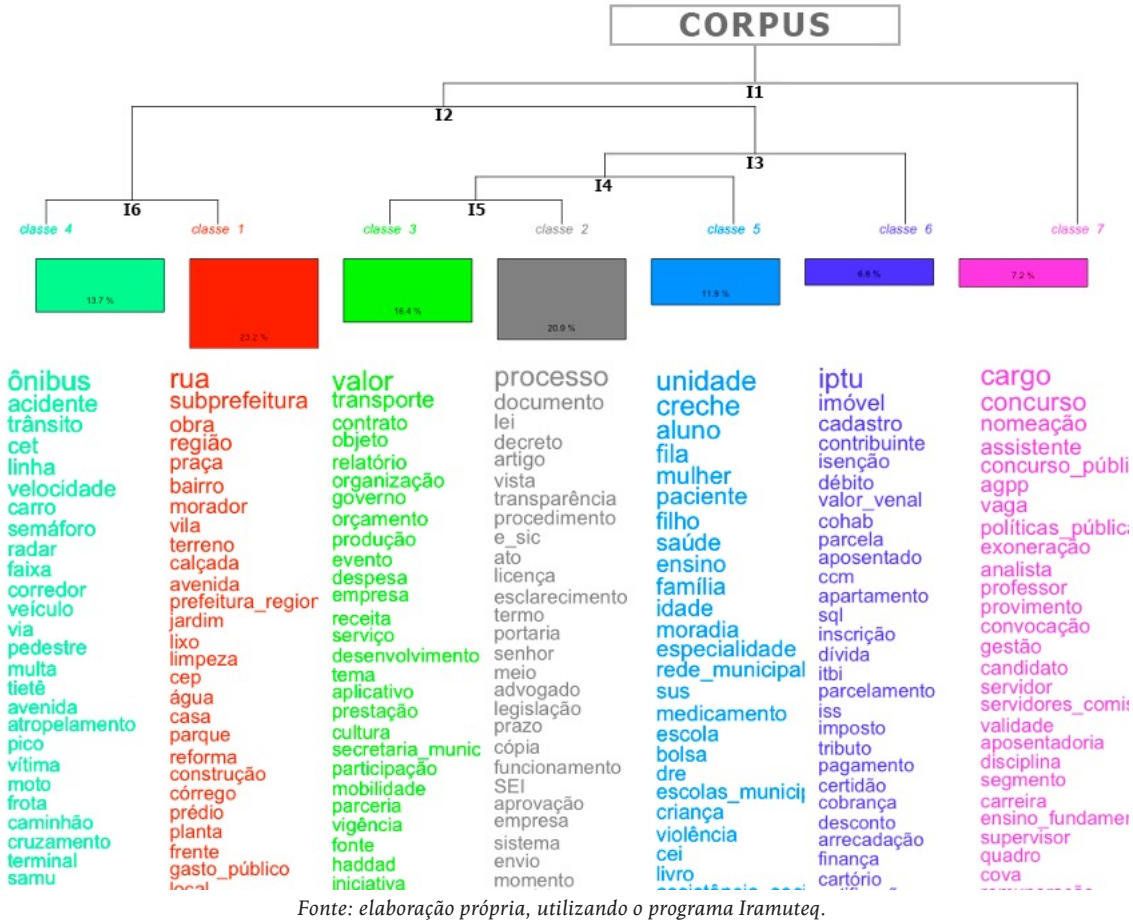
6. Foram consideradas apenas 13.499 respostas (34,3%) pois 25.883 (65,7%) solicitantes não informaram o distrito de residência, por ser um dado não obrigatório.

Cabe ser mencionado que é comum, quando há mudança de governo, também acontecerem mudanças nos nomes dos órgãos, o que dificulta o acompanhamento longitudinal dos dados. Por exemplo, a Secretaria Municipal de Habitação (SEHAB) possui pedidos a partir de 2018, sem dados anteriores, enquanto a Secretaria Municipal de Serviços (SES) possui pedidos até 2016, sem registros a partir de 2017.

Resultados da CHD

Após a execução na CHD resultaram 31.946 pedidos (81,1% da amostra) formados por 63.071 formas ativas (palavras, siglas ou expressões), com 3.102.034 ocorrências (vezes em que foram empregadas) e sete classes. 7.436 pedidos (18,9% da amostra) não apresentaram significância semântica para serem incluídos na CHD. O dendrograma apresentado na Figura 2 representa as iterações, as etapas da análise que resultaram em classe, e as principais palavras de cada uma delas, com base nos testes qui-quadrado ($\chi 2$) realizados na CHD. Observa-se que na primeira iteração (I1), a classe 7 foi separada das demais, significando que essa classe possui formas ativas mais exclusivas para si, e menos utilizadas nas demais. Em seguida, aconteceu a iteração (I2) onde as classes 6, 5, 2 e 3 foram separadas das classes 1 e 4, significando que essas duas possuem formas ativas mais exclusivas, e, como na primeira iteração, menos frequentes das demais. Do mesmo modo aconteceram as duas iterações seguintes, (I3) e (I4) separando a classe 6 das classes 5, 2 e 3, e separando a classe 5 das classes 2 e 3. As últimas iterações ocorrem quando restam pares de classes, como ocorreu com as classes 2 e 3 (I5), 1 e 4 (I6).

FIGURA 2 – RESULTADO DA CHD – CLASSE E PALAVRAS



Fonte: elaboração própria, utilizando o programa Iramuteq.

As sete classes temáticas resultantes, cujos nomes foram dados pelos pesquisadores, assemelham-se às que emergiram das pesquisas de Wang e Zhong (2020) e de Berliner et al. (2018) (Quadro 2).

