risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 16/02/2022
Aceitação/Acceptance: 10/05/2022

Requisitos para a ciência de dados: analisando anúncios de vagas de emprego com mineração de texto

André José Ribeiro Guimarães[1], Ricardo Mendes Júnior[1], Maria do Carmo Duarte Freitas[1]

andrejose@ufpr.br; ricardomendesjr@gmail.com; mcf@ufpr.br

[1] Universidade Federal do Paraná, Av. Prefeito Lothário Meissner, 632, 80210-170, Curitiba, Brasil

DOI: 10.17013/risti.46.54–70

Resumo: Esta pesquisa identifica os requisitos para cientistas de dados no Brasil em anúncios de emprego. Para analisar estes documentos, adota métodos de mineração de texto: n-grama, modelagem de tópico e agrupamento. Os resultados apontam uma concentração de vagas em São Paulo e revelam que a modalidade remota é a segunda mais ofertada. Além disso, destaca que os salários no Brasil estão abaixo da média de outros países, mesmo que as organizações procurem por profissionais experientes e com alto nível educacional. Quanto aos requisitos, há o predomínio de habilidades técnicas como *machine learning*, modelos estatísticos, python, banco de dados, dentre outras. Para as técnicas de mineração, demonstra que n-grama e o agrupamento são mais adequadas que a modelagem de tópicos.

Palavras-chave: Cientista de dados; Mineração de texto; Requisitos para cientista de dados; Competências.

Requirements for Data Science: analyzing job postings with text mining

Abstract: This research identifies in job postings the requirements for data scientists in Brazil. To analyze these documents, it adopts text mining methods of analysis: n-gram, topic modeling, and clustering. The findings point to a concentration of job opportunities in São Paulo while demonstrating that the remote modality is the second most offered. Additionally, it highlights that salaries in Brazil are below the average of other countries, even if organizations look for experienced professionals with an elevated level of education. About the requirements, there is a predominance of technical skills such as machine learning, statistical models, python, and database, among others. The results also demonstrate that n-gram and clustering are more suitable for text mining techniques than topic modeling.

Keywords: Data scientist; Text mining; Requirements for data scientist; Competencies.

1. Introdução

Uma das principais características do big data é a quantidade de dados gerada diariamente, que afeta todas as áreas da sociedade (Finzer, 2013; Stark & Hawamdeh, 2018). Juntamente com o avanço tecnológico e dos métodos de análise, a ciência de dados emerge como um campo que oferece alternativas mais precisas a questões complexas, além de melhorar a qualidade do processo decisório, seja no setor de negócios, governamental ou acadêmico-científico (Grossi et al., 2021; Provost & Fawcett, 2013a). Neste cenário, em busca de vantagens competitivas, as organizações procuram contratar profissionais com competências para coletar, gerenciar e analisar esses dados (Cao, 2019; Gottipati et al., 2021).

Esse profissional responsável por lidar com grandes conjuntos de dados é usualmente chamado de cientista de dados (Baumeister et al., 2020), um especialista interdisciplinar, capaz de atuar em todos as fases de um problema referente a dados, da coleta inicial às conclusões finais (Loukides, 2012). Desde que Davenport e Patil (2012) definiram o cientista de dados como a profissão mais "sexy" do século XXI, o interesse por este profissional só aumentou. De acordo com um relatório do LinkedIn (2020), cientista de dados ocupa a terceira posição do ranking de profissões emergentes. A lista é liderada pelo cargo de "especialista em inteligência artificial", outro profissional fortemente relacionado à ciência de dados.

Por outro lado, contar com um cientista de dados é uma tarefa árdua e dispendiosa para as organizações. As dificuldades começam por encontrar profissionais com pensamento computacional, habilidades científica e analítica, e se mantêm pelo complicado trabalho de reter os talentos da área (Davenport & Patil, 2012; Reis & Sá, 2020). É reconhecido que não há no mercado número suficiente de pessoas qualificadas para desenvolver projetos que envolvam técnicas complexas de análise, como *machine learning* e *deep learning*, nem para implementar mudanças organizacionais rumo a uma cultura orientada a dados (Cunha, 2018; Hall et al., 2016). Para agravar a situação, como a ciência de dados é uma área em desenvolvimento, não há entendimento ou clareza terminológica acerca das descrições dos diferentes profissionais desse cenário (Halwani et al., 2021). Esse fato traz prejuízos para todos os atores envolvidos: empresas, instituições de ensino e, claro, os próprios indivíduos.

Para colaborar para a superação desses problemas, foi realizada uma pesquisa que teve por objetivo busca identificar os principais requisitos apresentados em anúncios de emprego para cientistas de dados no Brasil. Não há registro de pesquisa semelhante no mercado brasileiro ou em língua portuguesa. Para atingir esse objetivo, as descrições de vagas para cientista de dados, coletadas em seis *websites* especializados, foram submetidas a um processo de mineração de texto. Esse tipo de mineração de dados emprega abordagens estatísticas, como técnicas de agrupamento e análise fatorial, para relevar relações ocultas, ou mesmo reforçar relações já conhecidas, simplificando a representação do conteúdo semântico presente em grandes volumes de dados não estruturados (Wolfram, 2017). As seguintes perguntas foram formuladas: quais são os requisitos mais recorrentes exigidos para o cargo de cientista de dados? É possível identificar tópicos dentre os documentos analisados? Os resultados vão ao encontro de

pesquisas anteriores? A mineração de texto se mostra uma técnica válida para analisar os anúncios?

Aumentar o entendimento sobre os requisitos e, por consequência, das competências dos profissionais de dados beneficia o mercado de trabalho, quem será empregado, a academia, mas também os profissionais que atuam na formação em ciência de dados (Halwani et al., 2021). Nesse sentido, espera-se que, ao responder as perguntas acima, a pesquisa possa contribuir com essa área que se mostra cada vez mais importante à sociedade. Ademais, entre os critérios de ineditismo, aponta-se a ausência de trabalho com foco no mercado brasileiro, ou mesmo, na língua portuguesa. Outro elemento de originalidade é a adoção da raspagem de dados e mineração de texto, processos automatizados para coleta e análise dos dados.

