



INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS TERESINA CENTRAL
TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

ADÉLIA MARA CARVALHO DE SOUSA

MINERAÇÃO DE TEXTO APLICADA À ANÁLISE DE VAGAS DE
DESENVOLVIMENTO DE SOFTWARE

TERESINA, PIAUÍ
2025

ADÉLIA MARA CARVALHO DE SOUSA

MINERAÇÃO DE TEXTO APLICADA À ANÁLISE DE VAGAS DE DESENVOLVIMENTO
DE SOFTWARE

Trabalho de Conclusão de Curso (monografia)
apresentado como exigência parcial para obten-
ção do diploma do Curso de Análise e Desen-
volvimento de Sistemas do Instituto Federal de
Educação, Ciência e Tecnologia do Piauí, Cam-
pus Teresina Central.

Orientador: Prof. Me. Rogério da Silva

TERESINA, PIAUÍ
2025

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD

Sousa, Adélia Mara Carvalho de

S725m Mineração de texto aplicada à análise de vagas de desenvolvimento de software / Adélia
Mara Carvalho de Sousa. - 2025.
65 f.: il. color.

Trabalho de Conclusão de Curso (Tecnologia em Análise e Desenvolvimento de Sistemas) -
Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Teresina Central, 2025.
Orientador : Prof Me. Rogério da Silva .

1. Desenvolvimento de software. 2. Análise de dados. 3. Mineração de texto. I. Título.

CDD - 004.21

ADÉLIA MARA CARVALHO DE SOUSA

MINERAÇÃO DE TEXTO APLICADA À ANÁLISE DE VAGAS DE DESENVOLVIMENTO
DE SOFTWARE

Trabalho de Conclusão de Curso (monografia)
apresentado como exigência parcial para obten-
ção do diploma do Curso de Análise e Desen-
volvimento de Sistemas do Instituto Federal de
Educação, Ciência e Tecnologia do Piauí, Cam-
pus Teresina Central.

Aprovada em: 12/02/2025.

BANCA EXAMINADORA

Prof. Me. Rogério da Silva(Orientador)
Instituto Federal do Piauí (IFPI)

Prof. Me. Ely da Silva Miranda
Instituto Federal do Piauí (IFPI)

Prof. Me. Duany Dreyton Bezerra Sousa
Instituto Federal do Piauí (IFPI)

TERESINA, PIAUÍ
2025

Dedico este trabalho às mulheres que vieram antes de mim, que sonharam, desafiaram limites e abriram caminhos em um mundo repleto de obstáculos. Àquelas que superaram barreiras, transformaram sistemas e impactaram não só a tecnologia, mas a própria sociedade. Que sua força, determinação e conquistas continuem a inspirar as próximas gerações a seguir em frente, construindo um futuro mais inclusivo e inovador.

AGRADECIMENTOS

Agradeço, primeiramente, à minha mãe, Maria da Paz, por seu apoio incondicional, sua paciência, compaixão e por todo o auxílio ao longo da minha vida. Mãe, você é meu maior exemplo de força e perseverança. Sem você, eu jamais teria chegado até aqui.

Ao meu pai, Adauto (*In Memoriam*), por todos os ensinamentos e pelo exemplo de trabalho, integridade e perseverança. Embora não esteja mais fisicamente presente, seu legado permanece comigo e continuará me guiando ao longo da vida.

À minha irmã, Adália, por estar sempre ao meu lado e por me ajudar nas decisões importantes. À minha sobrinha, Maju, cuja alegria e leveza tornaram mais felizes até mesmo os dias difíceis. A presença de vocês foi essencial para que eu seguisse em frente.

À minha família em Piracuruca, por sempre me acolher com carinho e fazer desse lugar um porto seguro. Em especial, à minha avó Alzira (*In Memoriam*), ao meu avô José e à minha tia Derci (*In Memoriam*), por terem acompanhado e celebrado comigo tantas conquistas.

A todas as pessoas que, em diferentes momentos da minha vida, estenderam a mão e me ofereceram carinho, compreensão e força. Algumas fizeram parte de capítulos importantes da minha história e deixaram sua contribuição para que eu chegasse até aqui.

Ao Programa de Bolsas TIM OBMEP e ao Programa de Iniciação Científica e Mestrado (PICME), pelas oportunidades e pelo suporte financeiro que me permitiram dedicar-me aos estudos e transformar este sonho em realidade.

Ao meu orientador, professor Rogério, por suas orientações, conselhos e apoio durante a construção deste trabalho. Sua dedicação, paciência e confiança foram fundamentais para o desenvolvimento desta pesquisa e para o meu crescimento acadêmico e pessoal.

Aos membros da banca examinadora, professor Ely e professor Duany, por aceitarem participar desta etapa tão importante e pelas valiosas contribuições para o aprimoramento deste trabalho.

Aos amigos que fiz ao longo dessa caminhada, agradeço pelo apoio, companheirismo, risadas e momentos que permanecerão em minha memória. Em especial, aos meus amigos Laís e Vinicius, por estarem ao meu lado em tantas etapas dessa jornada.

Ao time do Méliuz, pela oportunidade de aprendizado e pelas experiências que contribuíram para minha formação profissional. À equipe da V360, por acreditar no meu potencial e proporcionar oportunidades de crescimento e desenvolvimento.

Por fim, agradeço a todos que, de alguma forma, fizeram parte dessa caminhada. Seja por meio de palavras de incentivo, ensinamentos, apoio ou pequenos gestos de carinho, cada contribuição teve seu significado e fez parte do caminho que me trouxe até aqui.

A todos, minha mais sincera gratidão.

RESUMO

O avanço tecnológico aumentou significativamente o volume de dados processados, transformou o mercado de trabalho e tornou essenciais habilidades como programação e análise de dados. A pandemia da COVID-19 acelerou a demanda por trabalho remoto e expôs discrepâncias entre as competências dos candidatos e as exigências das empresas. Este trabalho utiliza mineração de texto para analisar vagas de emprego para desenvolvedores de software, com o foco em identificar tendências e propor estratégias para alinhar expectativas entre empregadores e candidatos. Seguindo a metodologia CRISP-DM, o estudo abrangeu as etapas de compreensão do negócio, coleta dos dados, preparação dos dados, modelagem, avaliação e implantação. Os dados foram obtidos por APIs e *web scraping*, utilizando Python. A análise foi realizada na plataforma Google Colab, utilizando técnicas de processamento de linguagem natural (PLN), como tokenização, remoção de *stopwords* e lematização. Na modelagem, empregou-se a técnica *bag-of-words* e o algoritmo k-means para análise exploratória e segmentação de dados. Os resultados apontaram uma alta demanda por Java, PHP e Python, com destaque para bancos de dados e testes unitários. Conclui-se que estratégias de capacitação são essenciais para reduzir a diferença entre as exigências das empresas e as qualificações dos profissionais. A atualização dos currículos acadêmicos, investimentos em treinamentos e certificações podem melhorar a empregabilidade e atender às demandas do setor. Além disso, descrições mais claras de vagas e a análise contínua das tendências do mercado, com o apoio de técnicas avançadas de PLN, podem auxiliar profissionais e empresas na tomada de decisões estratégicas.

Palavras-chaves: desenvolvimento de software; análise de dados; mineração de texto.

ABSTRACT

Technological advancement has significantly increased the volume of data processed, transformed the job market, and made skills such as programming and data analysis essential. The COVID-19 pandemic accelerated the demand for remote work and exposed discrepancies between candidates' skills and employers' requirements. This study analyzes job openings for software developers through text mining, aiming to identify trends and propose strategies to align expectations between employers and candidates. Following the CRISP-DM methodology, the study went through the stages of business understanding, data collection, data preparation, modeling, evaluation, and deployment. The data was obtained through APIs and web scraping with Python, and the analysis was carried out on the Google Colab platform using natural language processing (NLP) techniques, such as tokenization, stopword removal, and lemmatization. In the modeling phase, the bag-of-words technique and the k-means algorithm were applied for exploratory analysis and data segmentation. The results indicate a high demand for Java, PHP, and Python, with a strong emphasis on database management and unit testing. It is concluded that training strategies are essential to bridge the gap between companies' demands and professionals' qualifications. Updating academic curricula, investing in training programs and certifications can enhance employability and meet industry demands. Additionally, clearer job descriptions and continuous analysis of market trends, supported by advanced NLP techniques, can assist professionals and companies in making strategic decisions.

Keywords: software development; data analysis; text mining.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de aplicação de tokenização em uma frase.	19
Figura 2 – Exemplo do processo de stemização, onde palavras derivadas de "desenvolver" são reduzidas ao radical "desenvolv" e derivados de "dado" são reduzidos a "dad".	20
Figura 3 – Etapas do CRISP-DM conforme Silva <i>et al.</i> (2007)	23
Figura 4 – <i>Web Scraping</i> : Processo de Coleta e Organização de Dados	24
Figura 5 – Etapas do CRISP-DM	29
Figura 6 – Dados da coleta dos anúncios.	31
Figura 7 – Etapas da limpeza dos anúncios.	33
Figura 8 – Diagrama dos dados analisados neste trabalho.	34
Figura 9 – Etapas da análise do texto dos dados.	35
Figura 10 – Tokenização de Texto Utilizando o Módulo Punkt do NLTK	36
Figura 11 – Remoção de Caracteres Especiais com Expressões Regulares	36
Figura 12 – Remoção de Stopwords Utilizando o Módulo NLTK	37
Figura 13 – Exemplo de Stemização Utilizando o Algoritmo RSLP	37
Figura 14 – Percentual de anúncios por plataforma analisada	41
Figura 15 – Frequência dos salários	42
Figura 16 – Distribuição geográfica dos anúncios coletados por estado no Brasil. As tonalidades representam a quantidade de anúncios, variando de cores mais claras (menor concentração) para mais escuras (maior concentração).	44
Figura 17 – Nuvem da palavras dimensionadas pela importância (TD-IDF)	46
Figura 18 – Gráficos com valores de coerência para diferentes números de tópicos e variação do parâmetro alpha	49
Figura 19 – Visualização dos Tópicos Gerados pelo Modelo de Mineração de Texto	50
Figura 20 – Mapa de distância intertrópico	50
Figura 21 – <i>Silhouette Score</i> por número de <i>clusters</i>	52
Figura 22 – Palavras e volume de documentos por <i>cluster</i>	53
Figura 23 – Distribuição dos <i>Clusters</i> Formados pelo Algoritmo K-Means (k=3) sem <i>Outliers</i>	54
Figura 24 – Percentual de Vagas por Nível de Experiência.	56

LISTA DE TABELAS

Tabela 1	– Estatísticas descritivas do salário final	42
Tabela 2	– Termos mais importantes no <i>corpus</i> (TF-IDF)	43
Tabela 3	– Resumo das Métricas de Importância de Termos	45
Tabela 4	– Resumo das Métricas de Importância de N-gramas	47
Tabela 5	– Resumo das Métricas de Importância de Trigramas	48
Tabela 6	– Valores de TF-IDF e Ocorrências de Termos Relacionados a Níveis de Experiência	55
Tabela 7	– Principais Termos Técnicos Identificados no <i>Corpus</i>	56
Tabela 8	– Resumo de N-gramas (TF-IDF, Ocorrências e Documentos)	57
Tabela 9	– Comparação entre a frequência dos principais termos dos títulos e das descrições	58

LISTA DE QUADROS

Quadro 1 – Competências Essenciais	18
Quadro 2 – Principais Técnicas de Processamento de Linguagem Natural	21
Quadro 3 – Comparativo entre Web Scraping e API	25
Quadro 4 – Ferramentas e bibliotecas utilizadas e seus recursos	26
Quadro 5 – Descrição das Colunas Extraídas dos Anúncios de Emprego	32

LISTA DE ABREVIATURAS E SIGLAS

API	<i>Application Programming Interface</i> (Interface de Programação de Aplicações)
BERT	<i>Bidirectional Encoder Representations from Transformers</i> (Representações Bidirecionais de Codificadores de Transformadores)
BoW	<i>Bag-of-Words</i> (Saco de Palavras)
DBSCAN	<i>Density-Based Spatial Clustering of Applications with Noise</i> (Agrupamento Espacial Baseado em Densidade para Aplicações com Ruído)
GPU	<i>Graphics Processing Unit</i> (Unidade de Processamento Gráfico)
GPT	<i>Generative Pre-trained Transformer</i> (Transformador Generativo Pré-Treinado)
HTML	<i>HyperText Markup Language</i> (Linguagem de Marcação de Hipertexto)
JSON	<i>JavaScript Object Notation</i> (Notação de Objetos JavaScript)
K-Means	Algoritmo de Agrupamento K-Means
LDA	<i>Latent Dirichlet Allocation</i> (Atribuição de Dirichlet Latente)
NLTK	<i>Natural Language Toolkit</i> (Kit de Ferramentas para Processamento de Linguagem Natural)
OMS	Organização Mundial da Saúde
PLN	Processamento de Linguagem Natural
RSLP	Removedor de Sufixos da Língua Portuguesa
TDD	<i>Test-Driven Development</i> (Desenvolvimento Orientado a Testes)
TF-IDF	<i>Term Frequency–Inverse Document Frequency</i> (Frequência de Termo – Frequência Inversa de Documento)
TI	Tecnologia da Informação
TPU	<i>Tensor Processing Unit</i> (Unidade de Processamento de Tensores)
XML	<i>Extensible Markup Language</i> (Linguagem de Marcação Extensível)

SUMÁRIO

1	INTRODUÇÃO	13
2	FUNDAMENTAÇÃO TEÓRICA	16
2.1	MERCADO DE TRABALHO EM TECNOLOGIA DA INFORMAÇÃO . .	16
2.2	COMPETÊNCIAS PROFISSIONAIS EM DESENVOLVIMENTO DE SOFT- WARE	17
2.3	PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)	19
2.4	MINERAÇÃO DE TEXTO	21
2.5	METODOLOGIA CRISP-DM APLICADA À ANÁLISE DE DADOS TEX- TUAIS	22
2.6	FERRAMENTAS E TECNOLOGIAS PARA COLETA DE DADOS	23
2.7	FERRAMENTAS E TECNOLOGIAS PARA ANÁLISE DE DADOS	25
2.8	TRABALHOS RELACIONADOS	27
3	METODOLOGIA	29
3.1	COMPREENSÃO DO NEGÓCIO	29
3.2	COMPREENSÃO E COLETA DOS DADOS	30
3.3	PREPARAÇÃO DOS DADOS	32
3.3.1	Limpeza dos dados	33
3.3.2	Análise das descrições dos anúncios	34
3.3.3	Análise dos títulos dos anúncios	38
3.4	MODELAGEM E AVALIAÇÃO	39
4	RESULTADOS E DISCUSSÕES	41
4.1	VISÃO GERAL DAS DESCRIÇÕES	45
4.2	MODELAGEM DE TÓPICOS	49
4.3	CLASSIFICAÇÃO	52
4.4	COMPETÊNCIAS TÉCNICAS E SENIORIDADE	55
4.5	RECOMENDAÇÕES PARA ALINHAMENTO DE COMPETÊNCIAS E DEMANDAS DO MERCADO	59
4.6	LIMITAÇÕES DA PESQUISA	60
4.7	BASE DE DADOS E CÓDIGO UTILIZADO	61
5	CONCLUSÃO OU CONSIDERAÇÕES FINAIS	62
	REFERÊNCIAS	64

1 INTRODUÇÃO

Nas últimas décadas, o desenvolvimento de computadores e inovações tecnológicas resultou em um aumento significativo na quantidade de dados armazenados e processados. Esse avanço no acesso à informação provocou transformações importantes na sociedade, incluindo a forma de viver, trabalhar e pensar (MAYER-SCHÖNBERGER; CUKIER, 2013).