QUADRO 2 – VALIDAÇÃO EXTERNA E COMPARAÇÃO COM OUTROS ESTUDOS

ESTA PESQUISA	WANG E ZHONG (2020)	BERLINER ET AL. (2018)
São Paulo (Brasil)	Pequim (China)	México (País)
CLASSES TEMÁTICAS		
1) Bairros e Distritos;	1) Household Register;	1) Taxes and Finance;
2) Trâmite e documentos processuais;	2) Illegal Construction;	2) Environment and Land;
3) Contratações públicas;	3) Education;	3) Employees1: Salaries/Benefits;
4) Mobilidade Urbana;	4) Demolition;	4) Employees2: Functions/Qualif.;
5) Família: saúde, educação e assistência social;	5) City Management;	5) Employees3: Personnel;
6) Imóveis;	6) Housing;	6) Individual Needs;
7) Cargos e Concursos Públicos.	7) Traffic.	7) Commercial Information;
		8) Distributive Programs;
		9) Medical1: Contracts/Suppliers;
		10) Medical2: Purchases/Spending;
		11) Medical3: Inventories;
		12) Energy and Utilities;
		13) Health Statistics;
		14) Rules and Procedures;
		15) Education;
		16) Military, Police, and Crime;
		17) Budgets and Spending;
		18) Procurement1: Service Providers;
		19) Procurement2: Procedures/Docs;
		20) Procurement3: Anti-Corruption.

Fonte: elaboração própria.

Nos pontos a seguir as classes serão detalhadas, a partir dos significados agregados das palavras prevalentes em cada uma delas. As classes serão descritas, operacionalizadas e exemplificadas de forma sequencial, iniciadas pela classe 7, primeira à direita, para respeitar as iterações realizadas.

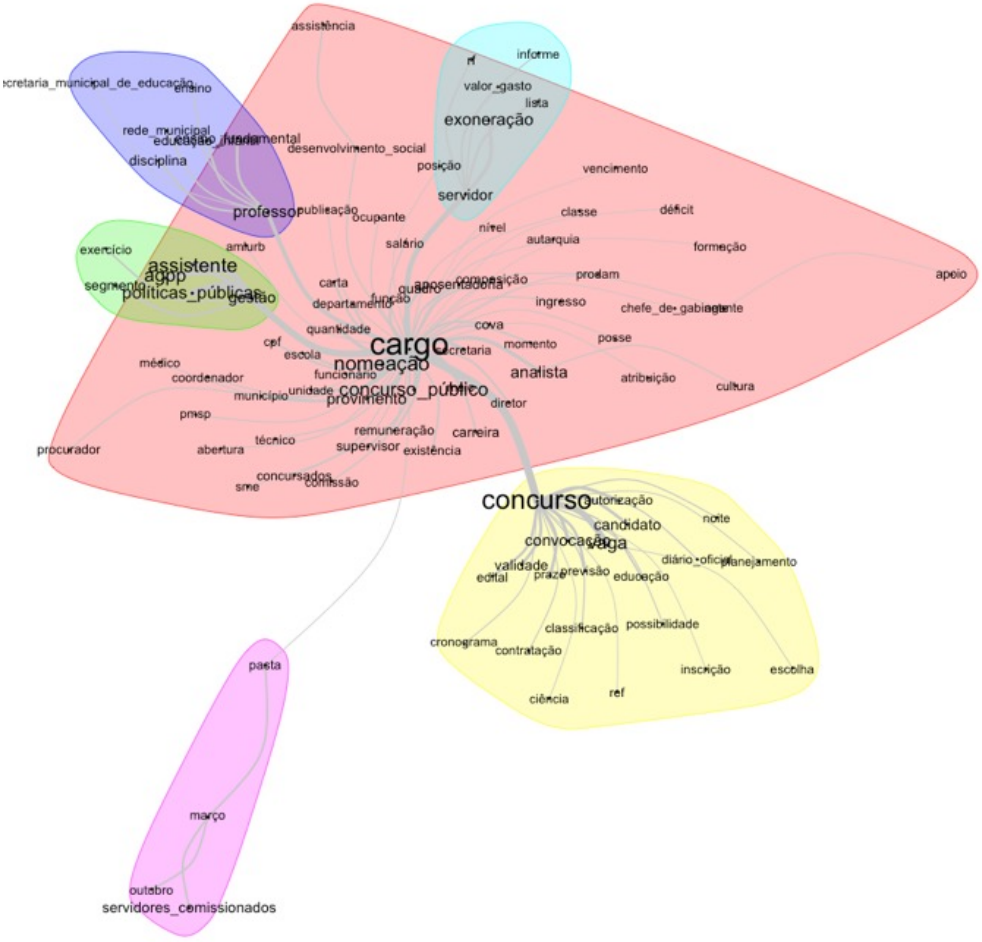
As classes foram nomeadas após a realização de uma análise amostral de 350 pedidos, correspondendo a 50 de cada uma (ver arquivo suplementar a esse artigo[7]). Observou-se que as palavras, expressões e siglas mais citadas geraram no seu entorno subclasses que podem ser consideradas para análise e tomada de decisões relacionadas aos temas.

7. O arquivo está disponível em: https://github.com/chfsantos/RevistaCGU_Ciencia_de_Dados_PMSP/blob/main/Amostra_das_classes.zip

Classe 7 – Cargos e Concursos Públicos

Composta por 2.307 pedidos de acesso à informação (7,2%), consistindo por palavras, expressões e siglas como: cargo (X^2 = 10.013,75), concurso público (X^2 = 3.420,53), AGPP (sigla de assistente de gestão de políticas públicas, cargo na Prefeitura Municipal de São Paulo) (X^2 = 3.252,40), e provimento (X^2 = 1.875,47).