2. Revisão da literatura

2.1. Ciência de dados

De maneira geral, a ciência de dados é compreendida como a extração metodológica de conhecimento a partir de quantidades massivas de dados (Dhar, 2013; NIST Big Data Public Working Group, 2015). Uma visão mais abrangente é apresentada por Chen et al. (2018) quando definem a ciência de dados como um campo interdisciplinar que visa beneficiar os seres humanos, por meio da combinação da metodologia científica e da tecnologia computacional voltada à gestão, acesso, análise e avaliação dos dados. Reforçando essa interdisciplinaridade, Grossi et al. (2021) afirmam que a ciência de dados combina diferentes teorias, práticas e modelos, configurando-se em um paradigma pervasivo que envolve múltiplas disciplinas, com potencial inovativo para a ciência, indústria, política e, consequentemente, para a vida das pessoas.

Para as organizações, a potencialidade dos dados se transforma em um diferencial competitivo (Metelo et al., 2021; Provost & Fawcett, 2013b), porém as obriga a repensarem suas práticas analíticas (Hall et al., 2016). Dentre as mudanças trazidas pelo desenvolvimento da ciência de dados, está a ampliação na demanda por um novo tipo de profissional, com competência para lidar com volumes de dados cada vez maiores e heterogêneos (Curty & Serafim, 2016). Dentre as características desse profissional, chamado cientista de dados, estão a forte formação técnica, o conhecimento em tecnologias focadas em dados e habilidades voltadas à melhoria dos processos organizacionais (Demchenko et al., 2016).

A valorização dos dados e, por consequência, da ciência de dados enfatiza a relevância desse profissional em muitas áreas sociais e econômicas (Brandt, 2016). Embora recente, a ocupação de cientista de dados passa por um processo de profissionalização evidenciado pelo crescente número de cargos ofertados. Com papéis divididos e diversificados, predominantemente orientados a dados, esse profissional impacta áreas relacionadas a pesquisa, inovação, economia e na sociedade em geral (Cao, 2019).

Para o desenvolvimento desse campo promissor e necessário, é imprescindível a formalização das competências requeridas para esse profissional (Cao, 2019; Demchenko et al., 2017; Saltz & Grady, 2017). No Brasil, ainda que de forma mais lenta que outros países, a profissionalização do cientista de dados é corroborada pelo surgimento de

cursos para a formação desse profissional, sobretudo na categoria *lato sensu* (Breternitz et al., 2015). Além disso, se em 2016 o número de vagas ativas para cientistas de dados no Brasil não ultrapassava uma centena na rede social LinkedIn (Curty & Serafim, 2016), a busca pelas expressões "data scientist" e "cientista de dados", realizada no dia 19 de fevereiro de 2022, demonstrou que esse número triplicou em seis anos, conforme demonstrado Figura 1.

2.2. Competências em ciência de dados

Por ainda ser uma área em desenvolvimento, muitos cientistas de dados possuem formação em cursos universitários já estabelecidos, como física, economia, engenharia, mas principalmente, estatística e ciência da computação (Baškarada & Koronios, 2017). Conforme antecipado por Davenport e Patil (2012), profissionais que agreguem formação acadêmica com habilidades computacionais e analíticas são raros e caros, uma vez que são disputados pelo mercado. Desse modo, mesmo que as competências do cientista de dados estejam fundamentadas nestas na estatística e ciência da computação, sua atuação envolve outros fatores que os diferenciam de profissionais relacionados, como estatísticos, analistas de dados ou analistas de *business intelligence* (Kim & Lee, 2016).

Entre estes fatores diferenciais, destaca-se o conhecimento de domínio, ou *expertise*, que é o conhecimento da área, relevante para o desenvolvimento de aplicações de análise de dados (Baškarada & Koronios, 2017; Demchenko et al., 2016; Hall et al., 2016). Além disso, Loukides (2012) destaca como habilidades interpessoais do cotidiano do profissional de dados a paciência, a motivação em construir produtos de dados, a vontade em explorar e gerar soluções de maneira contínua e incremental, além da busca constante por respostas. Para mais, dentre outras competências associadas à prática da ciência de dados, estão comunicação oral e escrita, além de questões sociais, éticas e legais que implicam em conhecimento sobre privacidade, segurança e propriedade de dados (Anderson et al., 2014).

3. Metodologia

No geral, a pesquisa seguiu a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*), um conjunto de tarefas voltadas a projetos de mineração de dados. Segundo a CRISP-DM (Chapman et al., 2000), o ciclo de vida de um projeto de dados é composto por seis estágios: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implementação. Naturalmente, em uma pesquisa científica, procedimento previamente realizado por outros autores (Bedregal-Alpaca et al., 2020), essas etapas precisam ser adaptadas, uma vez que o produto final não é necessariamente uma implementação. A seguir, as estratégias para busca, coleta e tratamento dos dados são descritas.

3.1. Estratégia de busca, coleta e tratamento de dados

As fontes de dados da pesquisa foram *websites* especializados na divulgação de vagas voltadas à área de tecnologia, selecionados a partir da experiência dos autores e do resultado de pesquisa, com a expressão "vagas de emprego", realizada no mecanismo de busca Google. Foram selecionados três *websites* estrangeiros com versões voltadas ao

público brasileiro (LinkedIn, Indeed e Infojobs) e três empresas brasileiras destinadas ao mesmo fim (Catho, Empregos e Vagas). Nesta etapa, correspondente ao entendimento de negócio proposto pela CRISP-DM, procurou-se por APIs (*Application Programming Interface*) nos *websites* definidos, a fim de se verificar a disponibilidade de dados de interesse da pesquisa pelas plataformas consultadas. Uma vez que nenhuma API encontrada fornecia os dados requeridos, optou-se por empregar técnicas de raspagem de dados para acesso e coleta do material. A raspagem de dados, também conhecida como *web scraping*, corresponde ao uso de programas que percorrem páginas HTML procurando informações desejadas na estrutura de marcação e compilando os dados para a formação de um conjunto passível de ser analisado (Lantz, 2015).