Os impactos na forma de trabalhar têm afetado a natureza das atividades e funções desempenhadas pelos profissionais, alterando significativamente a dinâmica do trabalho e as competências exigidas. Atividades anteriormente manuais passaram por processos de automação, o que mudou a lógica de trabalho. Por consequência, habilidades em programação, análise de dados e gerenciamento de sistemas tornaram-se cada vez mais essenciais no mercado de trabalho (VUALA; OUTROS, 2023; SAVEANU *et al.*, 2023).

Além disso, o mercado de trabalho foi afetado de forma significativa pela pandemia de COVID-19. Até o final de 2020, as vacinas ainda não estavam disponíveis, o que levou os governos a adotarem medidas como o distanciamento social e o isolamento, conforme recomendado pela Organização Mundial da Saúde (OMS). Nesse período, muitos trabalhadores perderam seus empregos, principalmente aqueles que não podiam realizar suas atividades remotamente e os que estavam no setor informal da economia (PAULA; AL., 2023).

No cenário internacional, houve uma redução inicial nos anúncios de vagas no início da pandemia. No entanto, posteriormente, houve um aumento na demanda por posições de trabalho remoto (HENSVIK; Le Barbanchon; RATHELOT, 2021). Essas mudanças contínuas no mercado de trabalho apresentam desafios e oportunidades tanto para os candidatos a vagas quanto para os empregadores.

Um dos desafios é a incompatibilidade entre as competências exigidas pelos empregadores e as possuídas pelos candidatos, o que agrava o desemprego. Isso é mais evidente em setores que passam por rápidas mudanças tecnológicas (BRAUNERHJELM; EKLUND; KÜGLER, 2023).

Desenvolvedores que adaptam suas habilidades para atender às mudanças constantes do setor têm maior probabilidade de conseguir emprego. Por exemplo, as empresas costumam priorizar os candidatos com habilidades técnicas avançadas, refletindo uma tendência mais ampla nos mercados de trabalho que favorecem a mão de obra qualificada (HADININGRUM; MARDIANA; ROCHIMAH, 2024).

Diante desse contexto, verifica-se a necessidade de uma análise sistemática das vagas de emprego para desenvolvedores de software, com o objetivo de identificar as habilidades mais demandadas pelos empregadores e, assim, proporcionar um alinhamento mais preciso entre as expectativas das empresas e as competências dos candidatos.

Além disso, compreender o mercado de trabalho está diretamente relacionado à capacidade de alinhar as expectativas dos candidatos com as demandas reais das empresas. Os anúncios de vaga de trabalho geram um volume significativo de informações sobre vagas e qualificações,

frequentemente registrado em anúncios e dados de recrutamento. No entanto, a falta de uma abordagem sistemática e padronizada para analisar essas informações pode dificultar a obtenção de *insights* valiosos.

Para investigar essas demandas do mercado de trabalho, este estudo emprega técnicas de mineração de textos e análise de dados aplicadas a anúncios de vagas online. Esse processo permite identificar padrões e tendências nas competências técnicas e comportamentais exigidas, contribuindo para um alinhamento mais eficiente entre profissionais e o mercado de tecnologia.

A análise das vagas pode oferecer uma compreensão mais aprofundada sobre quais habilidades e qualificações são valorizadas no setor, o que permite que instituições educacionais ajustem seus programas de formação e que empresas adaptem seus processos de recrutamento, contribuindo para um mercado de trabalho mais equilibrado. Além disso, empregadores podem identificar tendências emergentes e lacunas nas competências profissionais, possibilitando a formulação de estratégias mais eficazes de desenvolvimento e retenção de talentos.

Nesse contexto, a relevância desta pesquisa está em oferecer *insights* práticos que auxiliem candidatos a vagas em Tecnologia da Informação (TI) a alinhar suas competências às demandas do mercado, instituições de ensino a ajustarem seus currículos e empregadores a entenderem melhor as tendências e necessidades de competências técnicas e comportamentais. Essa abordagem contribui para o fortalecimento do mercado de tecnologia no Brasil.

Este trabalho tem como objetivo geral desenvolver uma análise sistemática das vagas publicadas no mercado de Tecnologia da Informação (TI) para desenvolvedores de software, utilizando técnicas avançadas de mineração de texto e aprendizado de máquina. O estudo visa alinhar as expectativas do mercado com as competências dos candidatos, apoiar instituições de ensino no aprimoramento curricular e oferecer ferramentas estratégicas para empregadores no recrutamento. Além deste objetivo geral, tem-se os seguintes objetivos específicos:

- Implementar e avaliar um *pipeline* de análise de dados textuais para a extração de padrões semânticos e tendências, utilizando técnicas de Processamento de Linguagem Natural (PLN) e aprendizado de máquina, contribuindo com uma abordagem replicável e eficiente para estudos futuros;
- Identificar e classificar as competências técnicas e comportamentais mais demandadas em vagas para desenvolvedores de software, utilizando métodos quantitativos para destacar habilidades prioritárias;
- Fornecer recomendações personalizadas baseadas nos dados analisados, direcionadas a candidatos, instituições educacionais e empregadores, destacando ações específicas que promovam um alinhamento mais eficaz entre as competências ofertadas e as demandas do mercado; e
- Examinar a distribuição geográfica das oportunidades de emprego na área de desenvolvimento de software, identificando regiões com maior concentração de vagas e suas características específicas.

Este trabalho está estruturado em 6 capítulos incluindo este capítulo de introdução. O Capítulo 2 aprofunda os conceitos essenciais relacionados à mineração de dados, análise de texto, dinâmica do mercado de trabalho no setor de tecnologia e os métodos utilizados para identificar competências profissionais.

Em seguida, no Capítulo 3, são apresentados a metodologia utilizada, o processo de coleta de dados, as ferramentas e técnicas aplicadas e a aplicação da metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM). A descrição objetiva dos métodos garante a transparência e a reprodutibilidade do estudo.

Já o Capítulo 4, apresenta a análise dos dados coletados, destacando padrões e tendências identificados nos anúncios de emprego no setor de tecnologia. Primeiramente, é feita uma visão geral das descrições das vagas, seguida pela modelagem de tópicos, que permite identificar os principais temas abordados nos anúncios. Em seguida, são discutidos os critérios de classificação adotados, bem como a relação entre competências técnicas e níveis de senioridade. Além disso, são abordadas as limitações da pesquisa e, por fim, é apresentada a base de dados utilizada, juntamente com o código desenvolvido para a extração e processamento das informações.

Por fim, o Capítulo 5 sintetiza as conclusões do estudo, reconhecendo suas limitações, destacando suas contribuições e sugerindo caminhos para pesquisas futuras. Este último capítulo busca consolidar o conhecimento gerado, avaliando o impacto das descobertas e propondo novas perspectivas para o aprofundamento do tema.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os conceitos fundamentais que sustentam esta pesquisa, com foco em temas como mineração de dados, análise de texto e a dinâmica do mercado de trabalho no setor de tecnologia. Além disso, são discutidos os principais métodos descritos na literatura para identificar competências profissionais mais demandadas, juntamente com um levantamento dos estudos e trabalhos acadêmicos relevantes ao tema. Essa fundamentação teórica oferece um embasamento sólido para a análise realizada no trabalho, destacando as contribuições e lacunas identificadas nos estudos existentes.

2.1 MERCADO DE TRABALHO EM TECNOLOGIA DA INFORMAÇÃO

O avanço tecnológico e o crescimento exponencial na geração de dados impactaram diretamente o setor de Tecnologia da Informação (TI), especialmente no perfil e as exigências para desenvolvedores de *software*. Esses profissionais passaram a enfrentar desafios mais complexos, sendo exigidos lidar com ambientes integrados por múltiplas fontes de dados, plataformas colaborativas e serviços em nuvem. A automação dos processos de desenvolvimento, implantação e manutenção de sistemas tecnológicos exige que esses profissionais dominem metodologias ágeis, práticas de integração contínua e ferramentas de automação de testes, consolidando um perfil técnico cada vez mais abrangente e multidisciplinar (GALL *et al.*, 2017).

Nesse contexto, a pandemia da COVID-19 acelerou ainda mais essa evolução. A necessidade de adaptação ao trabalho remoto e às operações digitais intensificou a busca por profissionais com habilidades avançadas em tecnologias emergentes, como computação em nuvem, segurança cibernética, análise de dados e desenvolvimento de soluções escaláveis. Empresas passaram a valorizar competências relacionadas à autonomia, colaboração virtual e flexibilidade, refletindo a importância das *soft skills* no contexto digital (SAVEANU *et al.*, 2023).

Esse cenário impulsionou o interesse dos profissionais em qualificação contínua. O aumento expressivo nas buscas por cursos online, certificações e especializações evidencia a conscientização sobre a necessidade de atualização constante para acompanhar as rápidas mudanças do mercado (CRIVEANU, 2022).

Além disso, novas oportunidades surgiram por meio de plataformas digitais de recrutamento, que ampliaram a visibilidade de vagas e democratizaram o acesso a processos seletivos. A forma como as empresas apresentam seus anúncios, incluindo informações sobre benefícios, formatos de trabalho e possibilidades de crescimento, tornou-se um fator decisivo para atrair talentos qualificados (KOŠTÁLOVÁ; LYÓCSA; ŠTEFÁNIK, 2022).

Paralelamente, a competitividade global aumentou, exigindo que os desenvolvedores de software não apenas dominem ferramentas técnicas, mas também possuam competências interpessoais, como comunicação eficaz, trabalho em equipe e pensamento crítico. A fluência

em idiomas estrangeiros, especialmente o inglês, também passou a ser um diferencial para atuar em ambientes internacionais ou com equipes distribuídas globalmente.

Diante desse panorama, a análise de vagas de emprego na área de TI, por meio de técnicas de mineração de texto, é fundamental para identificar as tendências e alinhar a formação de profissionais com as demandas do mercado. Esse entendimento permite que candidatos, instituições de ensino e empregadores adotem estratégias mais assertivas para suprir as lacunas de qualificação e acompanhar as constantes transformações do setor.

2.2 COMPETÊNCIAS PROFISSIONAIS EM DESENVOLVIMENTO DE *SOFTWARE*

A incompatibilidade entre os requisitos exigidos pelo mercado de trabalho e as competências apresentadas pelos candidatos é uma dificuldade evidente, especialmente em setores que passam por rápidas transformações tecnológicas (BRAUNERHJELM; EKLUND; KÜGLER, 2023). Esse descompasso reflete a urgência de uma compreensão mais aprofundada sobre as habilidades essenciais demandadas pelas empresas, tanto técnicas quanto interpessoais.

No contexto atual, as competências técnicas são essenciais no desenvolvimento de software, abrangendo a proficiência em linguagens de programação, a familiaridade com frameworks e uma base sólida em computação, incluindo estruturas de dados e algoritmos. Além disso, a capacidade de adaptação a diferentes ambientes de codificação, a compreensão de padrões de design de software e a implementação de soluções inovadoras são habilidades indispensáveis para atender às demandas do setor.

Garantir a qualidade e eficiência dos sistemas desenvolvidos também requer a aplicação de boas práticas de engenharia de software, como a escrita de testes automatizados, a adoção de metodologias ágeis e o domínio de matemática aplicada para a otimização de algoritmos e processamento de dados. Esses conhecimentos não apenas asseguram a robustez e escalabilidade das aplicações, mas também promovem a criação de soluções mais eficientes e alinhadas às necessidades do mercado.

Em complemento a essas habilidades, as competências interpessoais tornaram-se igualmente indispensáveis. Habilidades como comunicação eficaz, trabalho em equipe, liderança e resolução de conflitos são amplamente valorizadas, pois contribuem para um ambiente organizacional colaborativo e produtivo. A capacidade de adaptação, a criatividade e a inteligência emocional são diferenciais para profissionais que atuam em cenários dinâmicos, onde mudanças tecnológicas ocorrem de forma constante. Dessa forma, as competências interpessoais complementam as habilidades técnicas, criando um perfil profissional mais completo e preparado para atender às exigências do mercado.

O Quadro 1, adaptado de Assyne, Ghanbari & Pulkkinen (2022), sintetiza as competências essenciais, dividindo-as em técnicas e interpessoais. No grupo das técnicas, destacam-se a análise de requisitos, o design de *software*, a programação, a engenharia de qualidade e a manutenção. No campo das interpessoais, sobressaem-se as relações interpessoais, a cooperação e trabalho em equipe, a gestão de conflitos, a criatividade no ambiente de trabalho e o desenvolvimento pessoal.

Quadro 1 – Categorias de Competências e Competências Essenciais.

Competências Técnicas	
Análise de Requisitos	Promoção de novas ideias; conhecimento de negócios.
Design de <i>Software</i>	Designer experiente; conhecimento de padrões da indústria; conhecimento matemático em ciências fundamentais.
Programação	Competências de codificação em várias plataformas; habilidades de melhores práticas e padrões.
Testes e Engenharia de Qualidade	Escrever código de teste.
Manutenção	Habilidades em eletrônica e mecânica.
Competências Interpessoais	
Relações Interpessoais	Comunicar com o mundo exterior; ser sociável.
Cooperação e Trabalho em Equipe	Transferência de conhecimento; ver o quadro geral; habilidades de liderança.
Lidar e Resolver Conflitos	Humildade; consciência do cliente; compreender as necessidades do cliente.
Desenvolvimento no Ambiente de Trabalho	Destemido; criativo e corajoso; pensar fora da caixa.
Desenvolvimento Pessoal	Aplicar teorias no desenvolvimento de aplicações; ver oportunidade em sistemas; ter iniciativa; saber separar trabalho e estar disponível; ser autossuficiente.