FIGURA 3 – COOCORRÊNCIA DAS FORMAS ATIVAS DA CLASSE 7

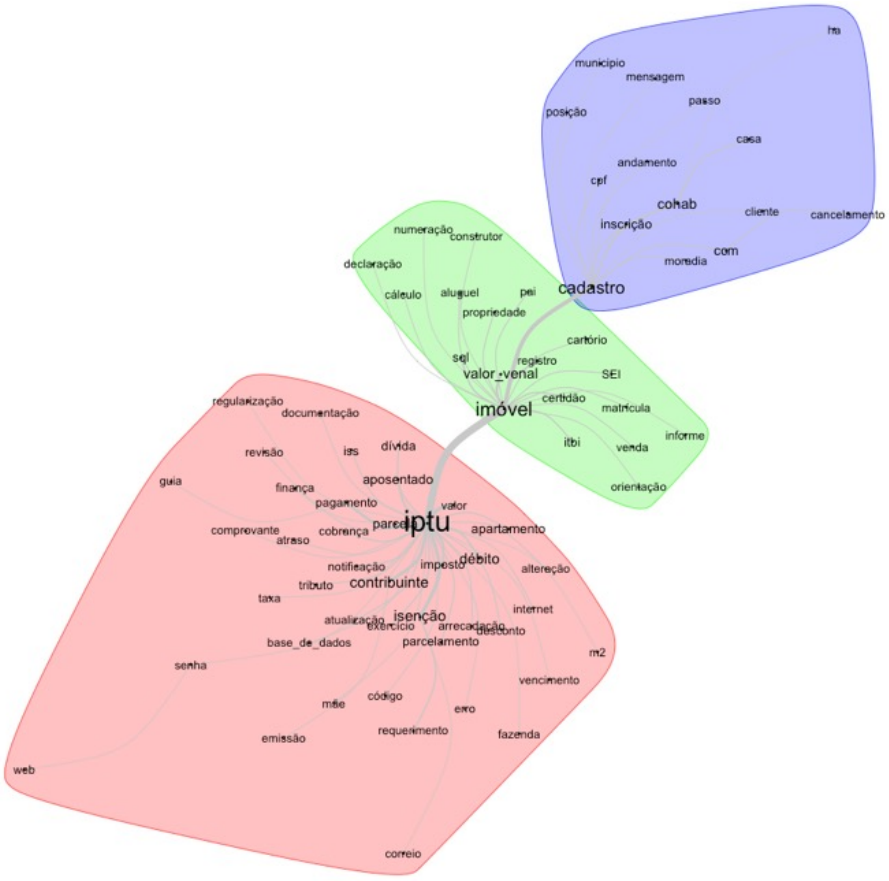


Fonte: elaboração própria

Classe 6 – Imóveis

Composta por 2.171 pedidos (6,8%), formada por palavras, expressões e siglas como: IPTU (X^2 = 9.781,26), imóvel (X^2 = 4.405,18), cadastro (X^2 = 3.130,69), contribuinte (X^2 = 1.944,35) e valor venal (X^2 = 1.540,36).

FIGURA 4 – COOCORRÊNCIA DAS FORMAS ATIVAS DA CLASSE 6

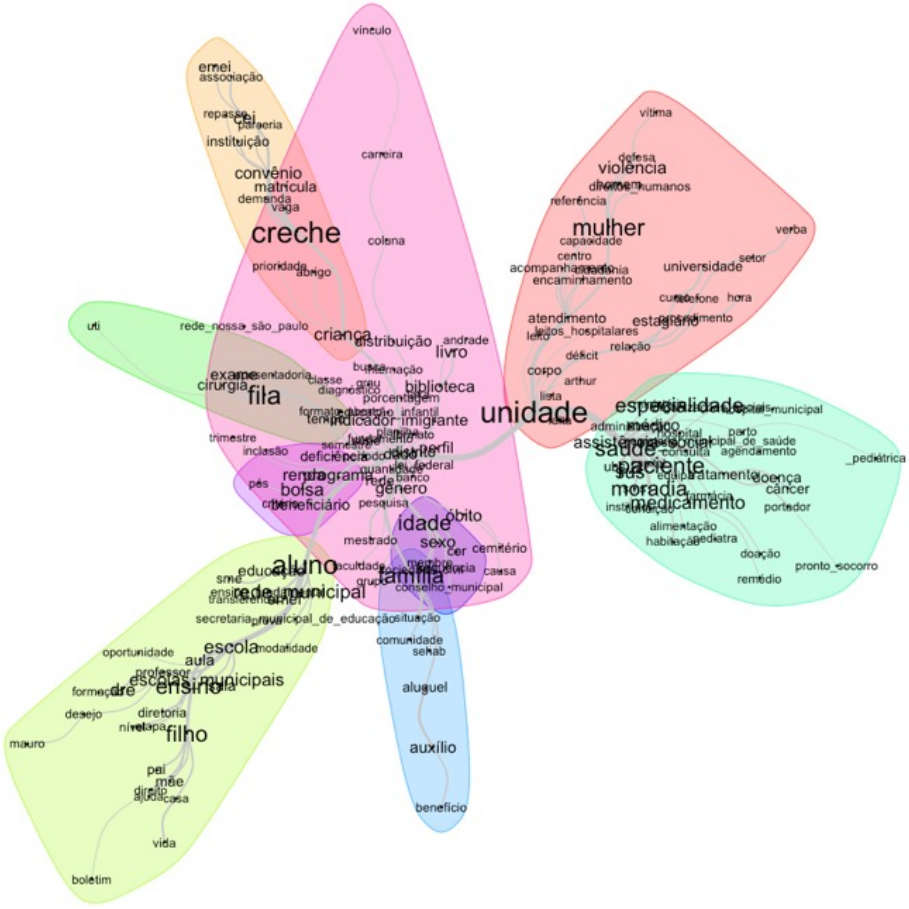


Fonte: elaboração própria

Classe 5 – Família: saúde, educação e assistência social

Composta por 3.788 pedidos (11,8%), formada por palavras, expressões e siglas como: unidade (X^2 = 1.518,01), creche (X^2 = 1.482,78), aluno (X^2 = 1.229,53), rede municipal (X^2 = 617,78) e SUS (Sistema Único de Saúde) (X^2 = 616,78).