Assim, como parte do entendimento dos dados, foram identificados quais informações dos anúncios de emprego poderiam ser extraídas pelo procedimento de raspagem: endereço URL do anúncio (*jobUrl*), título (*jobTitle*), resumo (*snippet*), descrição (*jobDescription*), nome da empresa (*companyName*), localidade da vaga (*companyLocation*), avaliação da empresa (*companyRating*), URL da empresa (*companyUrl*), remuneração (*salary*) e data de cadastro (*date*). Ainda que nem todos os seis *websites* e nem todas as vagas forneciam todas essas informações, sempre que possível, esses foram os dados coletados.

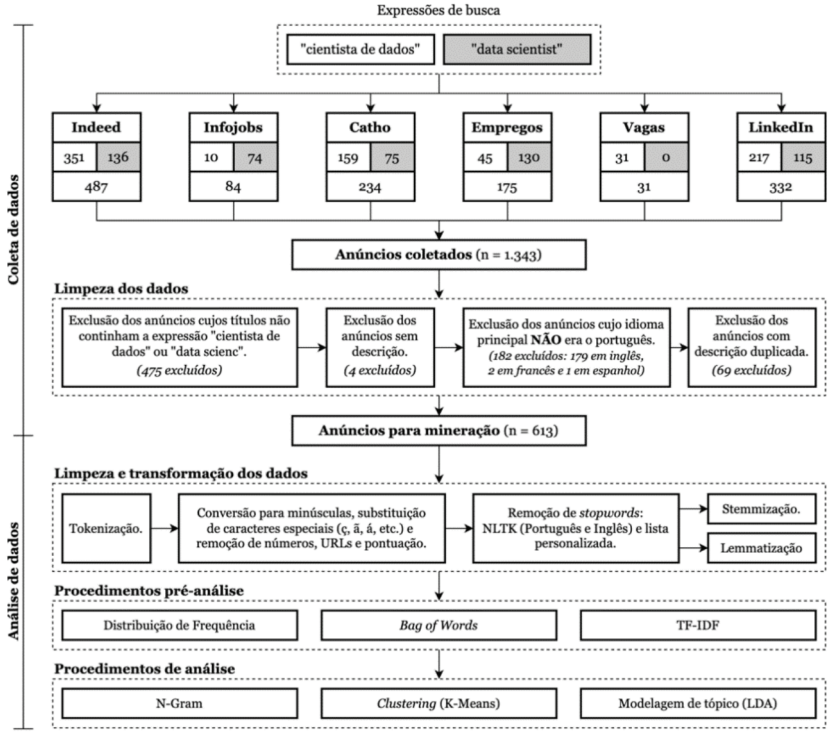

Figura 1 – Etapas da pesquisa, da coleta de dados aos procedimentos de análise

Os programas *web scraper* foram desenvolvidos em linguagem de programação Python, sendo adaptados para cada *website* acessado, visto que a estrutura HTML de cada um

é única. Além disso, para todos os *websites*, a raspagem foi organizada em duas etapas: 1) coleta de todos os anúncios com dados presentes na listagem (título, URL e resumo) e; 2) acesso individual a cada anúncio para coletar os demais dados disponíveis. Para buscar os anúncios, em cada *website* foram utilizadas as expressões "cientista de dados" e "data scientist" (dado que muitas empresas brasileiras utilizam termos em inglês em seus anúncios), filtrando por vagas no território brasileiro. Nesta etapa, entre 21 e 29 de setembro de 2021, foi possível coletar um total de 1.343 anúncios, cujos valores individuais de cada base são mostrados na Figura 1.

Uma vez que os mecanismos de busca dos *websites* não são atrelados aos títulos dos anúncios, percebeu-se que muitas vagas encontradas não se destinavam a cientista de dados. Por isso, como primeiro procedimento de tratamento, foram excluídos 475 anúncios que não continham nem "cientista de dados" e nem "data scien" em seus títulos. Em seguida, foram excluídos também quatro anúncios que não possuíam descrição, campo que será utilizado na mineração de texto, e 182 anúncios cujo idioma principal não era o português. Por fim, também foram removidos os anúncios com descrições duplicadas, uma vez que a mesma empresa poderia ter várias vagas para um cargo. Anúncios com descrições duplicadas afetariam a frequência com que os termos aparecem na mineração. Desse modo, ao final desses procedimentos, relacionados à "Preparação dos dados" da CRISP-DM, o *corpus* a ser analisado apresentava 613 documentos.

3.2. Protocolo de análise

Inicialmente, é necessário apresentar as principais características da amostra de anúncios que compuseram o *corpus* da pesquisa. Nesta etapa, além da participação de cada *website* utilizado, foram verificadas as localidades das vagas, os salários, o nível de escolaridade exigido, além do nível de experiência (Júnior, Pleno ou Sênior).

Em seguida, com o perfil das vagas conhecidos, procedeu-se à mineração de texto propriamente dita, que foi dividida em três passos:

1. **Limpeza e formatação dos dados:** ações específicas para a análise textual, conforme apresentado por Gajzer (2010). Primeiramente, os documentos (anúncios) passaram pelo processo de "tokenização" que é a separação das palavras (*tokens*). Cada documento foi convertido em uma lista de termos que, na etapa seguinte, foram contabilizados. Em seguida, todos os *tokens* foram convertidos para letras minúsculas e foram removidos caracteres especiais, números, URLs, dentre outros elementos textuais sem relevância para a análise. Assim, procedeu-se à remoção de *stopwords*, que são palavras muitos frequentes nos idiomas como artigos, preposições, conjunções e que possuem pouca carga semântica para a análise textual (Provost & Fawcett, 2013b). Por fim, para evitar variações como plural, foi adotado procedimentos de padronização de termos o processo chamado de ***stemmização***, que é a conversão de cada *token* para um termo menor (radical) (Bengfort et al., 2018). Por exemplo, os *tokens* "dados" e "dado" são convertidos para o *stem* "dad".
2. **Procedimentos de pré-análise:** com a remoção de elementos desnecessários e padronização dos termos restantes, é verificada a distribuição de frequência de palavra. Neste passo, foi computada a matriz de frequência conhecida como *Bag of Words* (BOW), que verifica a frequência de cada termo para todo o *corpus*,

e o TF-IDF (*Term Frequency* × *Inverse Document Frequency*), que normaliza a frequência dos *tokens* em um documento em relação ao restante do *corpus* (Bengfort et al., 2018).