Fonte: Adaptado de Assyne, Ghanbari & Pulkkinen (2022)

Essa combinação de competências técnicas e interpessoais é vital para que os profissionais estejam preparados para atender às exigências de um mercado de trabalho cada vez mais digitalizado e competitivo. Portanto, tanto instituições de ensino quanto empresas devem alinhar seus programas de formação e desenvolvimento para suprir essa lacuna, promovendo a integração entre conhecimento técnico e habilidades comportamentais.

Diante das transformações constantes no setor de TI, a compreensão das competências mais requisitadas pelo mercado tornou-se um desafio. Para lidar com essa complexidade e volume de informações, métodos automatizados de análise de dados textuais têm se mostrado ferramentas eficazes na extração de padrões e tendências. Entre essas abordagens, o Processamento de Linguagem Natural (PLN) se destaca por sua capacidade de estruturar informações extraídas de textos de linguagem humana.

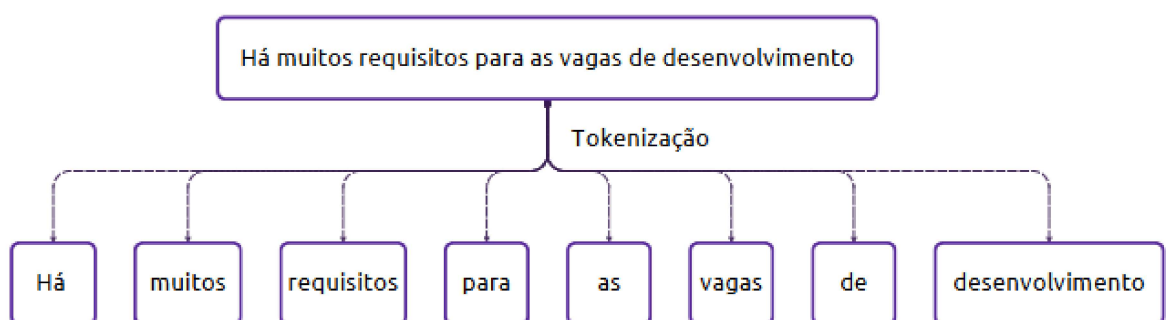
2.3 PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

Para Jurafsky & Martin (2021), o Processamento de Linguagem Natural (PLN) é uma subárea da inteligência artificial que busca desenvolver métodos para que máquinas compreendam, interpretem e gerem a linguagem humana. Essa tecnologia é essencial para transformar grandes volumes de dados textuais em informações estruturadas e acionáveis, possibilitando análises mais profundas e contextualizadas em diversos domínios, como assistentes virtuais, análise de sentimentos e tradução automática.

Ainda para Jurafsky & Martin (2021), o PLN integra linguística computacional com modelos de aprendizado de máquina, criando ferramentas eficientes para a interpretação de textos. Esse campo tem evoluído significativamente, impulsionado pela crescente disponibilidade de dados e avanços em algoritmos de aprendizado profundo. Entre as técnicas fundamentais de PLN, destacam-se Tokenização, Remoção de *Stopwords*, *Stemming* e Lematização, *Word Embeddings* e Modelos de *Deep Learning*.

A tokenização consiste em dividir textos em unidades menores, chamadas *tokens*, que geralmente correspondem a palavras individuais, mas também podem incluir expressões, sílabas ou caracteres, dependendo da abordagem adotada. Esse processo é essencial no Processamento de Linguagem Natural (PLN), pois possibilita a segmentação estruturada dos textos, facilitando tarefas como extração de informações, modelagem estatística e aprendizado de máquina (AMAZON WEB SERVICES, 2024; GUIMARÃES; JÚNIOR; FREITAS, 2022; ALMEIDA, 2020). A Figura 1 ilustra um exemplo do processo de tokenização, no qual uma frase é fragmentada em palavras individuais (*tokens*).

Figura 1 – Exemplo de aplicação de tokenização em uma frase.

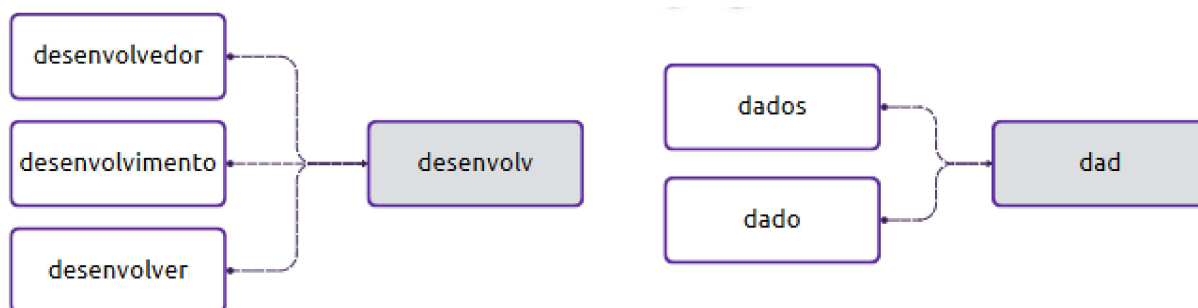


Fonte: Elaborado pelo autor (2025).

Outra técnica utilizada é a remoção de *stopwords* consiste em eliminar palavras de alta frequência e baixo valor semântico, como artigos e preposições, com o objetivo de reduzir o ruído nos dados textuais (ALMEIDA, 2020; GUIMARÃES; JÚNIOR; FREITAS, 2022). Palavras como "com", "para", "em" e "de" são exemplos típicos de termos considerados irrelevantes em análises de texto.

Além dessas técnicas, pode-se aplicar a stemização para reduzir os tokens às suas formas radicais, garantindo maior padronização no processamento textual. Por exemplo, as palavras "dados" e "dado" são convertidas para o radical "dad" (ALMEIDA, 2020). Esse processo é ilustrado na Figura 2, demonstrando como diferentes variações de uma palavra podem ser reduzidas a uma única forma base. No exemplo, "desenvolvedor", "desenvolvimento" e "desenvolver" são reduzidos ao radical "desenvolv", enquanto "dados" e "dado" são reduzidos a "dad".

Figura 2 – Exemplo do processo de stemização, onde palavras derivadas de "desenvolver" são reduzidas ao radical "desenvolv" e derivados de "dado" são reduzidos a "dad".



Fonte: Elaborado pelo autor (2025).

Para implementar essas técnicas, a biblioteca *Natural Language Toolkit* (NLTK) disponibiliza diversos algoritmos de *stemming*, incluindo o clássico *Porter Stemmer*, um dos primeiros e mais simples métodos de redução de palavras. Além disso, o NLTK oferece alternativas como o *Snowball Stemmer* (*Porter2*) e o *Lancaster Stemmer*, sendo este último mais rápido, porém mais agressivo na remoção de sufixos (RASCHKA, 2015).

Embora a stemização possa gerar formas não reais de palavras, como "dad" derivado de "dados", sua simplicidade e eficiência fazem dela uma abordagem amplamente utilizada em diversas aplicações. Em contrapartida, a lematização identifica a forma canônica e gramaticalmente correta das palavras, conhecida como *lemma*, mas exige algoritmos mais complexos e computacionalmente custosos. Apesar dessas diferenças, o impacto da stemização em tarefas de classificação de texto é frequentemente comparável ao da lematização, o que reforça sua viabilidade prática em diversas aplicações (RASCHKA, 2015; TOMAN; TESAR; JEZEK, 2006).

Destacam-se, entre as técnicas de PLN, as representações vetoriais densas de palavras, conhecidas como *word embeddings*. Modelos de representação vetorial de palavras, como o Word2Vec, são capazes de transformar termos linguísticos em vetores numéricos posicionados em um espaço multidimensional. Nesse espaço, palavras que compartilham significados ou contextos de uso similares tendem a estar mais próximas entre si, refletindo relações semânticas e sintáticas de maneira eficiente. Esse tipo de abordagem é fundamental para diversas aplicações em processamento de linguagem natural, pois permite que algoritmos interpretem e processem o significado das palavras com base em suas associações contextuais (MIKOLOV *et al.*, 2013).

Com o avanço das tecnologias de aprendizado profundo, surgiram modelos mais sofisticados

dos que transformaram a forma como a linguagem natural é processada. Arquiteturas avançadas, como o BERT e o GPT, ambos baseados no mecanismo de Transformers, revolucionaram o PLN ao capturar nuances contextuais e semânticas com alta precisão (DEVLIN *et al.*, 2019). Esses modelos são capazes de compreender o significado de uma palavra considerando o contexto em que ela aparece, o que resulta em melhorias significativas em diversas tarefas, como tradução automática, geração de texto e resposta a perguntas. O Quadro 2 resume essas técnicas, fornecendo uma visão geral das principais abordagens utilizadas no campo.

Quadro 2 – Principais Técnicas de Processamento de Linguagem Natural

Técnica	Descrição
Tokenização	Segmentação de texto em palavras ou sentenças para análise.
Remoção de <i>Stopwords</i>	Exclusão de palavras comuns e pouco informativas.
<i>Stemming</i>	Redução de palavras às suas raízes, sem considerar a forma gramatical.
Lematização	Redução de palavras à sua forma canônica (<i>lemma</i>).
<i>Word Embeddings</i>	Representação vetorial de palavras com significado semântico.
Modelos de <i>Deep Learning</i>	Arquiteturas avançadas (BERT, GPT) que capturam contexto e semântica.

Fonte: Elaborado pelo autor (2025).

Dessa forma, o PLN organiza e padroniza os textos para torná-los analisáveis por algoritmos, garantindo uma estrutura mais adequada para processamento. No entanto, a simples organização textual não é suficiente para extrair conhecimento relevante a partir de grandes volumes de dados. Nesse contexto, a mineração de texto desempenha um papel fundamental, que permite identificar padrões e tendências a partir dos dados estruturados. A combinação dessas abordagens possibilita um mapeamento mais preciso das competências exigidas no mercado de TI, bem como a análise de sua evolução ao longo do tempo.

2.4 MINERAÇÃO DE TEXTO

A mineração de texto é um processo intensivo em conhecimento que permite extrair informações relevantes de grandes coleções de documentos por meio de ferramentas de análise. Seu principal objetivo é identificar padrões em dados textuais não estruturados, utilizando técnicas como pré-processamento, algoritmos de descoberta de padrões e ferramentas de visualização. (FELDMAN; SANGER, 2007).

Diferentemente da mineração de dados, que lida com dados estruturados em bancos de dados relacionais, a mineração de texto trabalha com informações extraídas de fontes não estruturadas. São exemplos: artigos científicos, comentários em redes sociais, descrições de vagas de emprego e publicações online (FELDMAN; SANGER, 2007).

As técnicas de mineração de texto podem ser agrupadas em dois grandes enfoques: Análise de Conteúdo Manual e Análise Automatizada de Texto. A análise de conteúdo manual

envolve a leitura e categorização de textos por especialistas, sendo útil para volumes menores de dados ou análises qualitativas. Contudo, esse método é limitado pela subjetividade e pelo tempo de execução.

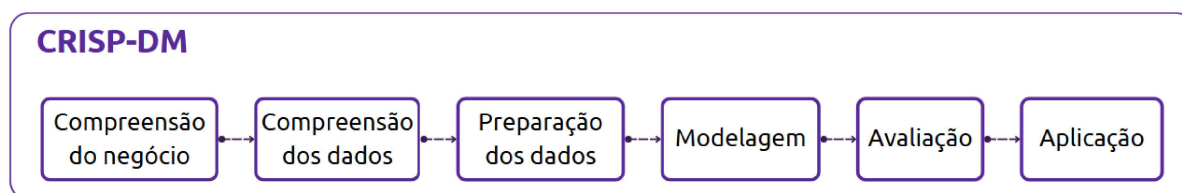
Por outro lado, a análise automatizada de texto utiliza algoritmos e modelos computacionais para processar grandes quantidades de dados de forma rápida e consistente. Esse tipo de análise é mais adequado para grandes bases de dados, como anúncios de emprego online e redes sociais. Dentre as principais técnicas automatizadas, destacam-se:

- *Bag-of-Words* (BoW): converte textos em vetores de frequência de palavras, desconsiderando a semântica e a ordem das palavras. Nesse modelo, um vocabulário é construído a partir dos documentos, e cada termo é representado pela frequência com que aparece em cada documento, gerando vetores esparsos. Embora simples, é eficaz para tarefas básicas de classificação de texto. (RASCHKA, 2015);
- TF-IDF (*Term Frequency-Inverse Document Frequency*): aprimora o BoW ao ponderar a frequência das palavras em relação à sua ocorrência em vários documentos, ajudando a identificar termos mais relevantes (RASCHKA, 2015);
- *Latent Dirichlet Allocation* (LDA): método probabilístico que identifica tópicos latentes em grandes volumes de texto, permitindo compreender a semântica dos documentos analisados (BLEI; NG; JORDAN, 2003);
- K-Means: algoritmo de agrupamento que organiza dados textuais em *clusters* com base em similaridades, sendo amplamente utilizado para segmentação de documentos (RASCHKA, 2015); e
- Além dessas técnicas, avanços recentes em Processamento de Linguagem Natural (PLN) e aprendizado profundo (*deep learning*), como o uso de modelos baseados em redes neurais (*Transformers*), revolucionaram a análise de grandes volumes de texto. Modelos como o BERT (*Bidirectional Encoder Representations from Transformers*) e o GPT (*Generative Pre-trained Transformer*) conseguem capturar o contexto semântico com maior precisão, sendo ferramentas promissoras para análise de textos complexos.

2.5 METODOLOGIA CRISP-DM APLICADA À ANÁLISE DE DADOS TEXTUAIS

A mineração de texto e o PLN têm ganhado destaque como estratégias para extrair conhecimento de grandes volumes de dados textuais. Uma das abordagens consolidadas para estruturar projetos dessa natureza é o CRISP-DM, amplamente reconhecido por sua aplicabilidade em diferentes setores e pela organização sistemática de suas etapas. Segundo GUIMARÃES, JÚNIOR & FREITAS (2022), o CRISP-DM pode ser compreendido como um conjunto de fases flexíveis, adaptável a diferentes necessidades e contextos de análise de dados.

Figura 3 – Etapas do CRISP-DM conforme Silva *et al.* (2007)



Fonte: Elaborado pelo autor (2025).