FIGURA 5 – COOCORRÊNCIA DAS FORMAS ATIVAS DA CLASSE 5

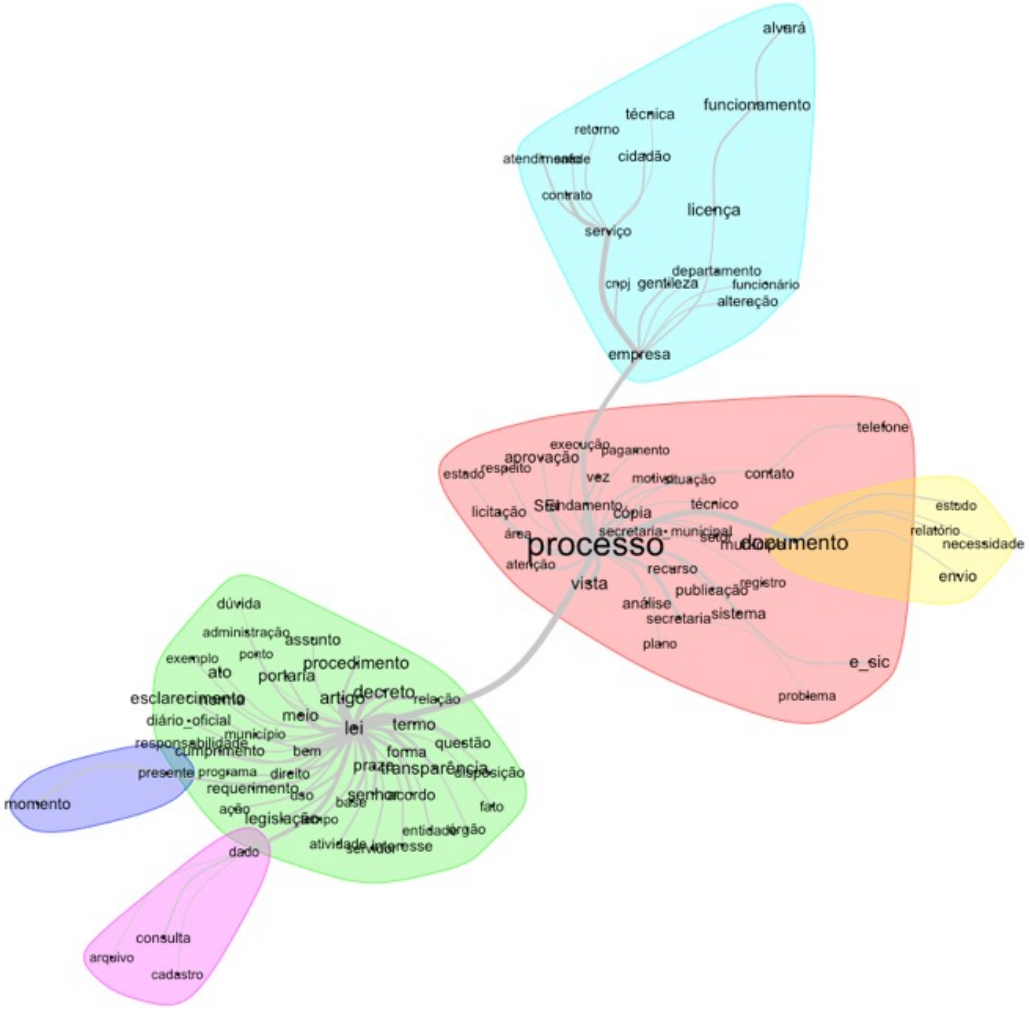


Fonte: elaboração própria

Classe 2 – Trâmite e documentos processuais

Composta por 6.665 pedidos (20,8%), constituída por palavras, expressões e siglas como: processo (X^2 = 3.195,27), documento (X^2 = 1.303,76), vista (X^2 = 722,25), processo administrativo (X^2 = 321,35) e despacho (X^2 = 287,61).