3. **Procedimentos de análise:** para identificar padrões nos anúncios coletados, foram adotados três tipos de análise: a) **N-grama:** apresentação dos termos (*stems*) mais frequentes, além das expressões formadas por dois e três *stems* (Raschka, 2016); b) **Modelagem de tópicos:** para extrair os principais temas presentes nos anúncios foi adotada a técnica *Latent Dirichlet Allocation* (LDA), que pertence à família de modelos probabilísticos e emprega uma abordagem Bayesiana de duas camadas para identificar padrões de coocorrência de palavras, definindo tópicos do *corpus* (Wesslen, 2018). Para a LDA foi utilizada a biblioteca Gensim (Řehůřek & Sojka, 2011) com o auxílio da biblioteca de visualização pyLDAvis (Mabey, 2018) e; c) **Agrupamento:** por fim, implementação do algoritmo de agrupamento K-Means por meio da ferramenta *Clustering Workbench* (Carrot² Clustering Engine, 2021). O algoritmo K-Means, que é bastante popular, começa com um número arbitrário de agrupamentos e posiciona as instâncias (documentos) conforme sua proximidade com os centroides dos grupos, sendo que o objetivo final é minimizar a soma dos quadros na estruturada encontrada (Bengfort et al., 2018).

Segundo Wesslen (2018), a adoção de modelos de aprendizagem de máquina por pesquisadores das Ciências Sociais tem emergido como uma das principais técnicas para descoberta de variáveis latentes, que antes só poderiam ser medidas sob suposições não testáveis. Nesse sentido, esta pesquisa explora esses recursos, algoritmos de *machine learning* para mineração de texto, para identificar os principais requisitos para ser contratado como um cientista de dados. Além disso, o emprego da mineração de texto possibilitou a extração de padrões e conhecimento de centenas de documentos cuja análise manual seria mais onerosa.

A linguagem de programação Python foi adotada na etapa de limpeza e formatação dos dados, nos procedimentos pré-análise e, também, nas próprias análises aplicadas. Para visualizar os termos no seu contexto original, empregou-se o software livre AntConc (Anthony, 2022).

4. Apresentação e análise dos resultados

4.1. Características da amostra

Dos 613 anúncios eleitos para análise, 79,28% foram obtidos em duas únicas fontes: Indeed (250 anúncios) e LinkedIn (236 anúncios), ambas de origem estrangeira. Esses valores confirmam a relevância do site Indeed para área da ciência de dados que já tinha sido apresentada por Kim e Lee (2016). Por outro lado, a empresa Infojobs, de origem espanhola, apresenta a menor contribuição com apenas sete anúncios analisados (1,14%). Em relação à Infojobs, também chama a atenção que dos 84 anúncios coletados inicialmente foram excluídos 77, principalmente, por não conterem em seus títulos "cientista de dados" ou "data scien". Em relação às empresas de origem brasileira (Catho, Empregos e Vagas), pode-se constatar que juntos compuseram quase 1/5 das vagas analisadas (19,58%, ou 120 anúncios).

Outra característica analisada foi quanto à localidade das vagas referentes aos anúncios coletados. A cidade de São Paulo concentrou quase 1/3 da amostra com 193 anúncios (31,48%). Na segunda posição, com 102 anúncios (16,64%), estão as vagas destinada a trabalho remoto (*home office*), seguida pela cidade do Rio de Janeiro com 51 anúncios (8,32%). Todas as demais cidades apresentaram menos de 22 anúncios, não ultrapassando 4,00% do total. Em relação aos estados brasileiros, São Paulo sozinho, com 260 anúncios, corresponde a 42,41% dos documentos analisados, impulsionado naturalmente pela demanda apresentada pela sua capital.

O salário oferecido foi outra informação analisada, porém apenas 98 anúncios (15,99%) faziam menção a esse item, dos quais 62 exibiam a mensagem "A combinar". Dessa maneira, as remunerações puderam ser analisadas em somente 36 anúncios ou 5,87% da amostra. Ainda assim, foi possível extrair alguns dados relevantes: o valor médio mensal oferecido é R$ 6.782,03 (desvio padrão de R$ 542,47), sendo o menor valor apresentado correspondendo à faixa salarial "De R$ 1.001,00 a R$ 2.000,00" e o maior, R$ 15.000,00. Considerando a cotação do Real[1], a moeda brasileira vale 0,20 dólares americanos e implica em uma média salarial mensal de U$ 1,356.40 ou um valor anual de cerca de U$ 16,276.87. Esse valor equivalente a 1/6 da média do salário anual de um cientista de dados iniciante nos EUA em 2019 (Burtch Works, 2021).

Em relação à experiência exigida, analisou-se os títulos das vagas procurando pelos termos Júnior (ou Jr), Pleno e Sênior. Dos 613 anúncios, 184 (30,02%) mencionavam em seu título ao menos um dos termos procurados. O nível sênior foi o mais procurado com 103 ocorrências (16,80%), seguido pelos profissionais plenos (70 anúncios, 11,42%) e, por fim, pelos profissionais do nível Júnior (11 anúncios, 1,79%). Dessa forma, segundo os anúncios que apresentam essa informação, há uma maior demanda por profissionais mais experientes.