No trabalho de Silva *et al.* (2007), são descritas seis etapas principais nesse modelo: (1) Compreensão do Negócio; (2) Compreensão dos Dados; (3) Preparação dos Dados; (4) Modelagem; (5) Avaliação; e (6) Aplicação. A Figura 3 ilustra essas fases e ressalta a interdependência entre elas. A seguir, apresenta-se uma descrição detalhada de cada etapa, conforme proposta por Silva *et al.* (2007).

As fases do CRISP-DM são essenciais para a condução estruturada de projetos de mineração de dados. A primeira fase, chamada Compreensão do Negócio, tem como objetivo alinhar as metas organizacionais a um plano de ação estruturado para a resolução do problema em estudo. Em seguida, a fase de Compreensão dos Dados envolve a identificação, coleta e análise preliminar das informações relevantes, avaliando sua qualidade, consistência e utilidade.

Em seguida, na fase de Preparação dos Dados, são realizadas atividades de pré-processamento, como limpeza, transformação e organização dos dados, garantindo que estejam prontos para a etapa de modelagem. A fase de Modelagem consiste na seleção e aplicação de técnicas de mineração de dados e algoritmos de aprendizado de máquina mais adequados ao contexto da pesquisa, além da implementação e execução dos modelos.

Posteriormente, na fase de Avaliação, os modelos desenvolvidos são analisados para verificar se os resultados obtidos estão alinhados com os objetivos iniciais, realizando ajustes quando necessário para melhorar o desempenho. Por fim, a fase de Aplicação integra os resultados ao processo decisório, gerando valor estratégico para a organização.

2.6 FERRAMENTAS E TECNOLOGIAS PARA COLETA DE DADOS

A análise de dados só se torna viável após a coleta eficiente das informações relevantes. Assim, a etapa de coleta de dados é fundamental para o sucesso de projetos de análise e mineração de dados (DUBEY; KUMAR, 2014). Existem diferentes abordagens para realizar essa coleta, sendo as mais comuns o *Web Scraping* e o uso de APIs (*Application Programming Interface*). A escolha entre esses métodos depende de fatores como a natureza dos dados, a complexidade da extração, e as restrições legais e éticas envolvidas.

O *Web Scraping* consiste na extração automatizada de dados de páginas da web sem o uso de APIs oficiais, por meio de programas que fazem requisições a servidores e processam os conteúdos retornados, como arquivos HTML, XML e JSON (LAWSON, 2015). Essa técnica

envolve a realização de requisições HTTP, geralmente utilizando métodos como GET ou POST, para acessar o conteúdo de sites e coletar informações relevantes Zhao (2017).

Figura 4 – *Web Scraping*: Processo de Coleta e Organização de Dados



Fonte: Elaborado pelo autor (2025).

Conforme ilustrado na Figura 4, o processo de *Web Scraping* pode ser dividido em duas etapas principais: Coleta de Dados e Extração e Organização dos dados. Inicialmente, ocorre a coleta. A etapa de coleta envolve o envio de requisições HTTP para capturar informações de sites-alvo. Ferramentas como Urllib2 e Selenium são amplamente utilizadas para automatizar esse processo.

Após a coleta, os dados são processados e organizados por meio de bibliotecas de *parsing*, como BeautifulSoup e Pyquery, que permitem navegar e extrair informações estruturadas de documentos HTML e XML. Essa técnica é amplamente aplicada em diversos contextos, como por exemplo:

- Comparação de Produtos: Monitoramento de preços e características de produtos em plataformas de comércio eletrônico (KHATTER *et al.*, 2022); e
- Análise de Sentimentos: Coleta de notícias e publicações em redes sociais para análise de sentimentos, como no estudo de Kaur (2022), em que dados extraídos por *Web Scraping* foram usados para identificar sentimentos expressos em notícias.

Contudo, o uso de *Web Scraping* levanta questões éticas e legais importantes. O envio de requisições em excesso pode ser interpretado como um ataque de negação de serviço (DoS), comprometendo a integridade de servidores (HIRSCHEY, 2014). Além disso, a extração de dados pode violar direitos autorais e os termos de serviço de sites. Para mitigar esses riscos, recomenda-se:

- Evitar a coleta de dados protegidos por direitos autorais ou que envolvam informações confidenciais;
- Garantir que o objetivo do *Web Scraping* não conflita com os interesses da página alvo;
- Manter a frequência de requisições moderada para não sobrecarregar servidores Krotov, Johnson & Silva (2020)

Por outro lado, as APIs oferecem uma alternativa mais estruturada e segura para a coleta de dados. APIs são conjuntos de protocolos e ferramentas que permitem a comunicação entre

sistemas e o acesso controlado a dados. Muitas plataformas, como LinkedIn, Twitter e Indeed, disponibilizam APIs oficiais para acesso a informações de forma legal e segura.

Ao contrário do *Web Scraping*, as APIs fornecem dados em formatos padronizados (como JSON e XML), permitindo a extração eficiente e organizada de grandes volumes de dados. APIs são projetadas para fornecer acesso direto e estruturado aos dados, reduzindo o risco de bloqueio ou mudanças no formato das páginas, o que garante maior estabilidade na coleta dos dados. Além disso, a extração de dados por APIs é mais rápida e precisa, já que os dados são disponibilizados de forma estruturada.

Embora as APIs ofereçam uma forma estruturada, segura e eficiente para a coleta de dados, elas também apresentam algumas limitações que devem ser consideradas. Nem sempre as APIs estão disponíveis para todas as plataformas ou oferecem acesso irrestrito às informações. Muitas APIs impõem restrições de uso, como limites de requisições por período, exigência de autenticação ou cobrança por acesso a determinados recursos. Além disso, nem todos os dados públicos exibidos em sites estão acessíveis por meio de APIs, o que pode limitar a profundidade da análise.

Quadro 3 – Comparativo entre Web Scraping e API

Critério	Web Scraping	API
Acesso a Dados	Amplo, porém sujeito a restrições legais	Limitado às permissões da API
Estrutura dos Dados	Não estruturado (HTML, XML)	Estruturado (JSON, XML)
Velocidade de Coleta	Variável, depende do site	Alta, com resposta rápida
Facilidade de Implementação	Requer parsing e adaptação constante	Mais simples e documentada
Risco Legal	Alto, sujeito a violações de termos	Baixo, geralmente regulamentado
Limites de Requisições	Não definido, risco de bloqueio	Definido, com limites por período

Fonte: Elaborado pelo autor (2025).

O Quadro 3 apresenta uma visão geral das principais características, vantagens e limitações das técnicas de coleta a partir de *Web Scraping* e de APIs. A escolha adequada do método de coleta de dados depende dos objetivos da pesquisa, da disponibilidade de dados e das restrições legais.

2.7 FERRAMENTAS E TECNOLOGIAS PARA ANÁLISE DE DADOS

De acordo com Raschka (2015), avanço da Ciência de Dados e do Processamento de Linguagem Natural impulsionou o desenvolvimento de bibliotecas e plataformas robustas, capazes de lidar com grandes volumes de dados de maneira eficiente. A linguagem Python tem se destacado nesse cenário por oferecer uma ampla variedade de ferramentas especializadas que

facilitam a manipulação, organização e análise de dados. Bibliotecas como NumPy e Pandas são amplamente utilizadas para o tratamento de dados estruturados, proporcionando alta performance na realização de cálculos complexos e na manipulação de grandes conjuntos de dados.

Para análises mais avançadas e processamento de grandes volumes de informações, ferramentas baseadas em computação em nuvem, como o Google Colab, têm desempenhado papel fundamental. Essa plataforma permite a execução de tarefas de alto desempenho, oferecendo recursos de hardware avançados, como Unidades de Processamento Gráfico (GPUs) e Unidade de Processamento de Tensores (TPUs), sem exigir investimentos iniciais elevados em infraestrutura computacional (KARAGÖL; KARAGÖL, 2024). Essa acessibilidade torna possível a realização de experimentos complexos, especialmente em projetos de mineração de textos e análise de dados.

Quadro 4 – Ferramentas e bibliotecas utilizadas e seus recursos

Ferramenta/Biblioteca	Recurso/Funcionalidade
NumPy	Realização de cálculos numéricos eficientes e manipulação de arrays multidimensionais.
Pandas	Manipulação, organização e análise de grandes conjuntos de dados estruturados.
Google Colab	Execução de tarefas de alto desempenho com acesso a GPUs/TPUs em nuvem.
NLTK	Realização de tokenização, stemming, lematização, análise sintática e de sentimentos.
Punkt Tokenizer	Tokenização precisa de sentenças e palavras baseada em modelos estatísticos.
Stopwords (NLTK)	Remoção de palavras irrelevantes para melhorar a análise textual.
SLP (Removedor de sufixos da língua portuguesa)	Stemização para reduzir palavras à forma base, otimizando a análise semântica.
Scikit-learn	Implementação de algoritmos de aprendizado de máquina, como BoW, TF-IDF e K-Means.
Gensim	Modelagem de tópicos e representações vetoriais, como LDA e Word2Vec.
LDavis	Visualização interativa de modelos LDA para interpretação de tópicos.
TensorFlow	Desenvolvimento de modelos avançados de aprendizado profundo, como Transformers.
PyTorch	Criação e treinamento de modelos de aprendizado profundo com alta performance.

Fonte: Elaborado pelo autor (2025).

No contexto da análise de dados textuais, o pré-processamento é uma etapa crucial para garantir a qualidade dos dados antes da aplicação de algoritmos de mineração e aprendizado de máquina. O Quadro 4 mostra um compilado das ferramentas disponíveis para o processamento e modelagem dos dados. Diversas ferramentas são essenciais nesse processo, destacando-se a

biblioteca NLTK, o Punkt Tokenizer, as listas de *Stopwords* e o SLP (Removedor de Sufixos da Língua Portuguesa).

A NLTK é uma das bibliotecas mais completas e utilizadas para tarefas de PLN em Python. Ela disponibiliza um conjunto abrangente de ferramentas para a tokenização de textos, análise sintática, classificação de conteúdo, *stemming*, lematização e análise de sentimentos, sendo altamente modular e extensível (BIRD; KLEIN; LOPER, 2009).

A versatilidade dessa biblioteca facilita o desenvolvimento de pipelines de pré-processamento, essenciais para a extração de informações significativas em projetos de análise textual. Dentro da NLTK, o *Punkt Tokenizer* se destaca como uma ferramenta eficaz para a segmentação de textos em sentenças e palavras. Baseado em modelos estatísticos, ele identifica limites de frases de maneira adaptativa, sendo especialmente eficiente em textos informais e com ambiguidades linguísticas.

Complementando esse processo, as listas de *stopwords* da NLTK são fundamentais para a remoção de palavras de baixo valor semântico, como preposições e conjunções, tornando o processamento mais enxuto e eficiente (SILVA; LOPES, 2013). Para tratar as particularidades da língua portuguesa, o RSLP (Removedor de Sufixos da Língua Portuguesa) desempenha um papel relevante. Sua principal função é realizar a stemização (*stemming*), reduzindo palavras a suas formas base ou raízes. Esse processo é crucial para uniformizar diferentes flexões de uma palavra e melhorar a análise semântica.

Além dessas bibliotecas, a Scikit-learn desempenha papel central na aplicação de algoritmos de aprendizado de máquina, sendo amplamente utilizada em técnicas como *Bag-of-Words* (BoW), TF-IDF e algoritmos de agrupamento como o K-Means (RASCHKA, 2015). Para modelagem de tópicos e representações vetoriais de textos, a biblioteca Gensim se destaca pela implementação de modelos como *Latent Dirichlet Allocation* (LDA) e Word2Vec, que são fundamentais para a análise semântica (ŘEHŮŘEK; SOJKA, 2011).

A integração dessas ferramentas em fluxos de análise de dados textuais possibilita a construção de modelos robustos e precisos. O pré-processamento adequado, combinado com técnicas avançadas de mineração de dados e aprendizado de máquina, potencializa a geração de insights significativos e a produção de resultados mais consistentes e relevantes.

2.8 TRABALHOS RELACIONADOS

No Brasil, estudos como o de GUIMARÃES, JÚNIOR & FREITAS (2022) demonstram a eficácia da aplicação de métodos combinados para a análise de requisitos em ciência de dados. Técnicas como o BoW, o algoritmo de agrupamento K-Means e o modelo probabilístico LDA têm se mostrado relevantes para a identificação de padrões em descrições de vagas de emprego. Essas abordagens permitem mapear as habilidades técnicas e comportamentais mais demandadas pelo mercado, contribuindo para uma compreensão mais aprofundada das exigências do setor de tecnologia.

Em âmbito internacional, Baumeister, Barbosa & Gomes (2020) explorou a aplicação

da técnica *Centering Resonance Analysis* (CRA) na análise de dados textuais de anúncios de emprego. A CRA avalia a coesão textual ao mapear conexões entre palavras e conceitos, proporcionando uma análise semântica mais profunda. Diferentemente de métodos tradicionais, como o BoW, que ignoram a ordem e o contexto das palavras, a CRA é capaz de capturar nuances semânticas mais complexas, tornando-se uma ferramenta poderosa para a análise de textos sofisticados.

A integração dessas técnicas de análise textual — tanto aquelas baseadas em frequência de termos (BoW, TF-IDF) quanto as que exploram padrões semânticos mais complexos (LDA, CRA) — permite extrair insights estratégicos sobre as competências mais exigidas, tendências emergentes e lacunas de qualificação no mercado de trabalho. Essa abordagem integrada oferece suporte para a formulação de estratégias mais eficazes de recrutamento e desenvolvimento profissional, promovendo um alinhamento mais preciso entre as demandas do mercado e as competências dos candidatos.

Além disso, o estudo de Montandon *et al.* (2021) analisou mais de 20.000 anúncios de emprego publicados no Stack Overflow Jobs para identificar as principais *hard skills* e *soft skills* exigidas pelo mercado de tecnologia. Aplicando técnicas de PLN e classificação automática, os autores extraíram e categorizaram habilidades a partir das descrições das vagas. Os resultados apontaram que linguagens de programação são o principal requisito técnico, com destaque para Python, JavaScript e Java, enquanto comunicação, colaboração e resolução de problemas figuram entre as competências comportamentais mais valorizadas pelas empresas. Além disso, o estudo identificou que frameworks e bibliotecas, como React, Angular e Django, também são altamente demandados, especialmente para cargos de desenvolvimento full stack e frontend. Esse estudo reforça a relevância da combinação de métodos estatísticos e aprendizado de máquina para a extração de padrões significativos em grandes volumes de dados textuais.