FIGURA 6 – COOCORRÊNCIA DAS FORMAS ATIVAS DA CLASSE 2

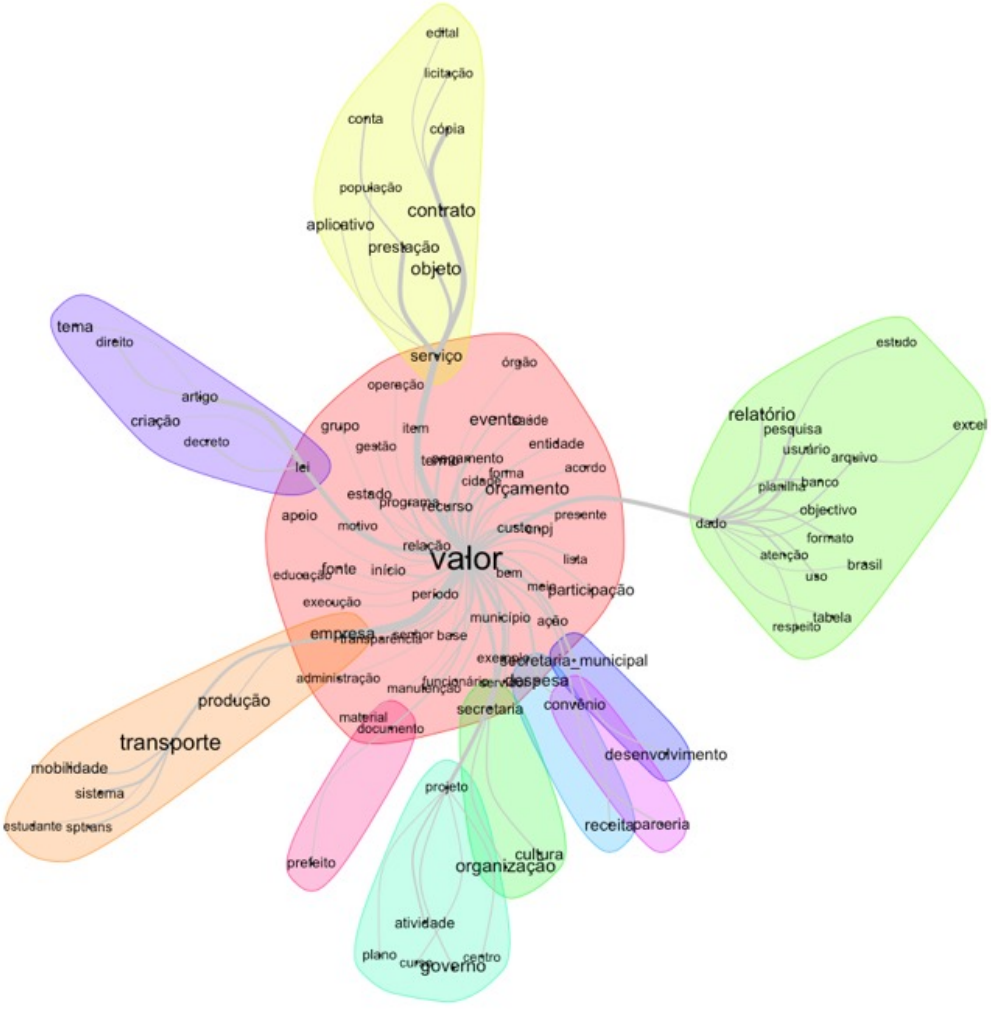


Fonte: elaboração própria

Classe 3 – Contratações públicas

Composta por 5.239 pedidos (16,4%), formada por palavras, expressões e siglas como: valor (X^2 = 2.600,92), transporte (X^2 = 1.187,29), contrato (X^2 = 723,29), objeto X^2 = 673,31) e orçamento (X^2 = 562,12).

FIGURA 7 – COOCORRÊNCIA DAS FORMAS ATIVAS DA CLASSE 3

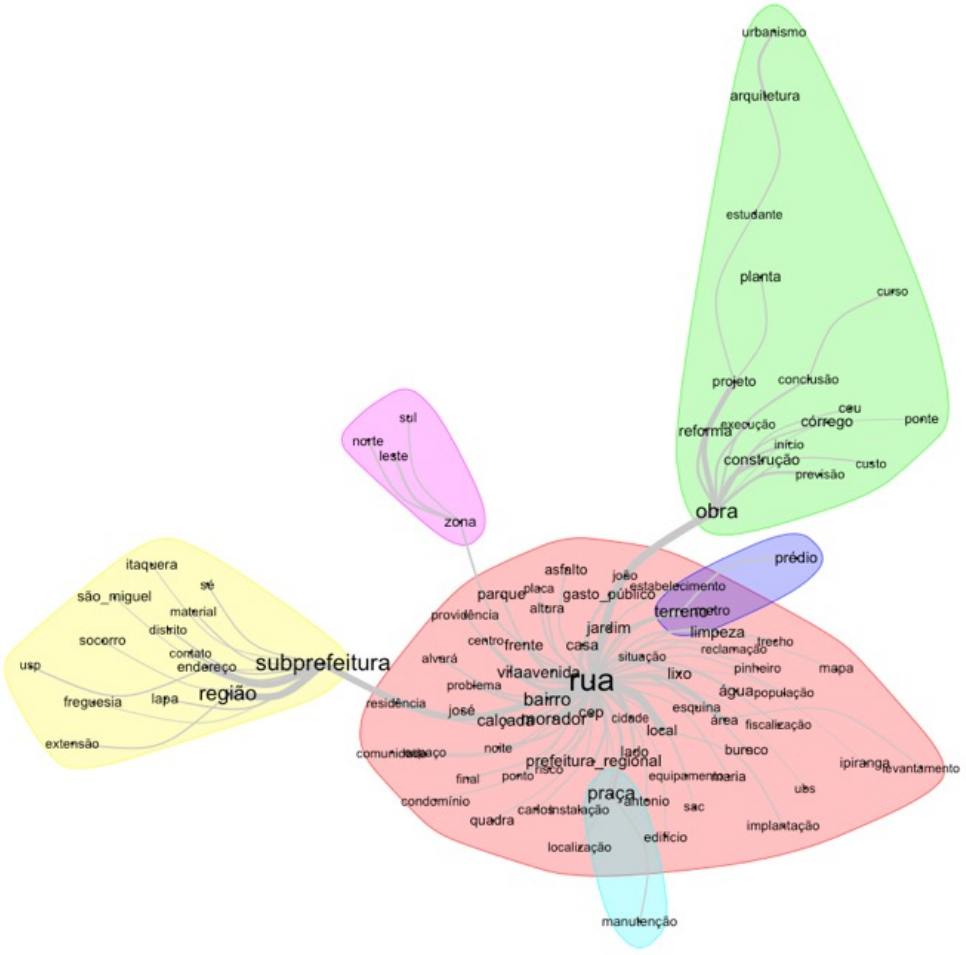


Fonte: elaboração própria

Classe 1 – Bairros e distritos

Composta por 7.404 pedidos (23,2%), constituída por palavras, expressões e siglas como: rua (X^2 = 3.551,54), obra (X^2 = 1.531,77), região (X^2 = 1.523,25), bairro (X^2 = 1.079,21) e prefeitura regional (X^2 = 641,82).

FIGURA 8 – COOCORRÊNCIA DAS FORMAS ATIVAS DA CLASSE 1



Fonte: elaboração própria

Classe 4 – Mobilidade urbana

Composta por 4.372 pedidos (13,7%), formada por palavras, expressões e siglas como: ônibus (X^2 = 4.926,60), trânsito (X^2 = 2.243,07), CET (X^2 = 2.198,87), radar (X^2 = 1.306,42) e mobilidade urbana (X^2 = 233,45).

FIGURA 9 – COOCORRÊNCIA DAS FORMAS ATIVAS DA CLASSE 4

Fonte: elaboração própria

Relação entre Classes e Variáveis

A partir do Corpus contendo os pedidos de acesso e suas variáveis foi possível se obter a relação entre as variáveis associadas aos pedidos de acesso à informação e as classes resultantes da CHD.

Classes x ano

Analisando a Tabela 4, que representa o qui quadrado (X^2) de associação dos anos com cada uma das classes, observa-se que a predominância das classes não foi a mesma ao longo dos anos.