Para concluir esta etapa, foram identificados os anúncios que citavam algum nível de escolaridade. Primeiramente, verificou-se que 209 anúncios, ou 34,09% do *corpus*, continham a expressão "graduação" ou "ensino superior". Se por um lado, 1/3 dos anúncios apontam que os contratantes valorizam a educação superior, por outro, expõe que 2/3 não expressam a educação formal como item fundamental à seleção de candidatos. Já os cursos *strictu sensu* são menos exigidos, sendo que mestrado aparece em 53 anúncios (8,65%) e doutorado, em 38 anúncios (6,20%). Esses valores são muito inferiores aos apresentados por Kim e Lee (2016), que apontam que de 1.240 anúncios relativos a cientista de dados, 645 (52,01%) citam o nível de mestre e 561 (45,20%), o nível de doutoramento como diferencial.

4.2. Termos mais frequentes

Como primeiro resultado da mineração de texto, são apresentados os resultados obtidos pela análise n-grama aplicada aos *stems* gerados. Os 10 termos mais frequentes 1-grama e 2-grama são listados em Tabela 1 e Tabela 2, respectivamente.

[1] Segundo cotação do Banco Central do Brasil, realizada em 03 de março de 2022, o valor comercial do dólar dos Estados Unidos da América é R$ 5,05 (https://www.bcb.gov.br/estabilidadefinanceira/historicocotacoes).

Termo	F	N	%
dad	2911	567	92,50%
model	1447	478	77,98%
conhec	1441	499	81,40%
experienc	1372	481	78,47%
desenvolv	1127	451	73,57%
analis	986	435	70,96%
negoci	938	379	61,83%
estatis	823	425	69,33%
are	783	406	66,23%
learning	703	364	59,38%

Tabela 1 – 1-grama

Termo	F	N	%
machin learning	580	348	56,77%
cient dad	275	195	31,81%
analis dad	272	196	31,97%
cienc dad	249	162	26,43%
model estatis	218	176	28,71%
banc dad	193	156	25,45%
desenvolv model	162	130	21,21%
dat scienc	160	105	17,13%
big dat	158	126	20,55%
cienc computaca	153	142	23,16%

Tabela 2 – 2-grama

A classificação 1-grama, formada pelos *stems*, confirmou "dado" como o temos mais recorrente nos anúncios, com 2.911 ocorrências e presente em 567 documentos analisados (92,50%). Em seguida, com menos da metade de ocorrência, vem o *stem* "model" com 1.447 aparições em 478 documentos distintos. Já "conhec", referente a conhecimento, está presente em mais anúncios (499), mas em menor frequência (1.441). De maneira geral, o *ranking* 1-grama enfatiza dado, modelo, conhecimento, experiência, desenvolvimento e análise como as palavras presentes em mais de 70% dos documentos analisados. Além disso, percebe-se entre as 20 palavras mais recorrentes, termos relativos ao contexto organizacional como "empresa", "negócios", "pessoas", "time" e "clientes". Nos anúncios, esses termos são adotados para descrever a própria empresa contratante, mas também para tratar do cotidiano do futuro contratado.

A lista com as termos 2-grama demonstra a expressão "machine learning" como a mais recorrente entre os anúncios, aparecendo 580 vezes em 348 documentos, ou seja, 56,77% do *corpus*. Em seguida, os *stems* de "cientista de dados" com 275 ocorrências em 195 anúncios (31,81%) e de "análise de dados", 272 ocorrência, 196 anúncios (31,97%). As outras expressões presentes em pelo menos 1/5 dos anúncios são: "ciência de dados", "modelo estatístico", "banco de dados", "desenvolvimento de modelo", "big data" e "ciência da computação". A adoção de expressão aumenta o valor semântico dos itens da classificação, especificando as competências mais recorrentes, como o desenvolvimento de modelos estatísticos, administração de banco de dados e formação em ciências da computação. Nesse sentido, outra expressão presente em mais de 20% dos anúncios é "seguro de vida", um dos benefícios mais recorrentes nos anúncios analisados e "superior completo". A Tabela 3 apresenta as principais expressões formadas por três *stems*.

Termo	F	N	%
model machin learning	97	77	12,56%
grand volum dad	65	53	8,65%
ensin superi complet	60	59	9,62%

Termo	F	N	%
machin learning deep	57	51	8,32%
learning deep learning	55	50	8,16%
banc dad relac	52	48	7,83%
estatis machin learning	50	48	7,83%
cienc computaca engenh	47	45	7,34%
lingu programaca python	47	40	6,53%
desenvolv model estatis	45	43	7,01%

Tabela 3 – 3-grama

Novamente, o termo mais frequente está relacionado à aprendizagem de máquina. A expressão referente a "modelos de *machine learning*" foi a única presente em mais de 10% do *corpus*, aparecendo em 77 anúncios. Ademais, ressalta-se que "machine learning" está presente em oito expressões 3-grama dentre as 20 mais recorrentes. A segunda expressão mais frequente foi "grande volume de dados", referente à "matéria-prima" do cientista de dados, seguido por "ensino superior completo", que ocorre em 9,62% dos anúncios. Mais uma vez, a educação formal está presente entre os termos mais recorrentes, porém em uma patamar muito abaixo do que demonstrado em outras pesquisas (Curty & Serafim, 2016; Kim & Lee, 2016).

4.3. Modelagem de tópicos

Inicialmente, procedeu-se pela verificação de coerência, medida utilizada para avaliar os modelos gerados, para definir a quantidade de tópicos a serem extraídos. Para isto, fez o cruzamento de possibilidades de número de tópicos (2, 3, 4, 5, 6, 10, 15, 20) com três de variações do valor de alfa (0,01, 0,1 e 1), que é um parâmetro que influencia na definição de tópicos para cada documento. O resultado desse cruzamento pode ser visto na Figura 2:

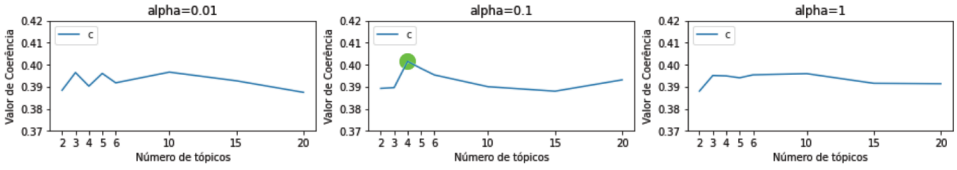

Figura 2 – Verificação dos valores de coerência para diferentes modelos

Os valores de coerência obtidos variaram entre 0,387 e 0,401, sendo este maior valor encontrado para a combinação de quatro tópicos e o valor de alfa de 0,10. Uma vez que o valor de coerência pode variar de 0 a 1, onde valores próximos a 0,7 são esperados, a coerência obtida é considerada baixa, mas não impede a utilização da LDA. Assim, com a definição do número de tópicos, procedeu-se à definição do modelo cuja visualização,

desenvolvida com a biblioteca LDAVis (Sievert & Shirley, 2014). Este procedimento indicou os quatro tópicos distantes, não havendo sobreposição, sendo que todos eles, segundo a distribuição marginal, são relevantes para grande parte do *corpus*. Porém, quando se verificou os 10 principais termos que compõem cada tópico, percebe-se que os termos encontrados são praticamente os mesmos, diferenciando-se apenas pela posição, ou seja, pelo peso que o termo representa a cada tópico. Termos únicos (entre os 10 principais) acontecem apenas nos três primeiros tópicos. No primeiro tópico, os *stems* "tecn", referente a técnica, e "client", de cliente, não aparecem nas primeiras posições dos demais tópicos. Para o segundo tópico, o termo exclusivo se refere a "projeto" e para o terceiro, o termo é "empr", referente a empresa. A Figura 3 traz os 10 principais termos de cada tópico, demonstrando as variações nas posições entre cada tópico definido.

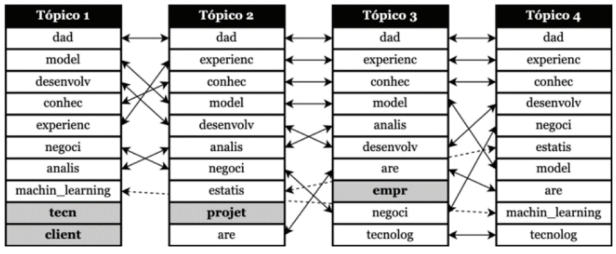

Figura 3 – Principais termos para cada tópico da LDA

A similaridade encontrada entre os tópicos definidos, que se repetiu em modelos com dois, três, cinco e seis tópicos, tornou incoerente a usual nomeação dos tópicos, vide todos "tratam" de temas muito parecidos. Resumidamente, os quatro tópicos tratam do mesmo tema, com poucas e sutis diferenças entre os pesos dos termos principais. Dessa forma, julga-se que os resultados obtidos pela LDA pouco contribuem para extrair informações da amostra analisada.

4.4.Agrupamento

Com a utilização da ferramenta *online Clustering Workbench* (Carrot² Clustering Engine, 2021), é possível aplicar algoritmos de agrupamento, dentre eles o K-Means, a partir de uma base de dados já trabalhada, de forma amigável. Os únicos parâmetros alterados na configuração padrão da ferramenta foram o número de *clusters*, definido em quatro para seguir o mesmo número obtido pela LDA, número de rótulos exibidos em cada grupo, definido como 10, e o idioma, alterado para português. Todos os 613 documentos foram agrupados nos seguintes *clusters*:

1. Time, *Data*, **Negócio**, **Novas**, Empresa, Processos, ***Learning***, ***Machine***, **Algoritmos**, **Atuar** *(367 documentos)*
2. Time, Modelos, **Diversidade**, **Crédito**, **Plano**, **Oferecemos**, **Saúde**, **Vida**, **Auxílio**, Máquina *(93 documentos)*
3. **Projetos**, Processos, **Serviços**, **Financeiro**, **Ciência**, Análise, **Experiência**, **Aprendizado**, Habilidades, Máquina *(77 documentos)*

4. *Data*, Modelos, **Tecnologia**, Empresa, **Avançado**, Análise, **Modelagem**, **Área**, **Analytics**, Habilidades *(76 documentos)*

O primeiro *cluster*, que engloba 59,87% do total de anúncios, apresenta exclusividade nos termos "negócio", "novas" (de "novas técnicas", "novas ferramentas", "novas práticas", "novas soluções", demonstrado relação com inovação e aprendizagem, conforme apontado por Cao (2019)), "machine", "learning", "algoritmos" e "atuar". Os termos sem negritos são aqueles presentes em mais de um grupo, como "time", referente aos profissionais já presentes na organização contratante, "data", dados no idioma inglês, "empresa", "processos", dentre outros.

Os outros três agrupamentos estão equilibrados em relação à quantidade de documentos englobados, 15,17%, 12,56% e 12,40%, respectivamente. O segundo agrupamento é marcado por termos relacionados a benefícios que a empresa oferece aos seus funcionários, bem como a ênfase à diversidade. O terceiro, apresenta os termos "Financeiro" e "Ciência" como destaques de exclusividade, enquanto o quarto apresenta "Tecnologia", "Modelagem" e "Avançado", termo geralmente utilizado juntamente com "Conhecimento".

Ainda que haja termos comuns a mais de um *cluster*, percebe-se uma maior distinção entre os grupos pela utilização do algoritmo *K-Means* do que pela aplicação do LDA. Logicamente, os objetivos das duas técnicas são distintos. Conforme, apontado por Bengfort et al. (2018), *clustering* procura estabelecer grupos em uma coleção de documentos, deixando cada documento em um grupo; enquanto a modelagem de tópico, busca abstrair os principais temas desta coleção, onde um único documento pode abranger mais de um tópico. A conclusão aqui é que o agrupamento se mostra mais definido que a abstração de tópicos.