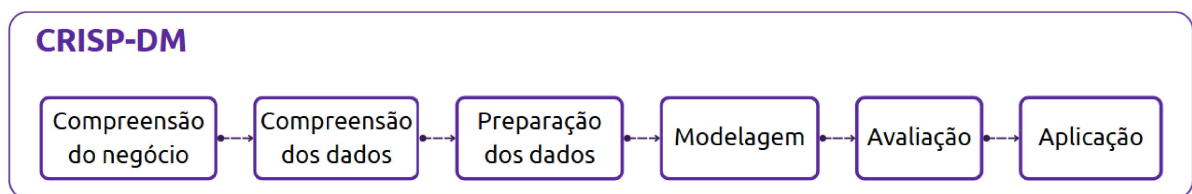
Ao considerar esses estudos, observa-se que a combinação de métodos tradicionais e avançados de análise de texto amplia significativamente a capacidade de compreender o mercado de trabalho. Essa perspectiva fundamenta a escolha metodológica deste estudo, que busca explorar técnicas de PLN para identificar padrões e tendências em descrições de vagas de emprego na área de desenvolvimento de *software*.

A seguir, a Seção 3 apresenta a metodologia aplicada neste estudo, incluindo a preparação dos dados. Essa etapa envolve técnicas de Processamento de Linguagem Natural (PLN) e algoritmos de modelagem, como LDA e K-Means, utilizados para extrair padrões relevantes e garantir a reprodutibilidade da investigação.

3 METODOLOGIA

A metodologia deste estudo baseou-se no modelo CRISP-DM, adaptado ao contexto acadêmico. Foram coletados anúncios de emprego em plataformas online, seguidos de pré-processamento para organização e padronização dos dados. Técnicas de processamento de linguagem natural (PLN) e algoritmos de modelagem, como LDA e K-Means, foram aplicados para identificar padrões e tópicos relevantes.

Figura 5 – Etapas do CRISP-DM



Fonte: Elaborado pelo autor (2025).

As etapas do CRISP-DM, ilustradas na 5, serão exploradas detalhadamente neste capítulo, estruturando a metodologia adotada para a análise dos dados textuais. Cada fase será descrita em profundidade, abordando os procedimentos aplicados para coleta, processamento e modelagem dos dados.

3.1 COMPREENSÃO DO NEGÓCIO

No contexto desta pesquisa, as etapas do CRISP-DM foram adaptadas para atender às especificidades do problema investigado. Considerando que o produto final deste estudo não é necessariamente uma implementação prática, mas sim um modelo conceitual ou análise de dados, as fases do CRISP-DM foram ajustadas de acordo com os objetivos acadêmicos.

Conforme discutido por Bedregal-Alpaca, Aruquipa-Velazco & Cornejo-Aparicio (2020), adaptações nas etapas são comuns em estudos científicos, onde o foco está na construção de conhecimento. Assim, as seis fases do CRISP-DM foram utilizadas como guia metodológico, com ajustes que garantem a adequação ao desenvolvimento da solução proposta neste trabalho.

No contexto deste projeto, as necessidades a serem atendidas não são empresariais, mas científicas, relacionadas diretamente à pesquisa em andamento. Durante esta etapa, foram realizadas ações fundamentais para estabelecer as bases do estudo, tais como:

- Fundamentação do tema, com a consolidação de conceitos e teorias relevantes que sustentam a pesquisa;
- Entendimento de trabalhos anteriores na área, com o objetivo de identificar lacunas, metodologias e abordagens aplicadas por outros autores; e

- Definição dos objetivos do estudo, através da definição do problemas a serem resolvidos e os resultados esperados.

Após a definição das bases conceituais do estudo, a próxima etapa envolve a aplicação do modelo CRISP-DM na estruturação do processo de análise de dados. A fase Compreensão dos dados Dados, é essencial para garantir a qualidade e a representatividade das informações utilizadas na pesquisa. Dessa forma, a seguir, são detalhados os procedimentos adotados para a identificação, seleção e organização das fontes de dados.

3.2 COMPREENSÃO E COLETA DOS DADOS

As etapas de coleta, preparação e análise de dados foram organizadas conforme a metodologia CRISP-DM, adaptada conforme sugerido por GUIMARÃES, JÚNIOR & FREITAS (2022) e Silva *et al.* (2007), garantindo rigor e reprodutibilidade ao processo aplicado. Neste trabalho, os dados utilizados tiveram origem em anúncios de vagas online. Para isso, foram catalogados e analisados anúncios publicados em *websites* especializados em divulgação de vagas de emprego.

Para a seleção das plataformas, foi realizada uma pesquisa no Google utilizando os termos "plataforma de vagas" e "melhores sites de vagas de emprego". As plataformas mais bem ranqueadas e frequentemente citadas nos resultados foram escolhidas para a coleta: LinkedIn, Indeed e Gupy.

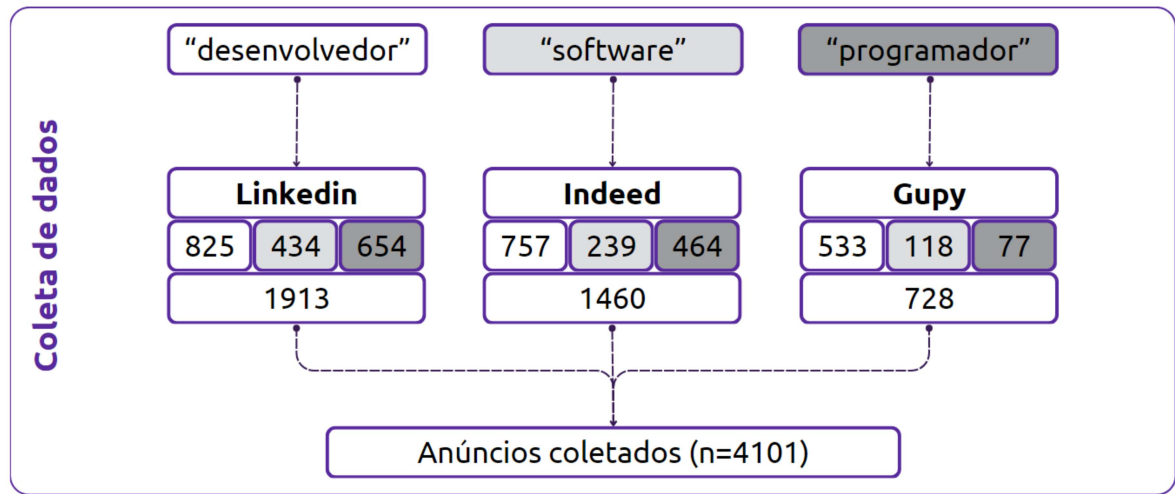
A coleta dos dados foi realizada por meio de uma busca automatizada em plataformas específicas de divulgação de vagas de emprego. A busca considerou três critérios principais: tempo de publicação, localização geográfica e palavras-chave relacionadas aos termos de interesse. Esses critérios foram definidos para assegurar que as vagas coletadas fossem recentes e pertinentes ao escopo deste trabalho.

Além disso, foi identificado que plataformas como a Gupy removem as vagas da plataforma assim que são encerradas. Nesse contexto, a utilização de intervalos de tempo mais longos para a coleta de dados poderia resultar na captura apenas das vagas ainda ativas, comprometendo a representatividade da amostra. Para evitar esse problema, optou-se por realizar a coleta de vagas recentes e de forma frequente, garantindo maior abrangência e atualidade dos dados analisados.

Para coletar dados, realizou-se uma análise inicial para identificar a disponibilidade de interfaces de programação de aplicativos (APIs) nas plataformas selecionadas. Apenas a plataforma Gupy ofereceu uma API que permitiu o acesso direto aos dados das vagas. As demais plataformas, LinkedIn e Indeed, não disponibilizavam APIs para consulta de vagas. Diante disso, optou-se pela utilização de técnicas de raspagem de dados (web scraping) para coletar as informações necessárias dessas plataformas.

A Figura 6 apresenta a distribuição dos anúncios coletados a partir dos três principais portais utilizados no estudo: LinkedIn, Indeed e Gupy. As palavras-chave “desenvolvedor”, “software” e “programador” foram empregadas para estruturar a busca automatizada, permitindo uma segmentação eficiente das vagas disponíveis em cada plataforma.

Figura 6 – Dados da coleta dos anúncios.



Fonte: Adaptado de (GUIMARÃES; JÚNIOR; FREITAS, 2022).

Os dados utilizados neste trabalho foram coletados no período de 13/12/2024 a 21/12/2024, com um total de 4101 anúncios coletados. No estudo realizado por GUIMARÃES, JÚNIOR & FREITAS (2022), foi coletada uma amostra de 1.343 anúncios, com a recomendação de que futuros estudos ampliassem o volume de vagas analisadas para aprimorar a aplicação do modelo proposto. Seguindo essa diretriz, este trabalho expande a base de dados utilizada, garantindo uma amostra mais abrangente.

Durante a fase de entendimento de dados, foram identificadas as principais informações extraídas dos anúncios de emprego. Essas informações foram selecionadas com base em sua relevância para a análise e organizadas de forma padronizada. Os dados mais importantes incluem o título da vaga, a descrição completa, a remuneração, a localização e o tipo de regime de trabalho.

Além disso, foram coletados dados complementares, como o nome da empresa contratante, as datas de publicação e criação do anúncio, bem como informações auxiliares, como o endereço URL externo da vaga, o site de origem, o termo de pesquisa utilizado e o link direto para o anúncio.

Após a análise preliminar, constatou-se que os websites analisados disponibilizavam os dados de forma heterogênea, variando tanto na estrutura quanto na nomenclatura dos campos. Diante disso, foi realizada uma avaliação detalhada para identificar as informações disponíveis em cada portal de vagas e estabelecer equivalências entre os diferentes formatos de dados. Posteriormente, aplicou-se um processo de padronização, uniformizando os campos nos dataframes a fim de garantir a consistência e integridade da estrutura dos dados. O Quadro 5 apresenta a descrição detalhada de cada atributo coletado após esse processo, destacando os principais elementos analisados no estudo.

Quadro 5 – Descrição das Colunas Extraídas dos Anúncios de Emprego

Nome do Campo	Significado
title	Título da vaga ou cargo oferecido.
description	Descrição completa das responsabilidades e requisitos da vaga.
salary	Informação sobre a remuneração oferecida para a vaga.
location	Localização geográfica onde a vaga está disponível.
work_type	Tipo de regime de trabalho (ex.: presencial, remoto, híbrido, meio período).
company_name	Nome da empresa contratante.
published_date	Data de publicação do anúncio da vaga.
created_at	Data de criação do anúncio no sistema.
external_job_url	Endereço URL externo do anúncio da vaga.
origin_site_name	Nome do site de origem de onde o anúncio foi extraído.
search_term	Termo de pesquisa utilizado para encontrar a vaga.
job_url	Link direto para o anúncio da vaga.

Fonte: Elaborado pelo autor (2025).

A estruturação dos atributos apresentados no Quadro 5 permitiu a extração de informações relevantes sobre as vagas analisadas. Os campos *title* e *description* foram utilizados na aplicação de técnicas de PLN e Mineração de Texto, com o objetivo de identificar padrões nas descrições das oportunidades de emprego. Além disso, os atributos *location* e *work_type* possibilitaram a segmentação geográfica das vagas e a classificação das modalidades de trabalho. O campo *salary* foi empregado para análise das faixas salariais disponibilizadas, enquanto as variáveis *company*, *created_at* e *published_date* permitiram examinar a frequência de publicação das vagas e a atuação das empresas no processo de recrutamento.

Além desses atributos principais, outros campos foram utilizados para segmentação das vagas por plataforma, como no caso do atributo *origin_site_name*, e como fonte de consulta para informações detalhadas caso haja necessidade futura, a exemplo do *job_url*. A próxima seção apresenta a metodologia empregada na preparação dos dados, descrevendo o papel desses atributos na construção das análises realizadas.

3.3 PREPARAÇÃO DOS DADOS

A preparação dos dados ocorreu em duas etapas principais: limpeza e transformação. Essas etapas foram fundamentais para converter os dados em formatos que os algoritmos pudessem interpretar e, consequentemente, melhorar o desempenho dos modelos. A combinação de diversas técnicas de pré-processamento foi aplicada para eliminar informações irrelevantes para a classificação textual, visando alcançar melhores resultados (ALMEIDA, 2020).

Para a realizar a análise de dados proposta neste trabalho, utilizou-se o Google Colab, pois a ferramenta fornece acesso gratuito a GPUs e TPUs, o que viabiliza a execução de tarefas de computação de alto desempenho sem a necessidade de alto custo operacional (KARAGÖL;

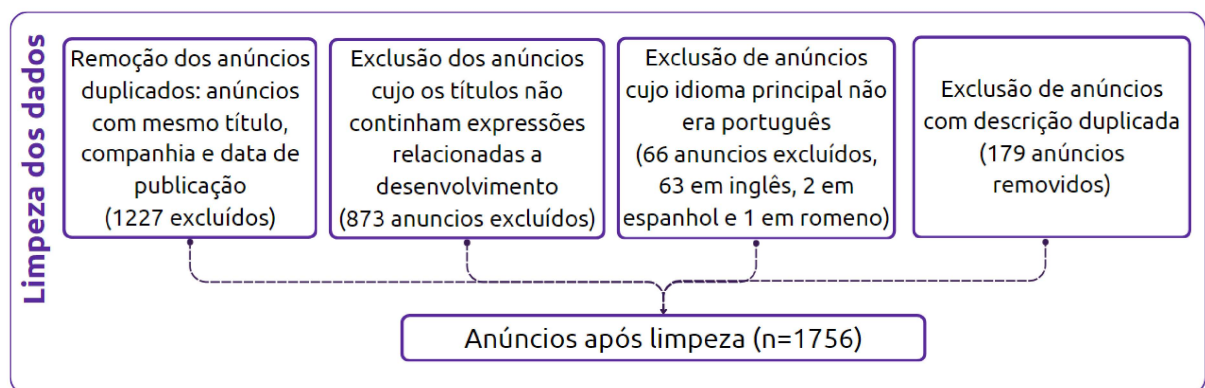
KARAGÖL, 2024). Todas as etapas de processamento foram implementadas em Python, abrangendo desde a coleta e pré-processamento de dados até as análises exploratórias e de modelagem.