TABELA 4 – ASSOCIAÇÃO ENTRE AS CLASSES DA CHD E ANOS

ANO	CLASSE 1	CLASSE 2	CLASSE 3	CLASSE 4	CLASSE 5	CLASSE 6	CLASSE 7
2012	-6,49	5,98	1,49	16,35	-1,04	-2,50	-15,90
2013	-2,69	4,25	18,56	6,04	-26,17	16,71	-57,34
2014	-5,98	4,17	-4,71	11,34	0,01	1,08	-2,87
2015	-0,47	-12,31	-0,91	52,77	5,13	3,95	-41,23
2016	4,55	-8,45	-15,17	21,13	-2,68	0,31	4,23
2017	3,67	-20,34	3,96	5,70	1,82	-59,16	13,96
2018	30,30	-3,80	-30,68	-18,39	13,26	0,65	5,63
2019	-30,38	69,71	35,17	-139,29	-3,98	3,57	13,84

Fonte: elaboração própria.

Por exemplo, a Classe 2 (Trâmite e Documentos Processuais) apresentou no ano de 2019 (X^2 = 69,71) uma maior associação, comparando com os demais anos, enquanto à Classe 4 (Mobilidade Urbana), observa-se um comportamento positivo desde 2012 (X^2 = 16,35), com um crescimento em 2015 (X^2 = 52,77), passando, podemos dizer, a uma desassociação média em 2018 (X^2 = -18,39) e alta em 2019 (X^2 = -139,28), denotando que, nesses dois anos, o tema foi menos mencionado. Do mesmo modo, a classe 6 (Imóveis), apresentou uma desassociação no ano de 2017 (X^2 = -59,16). Observa-se que a Classe 7 (Cargos e Concursos Públicos) apresentou baixa associação com os anos 2013 (X^2 = -57,34) e 2015 (X^2 = -41,22). Essas variações indicam mudanças no interesse da população no assunto, que, associadas a outras informações, poderão colaborar na tomada de decisões relacionadas a políticas públicas afetas ao tema.

Classes x solicitantes

As associações entre as classes dos pedidos e os tipos de solicitantes estão representadas na Tabela 5.

TABELA 5 – ASSOCIAÇÃO ENTRE AS CLASSES DA CHD E SOLICITANTES

TIPO	CLASSE 1	CLASSE 2	CLASSE 3	CLASSE 4	CLASSE 5	CLASSE 6	CLASSE 7
Pessoa Física	-0,28	-138,81	-93,72	137,51	19,73	-0,01	148,56
Pessoa Jurídica	0,28	138,81	93,72	-137,51	-19,73	0,01	-148,56

Fonte: elaboração própria.

Observa-se que as classes 2 – Trâmite e documentos processuais (X^2 = 138,81) e 3 – Contratações públicas: projetos, compras, despesas e orçamentos (X^2 = 93,72) tiveram mais associação com pedidos de pessoas jurídicas, enquanto as classes 4 – Mobilidade Urbana (X^2 = 137,50) e 7 – Cargos e Concursos Públicos (X^2 = 148,56) mais associaram-se com pedidos de pessoas físicas.

Classes x Sexo dos Solicitantes

As associações entre as classes dos pedidos e os tipos de solicitantes estão representadas na Tabela 6.

TABELA 6 – ASSOCIAÇÃO ENTRE AS CLASSES DA CHD E O SEXO DO/A SOLICITANTE

SEXO	CLASSE 1	CLASSE 2	CLASSE 3	CLASSE 4	CLASSE 5	CLASSE 6	CLASSE 7
Feminino	1,86	-52,81	-86,94	-41,37	216,55	23,82	62,97
Masculino	-5,96	1,40	3,32	125,85	-140,14	-0,46	0,00

Fonte: elaboração própria.

Destacou-se uma associação entre a Classe 5 – Família: saúde, educação e assistência social (X^2 = 216,55) e solicitantes do sexo feminino, e uma associação entre a Classe 4 – Mobilidade Urbana (X^2 = 125,80) e solicitantes do sexo masculino. Associações como essas são relevantes, por exemplo, para os gestores públicos, por exemplo, decidirem como focar as decisões e mensagens relacionadas a tais políticas públicas.

Classes x órgãos destinatários dos pedidos de acesso à informação

A Tabela 7 mostra que os órgãos 60 (X^2 = 7.132,52) e 67 (X^2 = 2.176,75) foram destacadamente associados à Classe 4 (Mobilidade Urbana), enquanto os órgãos 10 (X^2 = 1.961,22) e 16 (X^2 = 1.504,90) tiveram maior associação à Classe 5 (Família: saúde, educação e assistência social). Continuando, temos que os órgãos 18 (X^2 = 5.643,42) e 62 (X^2 = 1.801,91) foram os com maior associação à Classe 6 (Imóveis), e os órgãos 10 (X^2 = 1.285,89) e 22 (X^2 = 2.609,22) representaram os mais associados à Classe 7 (Cargos e Concursos Públicos).

TABELA 7 - ASSOCIAÇÃO ENTRE OS ÓRGÃOS E AS CLASSES DA CHD

ÓRGÃO / NOME DO ÓRGÃO	CLASSE						
	1	2	3	4	5	6	7
67 - SPTrans - São Paulo Transportes S/A	-354,64	-80,30	91,52	2.176,75	-85,52	-122,21	-73,44
16 - SMS - Secretaria Municipal da Saúde	-65,64	-2,23	-8,31	-49,75	1.504,91	-175,35	-42,70
10 - SME - Secretaria Municipal de Educação	-285,86	-126,50	-33,96	-372,73	1.961,22	-151,88	1.285,89
22 - SG - Secretaria Municipal de Gestão	-228,64	0,05	18,73	-182,03	-40,66	-53,44	2.609,22
18 - SF - Secretaria Municipal da Fazenda	-432,02	25,06	0,27	-295,21	-182,20	5.643,42	-65,51
62 - COHAB - Companhia Metropolitana de Habitação	-18,68	-17,37	-76,70	-91,26	8,86	1.801,91	-38,34
60 - CET - Companhia de Engenharia de Tráfego	-148,63	-244,16	-157,73	7.132,52	-307,88	-197,05	-201,84

Informações como essas, somente obtida através de análises como a CHD, apresentam aos órgãos oportunidade de trabalharem em conjunto, tanto os pedidos, análises, decisões, respostas e transparência ativa, quanto as políticas públicas associadas a eles, em busca de soluções, também tomadas em parceria.