5. Conclusões

Para identificar os principais requisitos presentes nos anúncios de emprego para cientistas de dados no Brasil, esta pesquisa coletou anúncios em *websites* especializados e empregou técnicas de mineração de texto. As análises preliminares revelaram que as vagas ofertadas se concentram no estado de São Paulo, especialmente em sua capital, mas que o trabalho na modalidade remota já é a segunda opção mais frequente. Além disso, verificou-se que a divulgação da remuneração ofertada não é uma prática adotada pela maioria das empresas e a média salarial obtida está muito abaixo dos valores apresentados por outras pesquisas (Burtch Works, 2021; Kaggle, 2021). Ainda em relação às características gerais dos anúncios, a pesquisa aponta para a busca por profissionais mais experientes, de nível sênior, se considerados os anúncios que apresentaram esta informação. Todavia a exigência por educação formal, especialmente por mestrado e doutorado, ficou aquém dos resultados apresentados por Kim e Lee (2016) e corroborou com Baumeister, Barbosa e Gomes (2020), que indicam que as companhias procuram por educação formal, mas esse não é um tema central nos anúncios de emprego.

Em relação aos procedimentos de mineração de texto, as classificações de 1, 2 e 3-grama destacaram conceitos mais técnicos como machine learning (aprendizagem de máquina), modelos estatísticos, análise e ciência de dados, inteligência artificial, ciência

da computação, tecnologia, python, banco de dados, modelos preditivos, processamento de linguagem natural, dentre outros. Por outro lado, também destacou conceitos relacionados ao ambiente organizacional, como negócios, pessoas, time, e ensino superior completo. Além disso, também houve a presença de termos relacionados a características e benefícios, como home office, seguro de vida, plano de saúde e plano odontológico.

As técnicas de modelagem de tópicos (LDA) e agrupamento (K-Means), embora tenham apresentado desempenhos diferentes, reforçaram os termos chave para a descoberta de temas centrais ou classificação dos anúncios. Novamente, dados, *machine learning*, análise, experiência, conhecimento, tecnologia, modelagem estatística, entendimento de negócio, projeto, desenvolvimento (software) e busca pelo novo foram termos que mais representam os requisitos para a ciência de dados. Esse resultado vai ao encontro de pesquisas preliminares como Kim e Lee (2016), Meyer (Meyer, 2019), Halwani et al. (2021) e Gottipati et al. (2021), por exemplo. Em contrapartida, diferentemente do que outras pesquisas apontam, não foi possível identificar a frequência de habilidade interpessoais, como comunicação oral e escrita (Baumeister et al., 2020; Cao, 2019; Kim & Lee, 2016) ou conhecimentos referentes a questões sociais, éticas e legais relacionadas a privacidade e segurança de dados (Anderson et al., 2014; Curty & Serafim, 2016).

Dessa forma, julga-se que as ferramentas contribuíram para um esclarecimento acerca do que é solicitado a um candidato a cientista de dados no Brasil. Contudo, para aprimorar o desempenho das técnicas utilizadas, em especial em relação à LDA, há sugestões para pesquisas futuras. Dentre elas, a recomendação inicial seria de aumentar o tamanho do *corpus*, visto que aplicações de NLP produzem resultados mais confiáveis diante de conjuntos de dados grande maiores (Wolfram, 2017). Além disso, a incorporação de outros perfis profissionais como engenheiro de dados, analista de dados, estatístico e outros. Este procedimento possibilitaria a rotulagem das vagas, permitindo a avaliação da acurácia dos algoritmos de agrupamento e também da modelagem de tópicos, uma vez que a expectativa aponta para conjuntos diferentes de requisitos entre as vagas.

Por fim, ainda que se julgue as técnicas de mineração de texto apropriadas para o objetivo da pesquisa, sugere-se para pesquisas futuras a aplicação de metodologias mistas, como análise de conteúdo, empregadas em pesquisas preliminares (Baumeister et al., 2020; Kim & Lee, 2016; Meyer, 2019). Assim, utilizando-se os mesmos dados, os resultados poderiam ser comparados e avaliados.

Financiamento

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

Referências

Anderson, P., McGuffee, J., & Uminsky, D. (2014). Data science as an undergraduate degree. *SIGCSE 2014 - Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, (pp. 705–706). https://doi.org/10.1145/2538862.2538868

Anthony, L. (2022). *AntConc (Version 4.0.4) [Computer Software]*. https://www.laurenceanthony.net/software/antconc/

Baškarada, S., & Koronios, A. (2017). Unicorn data scientist: the rarest of breeds. *Program*, *51*(1), 65–74. https://doi.org/10.1108/PROG-07-2016-0053

Baumeister, F., Barbosa, M. W., & Gomes, R. R. (2020). What is required to be a data scientist? Analyzing job descriptions with centering resonance analysis. *International Journal of Human Capital and Information Technology Professionals*, *11*(4), 21–40. https://doi.org/10.4018/IJHCITP.2020100102

Bedregal-Alpaca, N., Aruquipa-Velazco, D., & Cornejo-Aparicio, V. (2020). Técnicas de Data Mining para extraer perfiles comportamiento académico y predecir la deserción universitaria. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (*E27)*, 592–604. https://www.proquest.com/scholarly-journals/técnicas-de-data-mining-para-extraer-perfiles/docview/2385757429/se-2

Bengfort, B., Bilbro, R., & Ojeda, T. (2018). *Applied Text analysis with Python: Enabling Language-Aware Data Products with Machine Learning*. O'Reilly Media, Inc.

Brandt, P. S. (2016). The emergence of the data science profession. In *Doctor of Philosophy in the Graduate School of Arts and Sciences*. Columbia University. https://doi.org/10.7916/D8BK1CKJ

Breternitz, V. J., Lopes, F. S., & Silva, L. A. da. (2015). Big Data/Analytics: formação e gestão de cientistas de dados. *CONTECSI - International Conference on Information Systems and Technology Management*, (pp. 1–8). http://www.contecsi.tecsi.org/index.php/contecsi/12CONTECSI/paper/view/1960

Burtch Works. (2021). *The Burtch Works Study: Salaries of Data Science & Analytics Professionals*. https://www.burtchworks.com/big-data-analyst-salary/big-data-career-tips/the-burtch-works-study/

Cao, L. (2019). Data Science: Profession and Education. *IEEE Intelligent Systems*, *34*(5), 35–44. https://doi.org/10.1109/MIS.2019.2936705

Carrot² Clustering Engine. (2021). *Clustering Workbench*. https://search.carrot2.org/#/workbench

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS*.