3.3.1 Limpeza dos dados

Durante o pré-processamento, destacaram-se as bibliotecas Pandas e NumPy, utilizadas para manipulação, limpeza e organização dos dados coletados. Essas bibliotecas permitiram lidar com grandes volumes de informações, viabilizando a preparação do conjunto de dados para as etapas posteriores de modelagem e análise. A adoção de Python e de tais bibliotecas foi fundamental para garantir eficiência e escalabilidade ao longo de todo o processo metodológico.

Após a análise preliminar dos dados coletados na fase de compreensão dos dados, verificou-se que alguns dados foram selecionados de maneira incorreta. Para evitar que os resultados fossem comprometidos por um conjunto de dados inconsistente, foi aplicado um processo de limpeza de dados, que consistiu na filtragem de anúncios capturados de forma inadequada. Esse processo envolveu as seguintes etapas, ilustradas na Figura 7:

Figura 7 – Etapas da limpeza dos anúncios.



Fonte: Adaptado de (GUIMARÃES; JÚNIOR; FREITAS, 2022).

Inicialmente, foi feita a remoção dos anúncios duplicados. Consideraram-se duplicados os anúncios com título, nome da companhia e data de publicação idênticos. Como resultado, 1.227 anúncios foram excluídos. Posteriormente, anúncios com descrições duplicadas também foram removidos, uma vez que a publicação de um mesmo anúncio em diferentes plataformas, com variações no nome da empresa, afeta diretamente a eficácia do filtro de duplicações aplicado na etapa anterior. Esse filtro considerava apenas o nome da companhia, o que tornou necessária uma verificação adicional pelo conteúdo da descrição. Assim, 305 anúncios foram excluídos do conjunto de dados.

Em seguida, foram removidos os anúncios cujos títulos não continham expressões relacionadas à área de desenvolvimento. Os termos utilizados para a filtragem foram selecionados com base em cargos e funções, áreas de atuação e tecnologias, conforme listado abaixo:

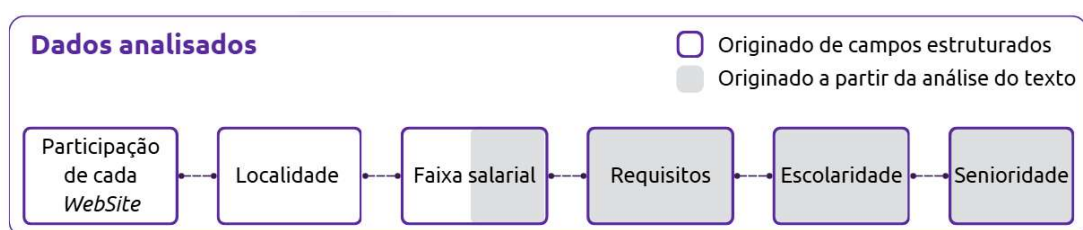
- Cargos e Funções: Desenvolvedor, Programador, Analista de Desenvolvimento, Dev, Developer, Engineer.
- Áreas de Atuação: Backend, Frontend, Back-end, Front-end, Back end, Front end.
- Tecnologias e Conceitos: Dados, *Software*, Computação, Informática, Data, Cloud, Tech.

Por fim, foram excluídos os anúncios cujo idioma principal da descrição não era o português, uma vez que a mineração de texto foi realizada com base nas informações contidas nesse campo. Anúncios em outros idiomas poderiam comprometer o desempenho das técnicas de processamento de linguagem natural aplicadas. Ao todo, 66 anúncios foram removidos: 63 em inglês, 2 em espanhol e 1 em romeno. Para a identificação do idioma foi utilizada a biblioteca *Langdetect*.

3.3.2 Análise das descrições dos anúncios

O estudo iniciou-se com a análise das principais características estruturadas e não estruturadas dos anúncios de emprego que compuseram o *corpus* da pesquisa. Essa etapa foi fundamental para contextualizar o perfil das vagas coletadas e compreender a distribuição das oportunidades no mercado de trabalho. Foram examinados campos estruturados dos anúncios, como a plataforma de origem, localização geográfica e salário, bem como informações não estruturadas extraídas por meio de mineração textual, como requisitos das vagas e nível de experiência exigido. A Figura 8 ilustra os dados analisados e informa se eles são originados de campos estruturados ou a partir da análise do texto.

Figura 8 – Diagrama dos dados analisados neste trabalho.



Fonte: Elaborado pelo autor (2025).

A análise dos campos estruturados revelou padrões quantitativos relevantes, possibilitando a identificação, por exemplo, a participação de cada *website* na distribuição dos anúncios, as localidades com maior concentração de vagas, os níveis de escolaridade mais demandados e as faixas salariais. Já a mineração textual se mostrou essencial para capturar nuances relevantes do mercado, como a extração de informações salariais descritas no corpo do anúncio e a identificação de requisitos específicos e senioridades dos cargos ofertados.

Inicialmente, avaliou-se a participação de cada *website* utilizado na coleta de dados, com o objetivo de identificar a distribuição dos anúncios entre as plataformas consultadas. Essa

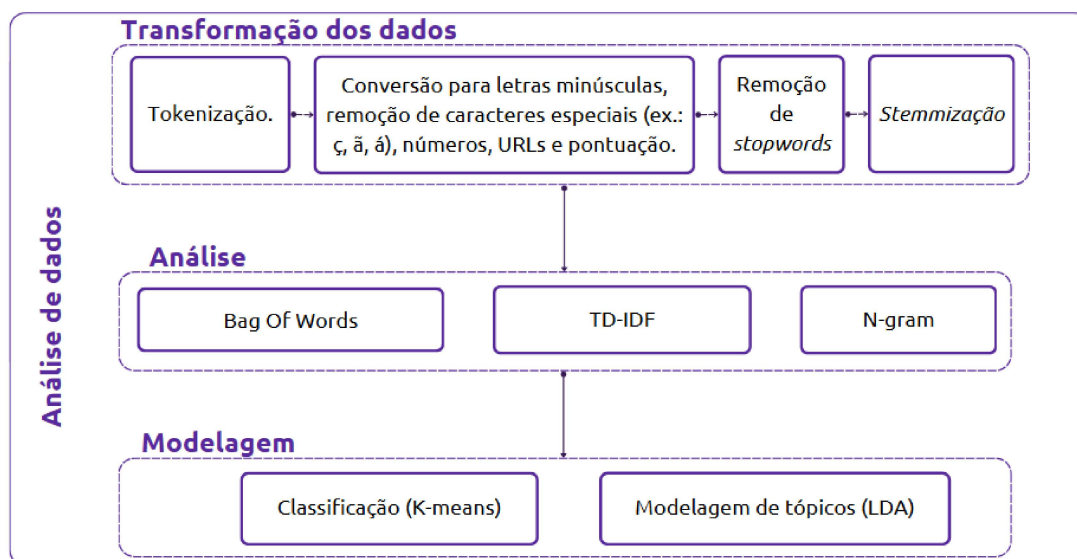
análise permitiu observar quais sites concentram a maior quantidade de vagas na área de interesse, além de revelar tendências específicas de publicação.

Em seguida, as localidades das vagas foram examinadas, o que possibilita a identificação da distribuição geográfica das oportunidades. Esse levantamento foi essencial para compreender quais regiões apresentam maior demanda por profissionais da área de tecnologia. O nível de escolaridade exigido pelas empresas também foi considerado, permitindo compreender os requisitos acadêmicos mais demandados para os cargos ofertados.

Por fim, a análise da faixa salarial, foi necessário aplicar etapas de transformação textual nas descrições, pois nem todas as vagas coletadas tinham o salário explícito como atributo, tornando-se necessário aplicar a engenharia de atributos no texto para coletar dados de salários informados na descrição das vagas. A avaliação dos salários forneceu uma visão detalhada sobre a remuneração média e a variação salarial entre diferentes cargos e níveis de experiência.

Para extrair os requisitos das vagas, também foram aplicadas etapas de transformação textual específicas. A Figura 9 ilustra esse processo e detalha as principais etapas da análise textual.

Figura 9 – Etapas da análise do texto dos dados.



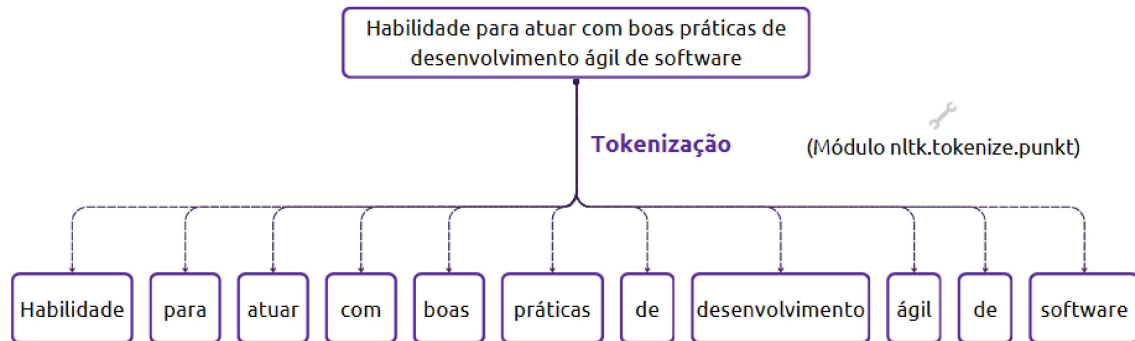
Fonte: Adaptado de (GUIMARÃES; JÚNIOR; FREITAS, 2022).

Para o processamento de linguagem natural, foi utilizada a biblioteca NLTK, amplamente reconhecida por sua robustez e facilidade de uso em tarefas de PLN. Essa ferramenta possibilitou a execução eficiente das etapas de pré-processamento de texto, fundamentais para a análise de dados deste estudo.

A primeira etapa consistiu na tokenização, implementada por meio do recurso Punkt da NLTK. Esse recurso permitiu segmentar o texto em sentenças e palavras de forma precisa, facilitando a preparação dos dados para as etapas subsequentes. A Figura 10 demonstra o resultado da aplicação desse processo na frase "Habilidade para atuar com boas práticas de desenvolvimento ágil de *software*", evidenciando como o texto foi segmentado em tokens para

melhor processamento computacional.

Figura 10 – Tokenização de Texto Utilizando o Módulo Punkt do NLTK

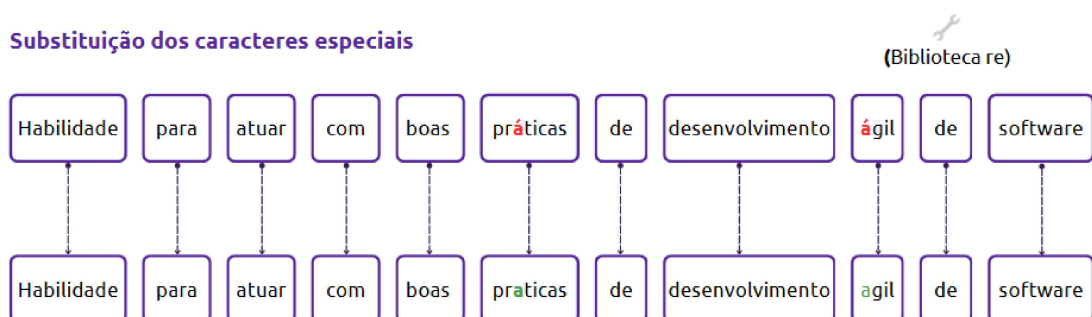


Fonte: Elaborado pelo autor (2025).

Após a tokenização, os tokens gerados passaram por um processo de normalização, no qual foram removidos caracteres especiais, como acentos (ç, ã, á), números, URLs e pontuações, que poderiam interferir nos resultados das análises textuais. Para garantir essa limpeza, foi utilizada uma expressão regular com a função `re.sub()` da biblioteca `re`. Esse procedimento eliminou pontuações, símbolos e caracteres não alfanuméricos, transformando os tokens da etapa anterior e preservando apenas palavras relevantes para a análise.

A remoção desses elementos é essencial para reduzir ruídos no conjunto de dados, garantindo que apenas informações significativas sejam consideradas nas etapas subsequentes do processamento. Observa-se na Figura 11 a remoção dos caracteres especiais das palavras "Práticas" e "ágil" foram substituídas por "praticas" e "agil", respectivamente, evidenciando a normalização dos caracteres acentuados.

Figura 11 – Remoção de Caracteres Especiais com Expressões Regulares

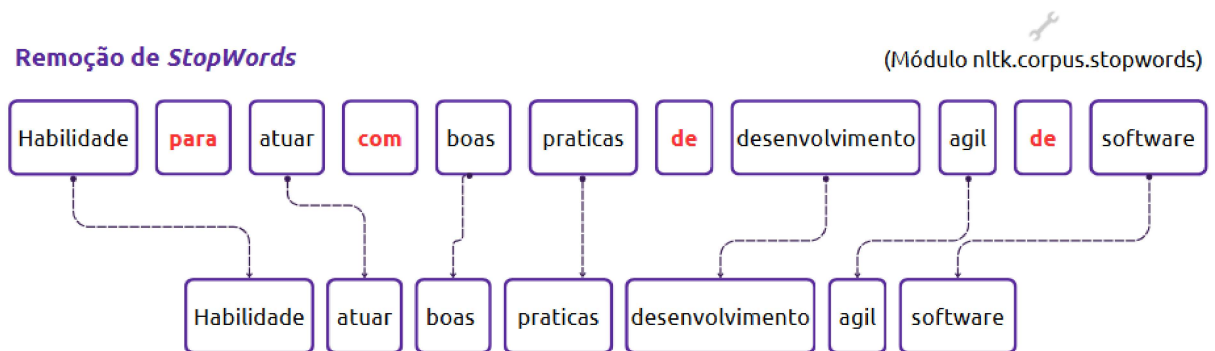


Fonte: Elaborado pelo autor (2025).

Após a normalização dos tokens, foi realizada a remoção de *stopwords*, utilizando o módulo Stopwords da biblioteca NLTK. Esse processo eliminou palavras irrelevantes para a análise, como artigos, preposições e pronomes, que não agregam valor semântico significativo ao processamento textual.

Além da remoção das *stopwords* padrão da língua portuguesa, foram eliminados termos personalizados, ajustados conforme as especificidades desta pesquisa. Os termos filtrados incluíram 'divulgavagasoficial', 'siga', 'trabalho', 'você', 'nosso', 'aqui', 'nbs' e 'insta'. A seleção dessas palavras baseou-se em sua alta frequência de ocorrência e na baixa relevância informativa para a análise. Esse processo teve como objetivo aprimorar a qualidade do conjunto de dados, reduzindo ruídos e garantindo que apenas informações essenciais fossem preservadas para as etapas subsequentes de processamento. A Figura 12 ilustra essa transformação, evidenciando a remoção de palavras funcionais e a manutenção dos termos mais relevantes para a análise."