CONCLUSÃO

A pesquisa realizada permitiu evidenciar a relevância do texto como dado para pesquisas, principalmente quando se dispõe de métodos e técnicas como os que a ciência de dados tem nos trazido nas primeiras décadas do século XXI. Com a CHD, uma das técnicas de mineração e classificação de textos, foi possível identificar, com mais clareza e de modo agregado, as temáticas mais abordadas nos pedidos de acesso à informação feitos à Prefeitura de São Paulo, entre os anos de 2012 e 2019, bem como dos órgãos destinatários das solicitações dos cidadãos. Destacaram-se os pedidos de acesso à informação relacionados a "Bairros e Distritos" (23,18%), "Trâmite e documentos processuais" (20,86%) e "Contratações Públicas" (16,40%). Sem a Ciência de Dados e suas ferramentas, não seria fácil interpretar e classificar quase 40.000 pedidos de acesso à informação, reunindo informações agregadas de apoio à tomada de decisão em contexto público.

Observando as Figuras 3 a 9 foi possível identificar, pelas diferentes cores, subclasses dentro de cada uma das classes, permitindo-se ir ainda mais fundo nos questionamentos apresentados, sendo cada cor uma dessas subclasses. Por fim, temos que a identificação de relações entre as classes temáticas e os órgãos a quem se destinaram os pedidos apresenta uma oportunidade para que o tratamento da informação e a tomada de decisão possa ser articulada em conjunto. Os problemas podem ser identificados em parceria, bem como as soluções podem ser buscadas em colaboração entre os órgãos.

A análise dos dados através da Classificação Hierárquica Descendente permite obter informação relevante para a tomada de decisão baseada em dados e evidências. Esta abordagem baseada em dados verificáveis e com qualidade favorece a existência de decisões fundadas e mais próximas das necessidades dos cidadãos.

Limitações do Trabalho e Sugestões de Investigações Futuras

Como limitações da investigação, tivemos a ausência de dados dos anos de 2020 a 2022, período em que o mundo enfrentou a pandemia da covid-19. Será relevante ver as prováveis alterações temáticas ocorridas. Percebeu-se também que alguns pedidos, tecnicamente, não são pedidos de acesso à informação nos termos da lei, mas pedidos de ajuda ou de solução de problemas. Outra limitação foi a existência de pedidos repetidos, para o mesmo órgão e para órgãos diferentes, que, apesar de terem sido identificados, não foram excluídos, podendo ter levado a ocorrência de algum viés (erro sistemático).

Para futuras explorações propomos a realização de estudos com dados de outras cidades e de outras esferas federativas, para verificar se há diferenças geopolíticas capazes de alterar a temáticas dos pedidos de acesso à informação, bem como para compreender a relação deste tipo de informação com os mecanismos de *accountability* disponíveis.

REFERÊNCIAS

Alves, M. S. D. (2012). Do sigilo ao acesso: Análise tópica da mudança de cultura. *Revista do Tribunal de Contas do Estado de Minas Gerais, 85(esp.)*, 120-134. https://revista1.tce.mg.gov.br/Content/Upload/Materia/1683.pdf

Angélico, F., & Teixeira, M. A. C. (2012). Acesso à informação e ação comunicativa: Novo trunfo para a gestão social. *Desenvolvimento em Questão, 10(21)*, 7-27. https://doi.org/10.21527/2237-6453.2012.21.7-27

Bagozzi, B. E., Berliner, D., & Almquist, Z. W. (2019). When does open government shut? Predicting government responses to citizen information requests. *Regulation & Governance, 15(2)*, 280-297. https://doi.org/10.1111/rego.12282

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: A comparison of retrieval performances. *Lecture Notes on Software Engineering, 2(3)*, 262-267. https://doi.org/10.7763/lnse.2014.v2.134

Banisar, D. (2006). *Freedom of information around the world 2006: A global survey of access to government information laws*. Privacy International. https://www.humanrightsinitiative.org/programs/ai/rti/international/laws_papers/intl/global_foi_survey_2006.pdf

Berliner, D. Bagozzi, B. E, & Palmer-Rubin, B. (2018). What information do citizens want? Evidence from one million information requests in Mexico. *World Development, 109*, 222-235. https://doi.org/10.1016/j.worlddev.2018.04.016

Berliner, D., Bagozzi, B. E., Palmer-Rubin, B., & Erlich, A. (2021). The political logic of government disclosure: Evidence from information requests in Mexico. *The Journal of Politics, 83(1)*, 229-245. https://doi.org/10.1086/709148

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys, 50(3)*, 1-42. https://doi.org/10.1145/3076253

Centre for Law and Democracy. (2017). *Global Right to Information Rating*. http://www.rti-rating.org/country-data

Constituição da República Federativa do Brasil de 1988. (2022). Presidência da República. http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm

Controladoria-Geral da União. (2019). *Escala Brasil Transparente 360º: Metodologia e critérios de avaliação*. Controladoria-Geral da União. https://mbt.cgu.gov.br/static/Metodologia%20EBT.pdf

Convenção das Nações Unidas contra a Corrupção. (2003). Escritório das Nações Unidas contra Drogas e Crime. https://www.unodc.org/documents/lpo-brazil//Topics_corruption/Publicacoes/2007_UNCAC_Port.pdf

Declaração dos Direitos do Homem e do Cidadão. (1789). Assembleia Nacional. https://br.ambafrance.org/A-Declaracao-dos-Direitos-do-Homem-e-do-Cidadao

Declaração Universal dos Direitos Humanos. (1948). Organização das Nações Unidas. https://www.unicef.org/brazil/declaracao-universal-dos-direitos-humanos

Decreto nº 53.623, de 12 de dezembro de 2012. (2012, 13 de dezembro). Regulamenta a Lei Federal n° 12.527, de 18 de novembro de 2011, no âmbito do Poder Executivo, estabelecendo procedimentos e outras providências correlatas para garantir o direito de acesso à informação, conforme especifica. Prefeitura de São Paulo. http://legislacao.prefeitura.sp.gov.br/leis/decreto-53623-de-12-de-dezembro-de-2012