Chen, J., Ayala, B. R., Alsmadi, D., & Wang, G. (2018). Fundamentals of Data Science for Future Data Scientists. In S. Hawamdeh & H.-C. Chang (Eds.), *Analytics and Knowledge Management* (pp. 167–194). CRC Press.

Cunha, R. (2018). *Procuram-se cientistas de dados*. https://www.linkedin.com/pulse/procuram-se-cientistas-de-dados-rodrigo-cunha/

Curty, R. G., & Serafim, J. D. S. (2016). A formação em ciência de dados: uma análise preliminar do panorama estadunidense. *Informação & Informação*, *21*(2), 307–331. https://doi.org/10.5433/1981-8920.2016v21n2p307

Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*, *90*(10), 5. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Demchenko, Y., Belloum, A., Los, W., Wiktorski, T., Manieri, A., Brocks, H., Becker, J., Heutelbeck, D., Hemmje, M., & Brewer, S. (2016). EDISON data science framework: A foundation for building data science profession for research and industry. *Proceedings of the International Conference on Cloud Computing Technology and Science, CloudCom*, *0*(Dtw), 620–626. https://doi.org/10.1109/CloudCom.2016.0107

Demchenko, Y., Belloum, A., & Wiktorski, T. (2017). *EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF-DS)*. https://edison-project.eu/data-science-competence-framework-cf-ds/

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73. https://doi.org/10.1145/2500499

Finzer, W. (2013). The Data Science Education Dilemma. *Technology Innovations in Statistics Education*, *7*(2), 1–9. https://doi.org/10.5070/T572013891

Gajzler, M. (2010). Text and data mining techniques in aspect of knowledge acquisition for decision support system in construction industry. *Technological and Economic Development of Economy*, *16*(2), 219–232. https://doi.org/10.3846/tede.2010.14

Gottipati, S., Shim, K. J., & Sahoo, S. (2021). Glassdoor job description analytics - Analyzing data science professional roles and skills. *IEEE Global Engineering Education Conference, EDUCON*, *2021-April*(April), 1329–1336. https://doi.org/10.1109/EDUCON46332.2021.9453931

Grossi, V., Giannotti, F., Pedreschi, D., Manghi, P., Pagano, P., & Assante, M. (2021). Data science: a game changer for science and innovation. *International Journal of Data Science and Analytics*, *11*(4), 263–278. https://doi.org/10.1007/s41060-020-00240-2

Hall, P., Phan, W., & Whitson, K. (2016). *The Evolution of Analytics: Opportunities and Challenges for Machine Learning in Business*. O'Reilly Media, Inc.

Halwani, M. A., Amirkiaee, S. Y., Evangelopoulos, N., & Prybutok, V. (2021). Job qualifications study for data science and big data professions. *Information Technology & People*. https://doi.org/10.1108/ITP-04-2020-0201

Kaggle. (2021). *State of Machine Learning and Data Science 2021*. https://www.kaggle.com/kaggle-survey-2021

Kim, J. Y., & Lee, C. K. (2016). An empirical analysis of requirements for data scientists using online job postings. *International Journal of Software Engineering and Its Applications*, *10*(4), 161–172. https://doi.org/10.14257/ijseia.2016.10.4.15

Lantz, B. (2015). *Machine Learning with R* (2nd ed.). Packt Publishing.

LinkedIn. (2020). 2020 Emerging Jobs Report. In *LinkedIn*. https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/emerging-jobs-report/Emerging_Jobs_Report_U.S._FINAL.pdf

Loukides, M. (2012). *What is data science? The future belongs to the companies and people that turn data into products*. O'Reilly Media, Inc. https://doi.org/10.1201/b13101-3

Mabey, B. (2018). *pyLDAvis Documentation*.

Metelo, M., Bernardino, J., & Pedrosa, I. (2021). Avaliação de Ferramentas Open Source para Data Science usando a Metodologia OSSpal. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, *E46*, 588–607. https://www.proquest.com/docview/2647406894

Meyer, M. A. (2019). Healthcare data scientist qualifications, skills, and job focus: A content analysis of job postings. *Journal of the American Medical Informatics Association*, *26*(5), 383–391. https://doi.org/10.1093/jamia/ocy181

NIST Big Data Public Working Group. (2015). NIST Big Data Interoperability Framework: Volume 1, Definitions. https://doi.org/10.6028/NIST.SP.1500-1

Provost, F., & Fawcett, T. (2013a). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Provost, F., & Fawcett, T. (2013b). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

Raschka, S. (2016). *Python Machine Learning: Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics*. Packt Publishing.

Řehůřek, R., & Sojka, P. (2011). Gensim-python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2). https://radimrehurek.com/gensim/index.html

Reis, L. C. R., & Sá, M. I. da F. e. (2020). Big Data: Um novo campo de atuação para bibliotecários. *Prisma.Com*, *41*, 231–250. https://doi.org/10.21747/16463153/41a12

Saltz, J. S., & Grady, N. W. (2017). The ambiguity of data science team roles and the need for a data science workforce framework. In *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*, (pp. 2355–2361). https://doi.org/10.1109/BigData.2017.8258190

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, (pp. 63–70). https://doi.org/10.3115/v1/W14-3110

Stark, H., & Hawamdeh, S. (2018). Relating Big Data and Data Science to the Wider Concept of Knowledge Management. In *Analytics and Knowledge Management* (pp. 141–166). CRC Press.

Wesslen, R. (2018). *Computer-Assisted Text Analysis for Social Science: Topic Models and Beyond*. ArXivLabs. http://arxiv.org/abs/1803.11045

Wolfram, D. (2017). A pesquisa bibliométrica na era do big data: Desafios e oportunidades. *Bibliometria e Cientometria No Brasil: Infraestrutura Para Avaliação Da Pesquisa Científica Na Era Do Big Data*.