Figura 12 – Remoção de Stopwords Utilizando o Módulo NLTK

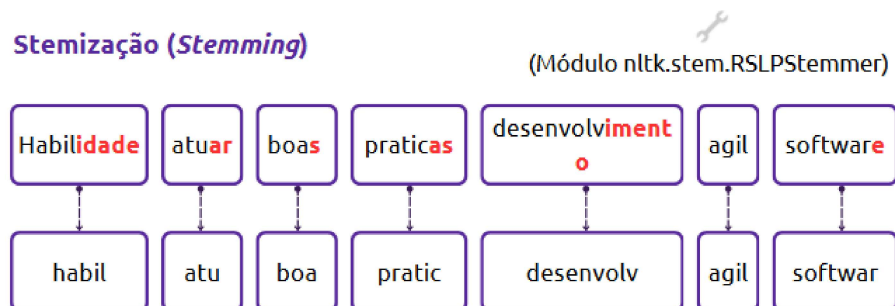


Fonte: Elaborado pelo autor (2025).

Por fim, foi aplicada a stemização (*stemming*) utilizando o algoritmo RSLP, disponibilizado pelo NLTK. Esse procedimento reduziu as palavras às suas raízes, unificando diferentes variações de um mesmo termo e padronizando o vocabulário (ALMEIDA, 2020).

Por exemplo, palavras como desenvolvedor, desenvolvimento e desenvolver foram reduzidas à forma base "desenvolv", enquanto dado e dados foram transformados em "dad". Essa técnica é essencial para reduzir a dimensionalidade do texto e melhorar a eficácia das análises linguísticas. A Figura 13 ilustra esse processo, demonstrando como as palavras foram reduzidas às suas formas radicais.

Figura 13 – Exemplo de Stemização Utilizando o Algoritmo RSLP



Fonte: Elaborado pelo autor (2025).

Os dados da etapa anterior foram convertidos em vetores numéricos utilizando a técnica de *bag-of-words* para representar as informações textuais. Esse processo de transformação é essencial para que os algoritmos de aprendizado de máquina possam lidar com dados textuais de maneira eficiente, permitindo a aplicação de métodos de classificação e análise de forma estruturada e escalável (RASCHKA, 2015). Para estruturar os dados textuais de forma adequada à aplicação dos algoritmos de análise, foi utilizada a classe `CountVectorizer` da biblioteca `scikit-learn`.

Abaixo, um exemplo de aplicação de n-gramas sobre os tokens reduzidos por stemming na frase original "Habilidade para atuar com boas práticas de desenvolvimento ágil de *software*":

- Unigramas (n=1):
 - habil, atu, boa, pratic, desenvolv, agil, softwar
- Bigramas (n=2):
 - habil atu, atu boa, boa pratic, pratic desenvolv, desenvolv agil, agil softwar
- Trigramas (n=3):
 - habil atu boa, boa pratic, boa pratic desenvolv, desenvolv agil softwar

Essas etapas garantiram que os dados textuais estivessem adequadamente preparados para o processamento e análise, otimizando a performance dos algoritmos de aprendizado de máquina e reduzindo a complexidade computacional. Além disso, a análise de unigramas, bigramas e trigramas permitiu identificar os termos mais recorrentes, incluindo habilidades exigidas para as vagas e benefícios oferecidos, fornecendo *insights* valiosos sobre os padrões presentes nos anúncios de emprego.

3.3.3 Análise dos títulos dos anúncios

Para complementar as informações obtidas, foi realizada uma análise específica dos títulos dos anúncios, que frequentemente sintetizam as informações mais relevantes sobre as vagas. Os títulos serviram como ponto de partida para compreender padrões linguísticos que associam habilidades técnicas às posições hierárquicas.

O primeiro passo envolveu a compreensão da estrutura e do conteúdo dos títulos das vagas de emprego. Essa etapa foi essencial para identificar padrões linguísticos e contextuais que pudessem revelar informações sobre as habilidades técnicas e a senioridade exigida. A partir dessa análise preliminar, foi possível perceber que os títulos das vagas frequentemente combinam a principal tecnologia ou habilidade com o nível hierárquico desejado, como nos exemplos "Desenvolvedor *Fullstack* JR" e "Engenheiro *Backend* Pleno".

Nesse contexto, foi aplicada para a análise dos títulos das vagas a mesma preparação de dados utilizada anteriormente na análise das descrições. Esse processo incluiu as seguintes etapas: tokenização, stemmização, remoção de stopwords e análise da frequência de unigramas e bigramas.

Após a preparação, foi realizada a análise de frequência dos termos extraídos. A partir dos unigramas, foi possível mapear as tecnologias, linguagens de programação e ferramentas mais mencionadas nos títulos das vagas. Já a análise de bigramas permitiu identificar combinações de termos que evidenciam a associação entre habilidades técnicas e níveis hierárquicos. Exemplos disso são expressões como "Desenvolvedor *Fullstack*" e "Engenheiro *Backend*", que demonstram a relação direta entre a tecnologia exigida e a senioridade da posição.

Esse processo analítico possibilitou compreender de forma aprofundada quais competências técnicas são mais valorizadas pelo mercado de trabalho e como elas se relacionam com diferentes níveis de senioridade, oferecendo uma visão clara das exigências das vagas de emprego analisadas.

A partir da extração e análise dos unigramas e bigramas montados com os dados do título, foram filtradas as características técnicas mais relevantes, permitindo um mapeamento das principais habilidades técnicas presentes nos títulos dos anúncios de emprego.

Tanto as descrições quanto os títulos forneceram uma visão detalhada das exigências do mercado de trabalho. Enquanto as descrições permitiram capturar nuances mais amplas, como expectativas relacionadas a experiência e remuneração, os títulos sintetizaram as competências técnicas e níveis hierárquicos em um formato mais direto. Juntas, essas análises garantem uma compreensão robusta das demandas e padrões do mercado.

3.4 MODELAGEM E AVALIAÇÃO

Embora o modelo de BoW ainda seja amplamente utilizado em tarefas de classificação de textos, ele não leva em consideração a estrutura das sentenças nem aspectos gramaticais. Como alternativa, destaca-se a *Latent Dirichlet Allocation* (LDA), um modelo probabilístico de tópicos que incorpora a análise de semântica latente das palavras (BLEI; NG; JORDAN, 2003). Por isso, foi aplicado o LDA através da biblioteca Gensim.

Utilizou-se o algoritmo K-Means para agrupar anúncios de emprego com base em características semelhantes. A implementação do K-Means foi realizada utilizando a classe KMeans da biblioteca scikit-learn, amplamente reconhecida por sua eficiência e facilidade de uso em tarefas de aprendizado de máquina. Essa ferramenta permitiu a configuração de parâmetros como o número de clusters e o número máximo de iterações, assegurando a convergência do modelo. O K-Means é extensivamente utilizado tanto no meio acadêmico quanto na indústria, destacando-se pela sua implementação simplificada e elevada eficiência computacional, especialmente quando comparado a outros algoritmos de agrupamento (RASCHKA, 2015).

A avaliação dos modelos foi conduzida com base nos resultados obtidos, os quais serão apresentados na seção de Resultados, utilizando-se de métricas específicas para cada técnica aplicada. Para o modelo LDA, será utilizada a métrica de Coerência de Tópicos (*Topic Coherence*), que avalia a qualidade semântica dos tópicos gerados ao medir a similaridade entre as palavras mais representativas de cada tópico. Valores de coerência próximos a 1 indicam maior coesão semântica entre os termos, enquanto valores mais baixos podem sugerir tópicos menos

interpretáveis (RÖDER; BOTH; HINNEBURG, 2015). A escolha dessa métrica se justifica por sua eficácia na validação da qualidade dos tópicos em modelos probabilísticos como o LDA.

Para o modelo de clusterização K-Means, a avaliação será realizada por meio da métrica de Índice de Silhouette (*Silhouette Score*). Essa métrica quantifica o grau de separação entre os clusters formados, com valores variando de -1 a 1. Valores acima de 0,5 são considerados bons, indicando clusters bem definidos e separados; valores entre 0,2 e 0,5 são aceitáveis, enquanto valores abaixo de 0,2 sugerem sobreposição entre os grupos (ROUSSEEUW, 1987)

O estudo focou na análise das descrições e títulos dos anúncios de emprego, extraindo informações valiosas para compreender as demandas do mercado de trabalho. Essa abordagem permitiu identificar características estruturadas e não estruturadas presentes nos dados coletados, como plataformas de origem, localização geográfica, salários, requisitos das vagas e níveis de experiência. Cada uma dessas dimensões foi examinada separadamente para oferecer uma visão completa e detalhada.

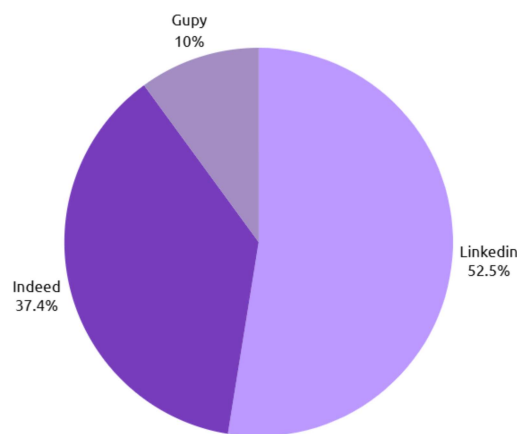
Além da análise de dados, esta pesquisa busca fornecer recomendações práticas para profissionais, instituições de ensino e empresas, com base nas competências identificadas como mais demandadas pelo mercado de desenvolvimento de *software*. Essas sugestões serão apresentadas de forma detalhada na seção de Resultados.

4 RESULTADOS E DISCUSSÕES

Após a aplicação das técnicas de mineração de texto descritas na metodologia, foram obtidos resultados que permitem identificar as principais tendências do mercado de trabalho para desenvolvedores de software. A seguir, são apresentados os dados coletados, acompanhados de uma análise crítica em relação ao referencial teórico.

A plataforma LinkedIn concentrou a maior parte dos anúncios coletados, totalizando 921 registros. Em seguida, a plataforma Indeed apresentou 656 anúncios, enquanto a Gupy contabilizou 176 anúncios. A Figura 14 demonstra o percentual dos anúncios coletados por vaga.

Figura 14 – Percentual de anúncios por plataforma analisada



Fonte: Elaborado pelo autor (2025).

Os dados do gráfico da Figura 14 destacam a predominância do LinkedIn como ferramenta de divulgação de oportunidades de trabalho, possivelmente devido à sua ampla rede de conexões profissionais e à alta popularidade no mercado corporativo. A Indeed, por ser um agregador de vagas bastante acessível, também apresentou uma representatividade significativa. A Gupy, por sua vez, teve menor participação, o que pode estar relacionado ao seu foco específico em processos seletivos automatizados para empresas parceiras.

Além disso, o alto volume de dados coletado nas plataformas digitais de divulgação de vagas, como LinkedIn, Indeed e Gupy, reflete a tendência observada por Košťálová, Lyócsa & Štefánik (2022), que destacam a importância das redes profissionais e das plataformas especializadas no recrutamento de profissionais de tecnologia.

No entanto, um dos desafios identificados foi a escassez de informações salariais nas descrições das vagas. Somente 183 anúncios (25,51% do total) traziam algum dado sobre salário. Desse total 67,21% (123 anúncios) mencionam o valor numérico do salário, enquanto os demais mencionavam a palavra "salário" mas sem informar o valor.

A Tabela 1 apresenta um resumo da análise descritiva dos salários ofertados nas oportunidades de trabalho coletadas. O salário médio registrado foi de R\$ 4.879,12, com um

desvio padrão de R\$ 3.443,80, indicando uma variação considerável nas remunerações. Os valores oscilaram entre um mínimo de R\$ 1.050,00 e um máximo de R\$ 16.500,00, evidenciando discrepâncias significativas entre as faixas salariais disponíveis.

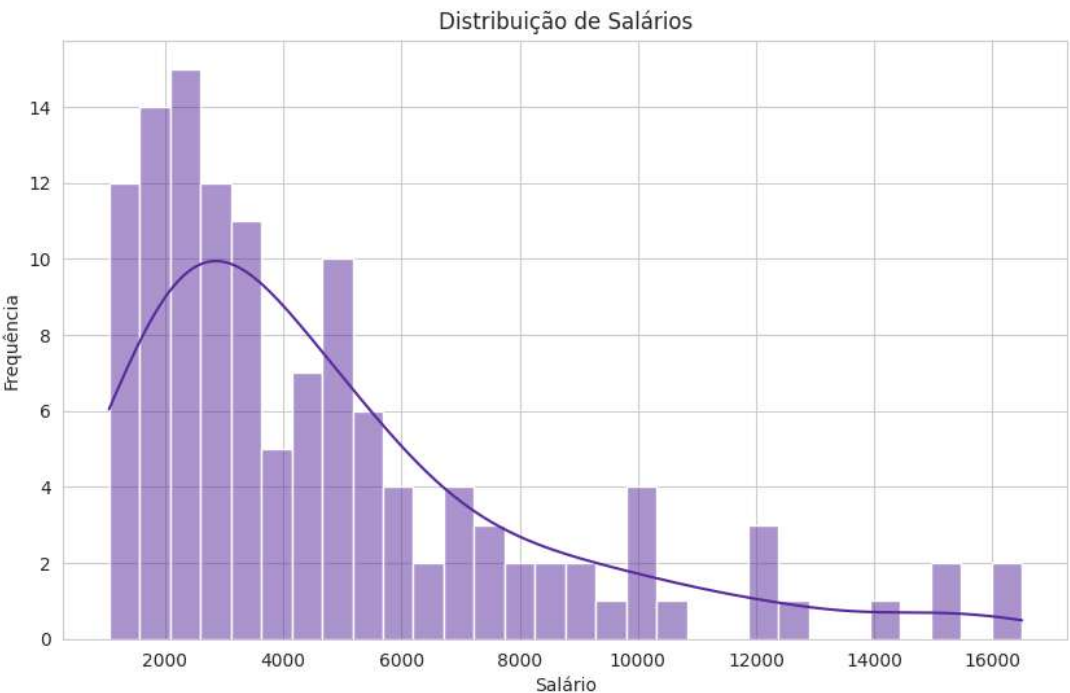
Tabela 1 – Estatísticas descritivas do salário final

Estatística	Valor (R\$)
Contagem (n)	123
Média	4.879,12
Desvio padrão	3.443,80
Mínimo	1.050,00
1º Quartil (25%)	2.500,00
Mediana (50%)	3.500,50
3º Quartil (75%)	6.000,00
Máximo	16.500,00

Fonte: Elaborado pelo autor (2024).