Duarte, J., & Theorga, A. B. (2012). O processo de implantação da Lei de Acesso à Informação em órgãos do Poder Executivo federal. *Comunicação & Informação, 15(2)*, 66-79. https://doi.org/10.5216/c&i.v15i2.24568

Flores, A. M., Pavan, M. C., & Paraboni, I. (2022). User profiling and satisfaction inference in public information access services. *Journal of Intelligent Information Systems, 58(1)*, 67-89. https://doi.org/10.1007/s10844-021-00661-w

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57(3)*, 535-574. https://doi.org/10.1257/jel.20181020

Hollibaugh, G. E., Jr. (2019). The use of text as data methods in public administration: A review and an application to agency priorities. *Journal of Public Administration Research and Theory, 29(3)*, 474-490. https://doi.org/10.1093/jopart/muy045

Instituto Brasileiro de Geografia e Estatística. (2010). Censo 2010: São Paulo: panorama. *Portal do Governo Brasileiro*. https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama

Kelleher, J. D., & Tierney, B. (2018). *Data science*. MIT Press.

Kotu, V., & Deshpande, B. (2018). *Data science: Concepts and practice* (2nd ed.). Morgan Kaufmann.

Lane, J., Gimeno, E., Levitskaya, E., Zhang, Z., & Zigoni, A. (2022). Data inventories for the modern age? Using data science to open government data. *Harvard Data Science Review*, 4.2, 1-45. https://doi.org/10.1162/99608f92.8a3f2336

Lei Modelo Interamericana sobre o Acesso à Informação Pública. (2010). Organização dos Estados Americanos. http://www.oas.org/dil/AG-RES_2607-2010_por.pdf

Lei nº 12.527, de 18 de novembro de 2011. (2011, 18 de novembro). Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei nº 8.112, de 11 de dezembro de 1990; revoga a Lei nº 11.111, de 5 de maio de 2005, e dispositivos da Lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. Presidência da República. http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm

Martin, F., & Johnson, M. (2015, 8-9 December). More efficient topic modelling through a noun only approach. *Proceedings of the Australasian Language Technology Association Workshop 2015*, 13, 111-115. https://aclanthology.org/U15-1013/

Mendes, A. M., Tonin, F. S., Buzzi, M. F., Pontarolo, R., & Fernandez-Llimos, F. (2019). Mapping pharmacy

journals: A lexicographic analysis. *Research in Social & Administrative Pharmacy, 15(12)*, 1464-1471. https://doi.org/10.1016/j.sapharm.2019.01.011

Porumbescu, G.; Meijer, A.; Grimmelikhuijsen, S. *Government transparency*: state of the art and new perspectives. London: Cambridge University Press, 2022.

Rydholm, L. (2013). China and the World's First Freedom of Information Act: The Swedish Freedom of the Press Act of 1766. *Javnost-The Public, 20(4)*, 45-63. https://doi.org/10.1080/13183222.2013.11009127

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *Plos One, 16(8)*, e0254937. https://doi.org/10.1371/journal.pone.0254937

Savage, A., & Hyde, R. (2014). Using freedom of information requests to facilitate research. *International Journal of Social Research Methodology, 17(3)*, 303-317. https://doi.org/10.1080/13645579.2012.742280

Schofield, A., Magnusson, M., & Mimno, D. (2017, 3-7 April). Pulling out the stops: Rethinking stopword removal for topic models. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2*, 432-436. https://aclanthology.org/E17-2069/

Soares, G. F. (2020). Ciência de dados aplicada à auditoria interna. *Revista da CGU, 12(22)*, 196-208. https://doi.org/10.36428/revistadacgu.v12i22.195

Sousa, Y. S. O. (2021). O uso do software Iramuteq: Fundamentos de lexicometria para pesquisas qualitativas. *Estudos e Pesquisas em Psicologia, 21(4)*, 1541-1560. https://doi.org/10.12957/epp.2021.64034

Walby, K., & Larsen, M. (2012). Access to information and freedom of information requests: Neglected means of data production in the social sciences. *Qualitative Inquiry, 18(1)*, 31-42. https://doi.org/10.1177/1077800411427844

Walby, K., & Luscombe, A. (2017). Criteria for quality in qualitative research and use of freedom of information requests in the social sciences. *Qualitative Research, 17(5)*, 537-553. https://doi.org/10.1177/1468794116679726

Wang, Z., & Zhong, Y. (2020). What were residents' petitions in Beijing-based on text mining. *Journal of Urban Management, 9(2)*, 228-237. https://doi.org/10.1016/j.jum.2019.11.006

Claudio Henrique Fontenelle Santos
chfs@iscsp.ulisboa.pt
ORCID: https://orcid.org/0000-0001-5237-2461

Doutorando em Administração Pública pelo Instituto Superior de Ciências Sociais e Políticas da Universidade de Lisboa, Mestre em Administração Pública pela Universidade Federal da Bahia, Especialista em Comunicação, Publicidade e Propaganda pela Universidade de Fortaleza, Bacharel em Arquitetura e Urbanismo pela Universidade de Fortaleza e Bacharel em Administração de Empresas pela Universidade Estadual do Ceará. Pesquisa sobre transparência, Acesso à informação, governo aberto e ouvidorias públicas.

Ana Lúcia Romão
anaromao@iscsp.ulisboa.pt
ORCID: https://orcid.org/0000-0003-2730-4007

Doutorada em Economia, Investigadora no Centro de Administração e Políticas Públicas e Professora Auxiliar no Instituto Superior de Ciências Sociais e Políticas da Universidade de Lisboa. É Vice-Presidente do Centro de Administração e Políticas Públicas, Coordenadora Adjunta da Unidade de Coordenação de Administração Pública e Coordenadora Científica da Pós-Graduação em Contabilidade e Gestão Pública. Tem desenvolvido investigação na área da administração pública, nomeadamente no domínio da gestão pública, do controlo da gestão pública e da contabilidade pública.