A mediana foi de R\$ 3.500,50, o que revela que metade das vagas oferece salários abaixo desse valor e metade acima. A análise dos quartis mostra que 25% das vagas oferecem até R\$ 2.500,00 e 75% oferecem até R\$ 6.000,00. Esses dados indicam uma distribuição assimétrica, com uma concentração de salários mais baixos e poucos casos com remuneração muito elevada, apenas 25% dos anúncios possuem salário acima de R\$ 6000,00. A distribuição da quantidade de anúncios por faixa salarial está disponível no gráfico da Figura 15 que evidencia a assimetria nos dados salariais analisados.

Figura 15 – Frequência dos salários



Fonte: Elaborado pelo autor (2025).

A Tabela 2 apresenta os resultados obtidos a partir da mineração textual, destacando os termos mais relevantes relacionados à flexibilidade e às condições de remuneração nos anúncios de emprego. A relevância dos termos foi calculada com base na métrica TF-IDF, permitindo identificar as palavras mais significativas nas descrições das vagas associadas ao tema salarial.

Tabela 2 – Termos mais importantes no *corpus* (TF-IDF)

Termo	Soma do TF-IDF	Frequência nos Documentos
combin	19,05	53
adic	10,75	21
compati	9,59	26
pag	9,21	24
experie	8,73	58
benefici	8,04	47
val	7,78	50
merc	7,54	28
ensin	7,42	37
plan	7,13	42
desenvolv	6,65	46

Fonte: Elaborado pelo autor (2025).

Conforme os dados da Tabela 2, observa-se que termos como "combin" (19,05) e "adic" (10,75) apresentaram alta frequência, indicando o uso comum de expressões como "a combinar" e "adicional". Esses resultados evidenciam a ênfase dos empregadores na flexibilidade salarial e na oferta de benefícios complementares.

Adicionalmente, outros termos intimamente associados à remuneração e às condições de trabalho também apresentaram relevância, como "pag" (9,21), "benefici" (8,04) e "plan" (7,13). Esses resultados indicam que os anúncios frequentemente incluem informações sobre benefícios oferecidos como plano de saúde, formas de pagamento e planos de carreira, com o objetivo de detalhar atrativos que vão além do salário nominal.

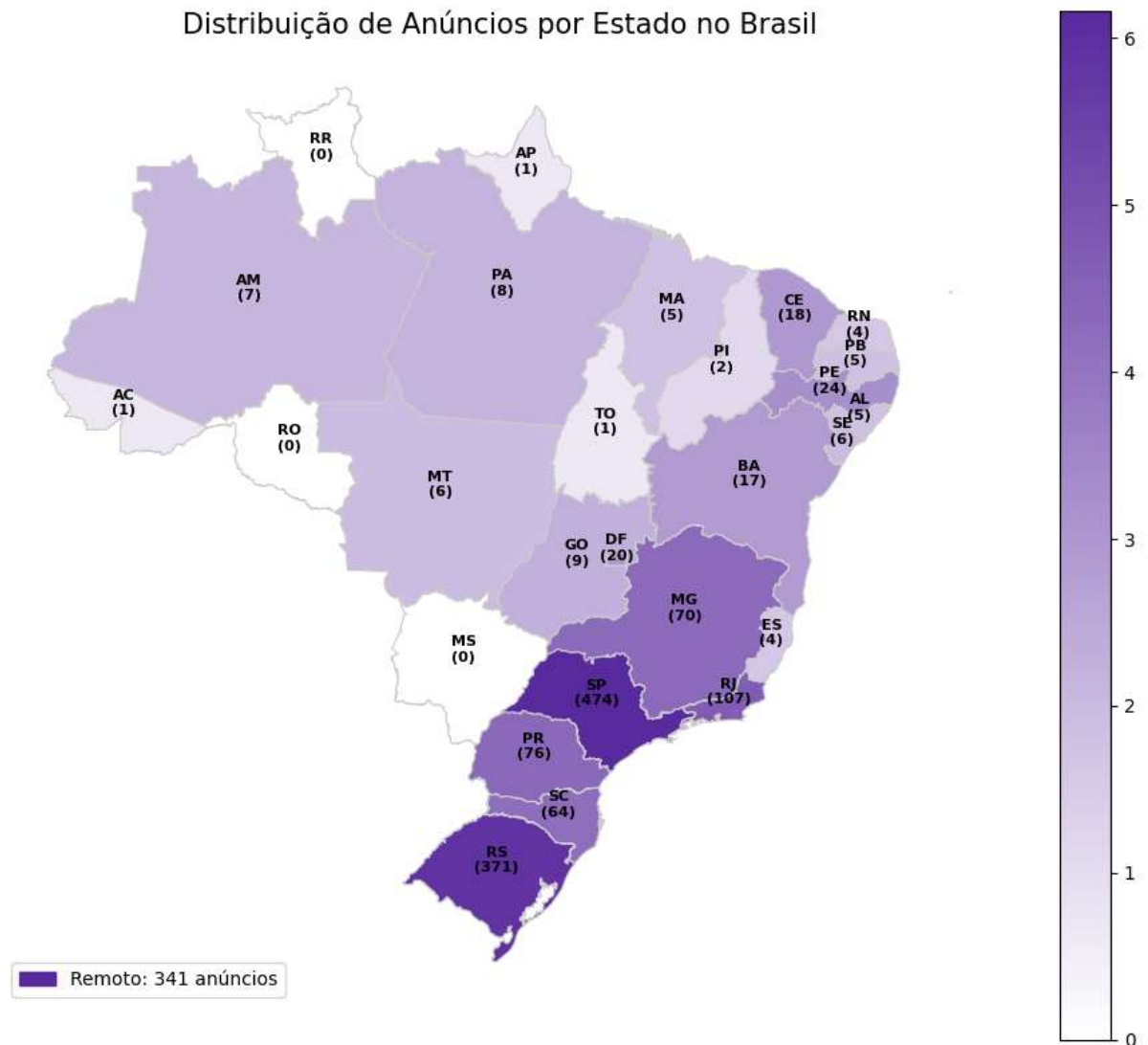
A presença dos termos "desenvolv" (6,65) e "experie" (8,73) sugere a relevância de fatores como desenvolvimento profissional e experiência prévia como critérios essenciais para a definição de remunerações diferenciadas. Esses elementos indicam que os empregadores buscam alinhar o salário às competências técnicas e à trajetória profissional dos candidatos, reforçando o vínculo entre qualificação e remuneração.

Portanto, os resultados da análise textual indicam que as descrições dos anúncios de emprego não se limitam à abordagem direta de salários, mas também enfatizam aspectos complementares, como benefícios adicionais, oportunidades de desenvolvimento e critérios para progressão salarial. Essa abordagem integrada reforça a atratividade das vagas, evidenciando a importância de fatores qualitativos na estratégia de captação de talentos pelas empresas.

A Figura 16 apresenta a distribuição geográfica dos anúncios coletados por estado no Brasil. Observa-se que o estado de São Paulo concentra o maior número de anúncios (473), seguido pelo Rio Grande do Sul (370) e Rio de Janeiro (107). Por outro lado, estados como

Roraima, Rondonia e Mato Grosso do Sul não tiveram nenhum anúncio coletado no período.

Figura 16 – Distribuição geográfica dos anúncios coletados por estado no Brasil. As tonalidades representam a quantidade de anúncios, variando de cores mais claras (menor concentração) para mais escuras (maior concentração).



Fonte: Elaborado pelo autor (2025).

As tonalidades mais escuras no mapa indicam uma maior densidade de anúncios, com concentração predominante nas regiões Sudeste e Sul do Brasil. Esse padrão pode estar associado a fatores como o desenvolvimento econômico dessas regiões, a maior densidade populacional e a presença significativa de empresas de tecnologia e serviços.

Além disso, possíveis divergências na distribuição dos anúncios podem ser explicadas pelo uso de plataformas locais ou pela preferência por divulgações presenciais, que podem impactar a visibilidade e o alcance dos dados apresentados.

Adicionalmente, observa-se um elevado volume de vagas no formato remoto e híbrido. Esse achado está em consonância com a análise de Hensvik, Le Barbanchon & Rathelot (2021),

que destaca como a pandemia de COVID-19 impulsionou a adoção do trabalho remoto no setor de tecnologia.

4.1 VISÃO GERAL DAS DESCRIÇÕES

A Tabela 3 apresenta um resumo das métricas de importância dos termos identificados nas descrições das vagas, com base no cálculo da importância atribuída pelo TF-IDF. A tabela também exibe o total de ocorrências de cada termo e o número de documentos em que esses termos aparecem.

Tabela 3 – Resumo das Métricas de Importância de Termos

Termo	Importância (TF-IDF)	Total de Ocorrências	Total de Documentos
desenvolv	133.33	5557	1413
experienc	94.53	3745	1192
conhec	84.15	3273	1200
sistem	66.45	1911	875
tecnolog	58.70	1964	1004
dad	53.94	1716	804
empr	51.67	1605	889
requisit	51.37	1433	955
equip	50.63	1679	843
softw	49.10	1299	676
tecn	47.35	1408	739
program	46.97	1060	631
soluco	45.47	1531	750
projet	45.28	1302	728
serv	42.78	1206	657
test	41.59	1206	686
programaca	39.40	798	548
trabalh	38.94	1214	709
are	38.11	1003	663
banc	37.61	987	679

Fonte: Elaborado pelo autor (2025).

Com base nos dados da Tabela 3, observa-se que o termo mais relevante foi "desenvolv", com um TF-IDF de 133,33, aparecendo em 1.413 documentos e registrado 5.557 vezes no *corpus* analisado. Esse resultado evidencia a centralidade de competências relacionadas ao desenvolvimento de software, englobando desde a criação de sistemas até a implementação de soluções tecnológicas. Outros termos relevantes especificam as exigências técnicas das vagas:

- "experienc" (94,53 TF-IDF; 3.745 ocorrências): destaca a importância de experiência prática, indicando que o mercado valoriza profissionais com vivência anterior em projetos de desenvolvimento.
- "conhec" (84,15 TF-IDF; 3.273 ocorrências): sinaliza a demanda por conhecimentos técnicos específicos, como linguagens de programação, frameworks e ferramentas de

desenvolvimento.

- "sistem" (66,45) e "tecnolog" (58,70): evidenciam a necessidade de compreensão de arquiteturas de sistemas e domínio de tecnologias emergentes.

Termos diretamente ligados a práticas e ferramentas de desenvolvimento também aparecem com destaque:

- "softw" (49,10) e "program" (46,97) indicam um foco em desenvolvimento de software e programação como habilidades essenciais.
- "test" (41,59) sugere a relevância de conhecimentos em testes de software, assegurando a qualidade e desempenho das aplicações.
- Termos como "banc" (37,61) e "dad" (53,94) demonstram a demanda por competências em bancos de dados e manipulação de dados, áreas essenciais no desenvolvimento de sistemas modernos.

Ainda de acordo com a tabela 3, "requisit" (51,37) aponta para a necessidade de profissionais capazes de compreender e implementar requisitos técnicos em projetos de software. Por fim, termos como "soluco" (45,47) e "projet" (45,28) refletem a busca por profissionais que atuem na solução de problemas e no gerenciamento de projetos, habilidades fundamentais em equipes de desenvolvimento. A Figura 17 demonstra uma visão de nuvem de palavras dos principais termos. Termos maiores possuem uma maior importância (TD-IDF).

Figura 17 – Nuvem da palavras dimensionadas pela importância (TD-IDF)



Fonte: Elaborado pelo autor (2025).

Os bigramas mais relevantes, apresentados na Tabela 4 mostram a forte presença de termos relacionados a bancos de dados, como "banc dad" (TF-IDF de 22,42), o que aprofunda a

importância de "dad" (TF-IDF de 53,94) já identificada na análise dos unigramas. enquanto "sql serv" (10,35) aponta para a necessidade de familiaridade com bancos de dados relacionais.

Tabela 4 – Resumo das Métricas de Importância de N-gramas

N-grama	Importância (TF-IDF)	Total de Ocorrências	Total de Documentos
banc dad	22.42	849	609
desenvolv softw	15.65	434	313
escolar ensin	14.24	177	175
rea atua	13.83	167	167
experienc desenvolv	13.37	386	298
ensin medi	13.21	175	169
informa ti	13.02	160	158
ensin superi	12.96	266	216
salari combin	12.69	219	206
ti tecnolog	12.18	142	142
tecnolog localiza	12.13	139	139
atua informa	12.13	139	139
desenvolv sistem	11.52	251	208
html cs	11.03	266	240
formaca academ	10.83	178	178
faz part	10.43	388	300
sql serv	10.35	215	175
test unitari	10.31	313	259
combin carg	10.30	158	158
segur vid	9.89	355	353

Fonte: Elaborado pelo autor (2025).

O bigrama "desenvolv softw" (15,65) reforça a centralidade do termo "desenvolv" (133,33), evidenciando o foco das vagas no desenvolvimento de software como atividade principal. Outras combinações, como "experienc desenvolv" (13,37) e "desenvolv sistem" (11,52), complementam essa tendência ao especificar que, além de habilidades técnicas, o mercado demanda experiência prática.

Na análise dos bigramas, requisitos de formação acadêmica também são recorrentes, com termos como "ensin medi" (13,21), "ensin superi" (12,96) e "formaca academ" (10,83), sugerindo que a maioria das vagas exige, no mínimo, o ensino médio completo. Além disso, o bigrama "html cs" (11,03) destaca a busca por profissionais com conhecimentos em desenvolvimento web.

Além disso, bigrama "test unitari" (10,31 TF-IDF) especifica que a exigência por testes de software se refere, em grande parte, a testes unitários, indicando a valorização de práticas de garantia de qualidade no desenvolvimento. Isso evidencia a preocupação do mercado com boas práticas, como TDD (Test-Driven Development) e automação de testes, além da expectativa de que os candidatos tenham familiaridade com frameworks específicos para testes.

Nesse contexto, a análise dos trigramas mais relevantes nas descrições de vagas de desenvolvimento de software revela combinações de termos que oferecem uma compreensão mais detalhada das exigências do mercado. Esses agrupamentos permitem identificar contextos mais específicos nas demandas das empresas.

Tabela 5 – Resumo das Métricas de Importância de Trigramas

N-grama	Importância (TF-IDF)	Total de Ocorrências	Total de Documentos
rea atua informa	10.81	139	139
informa ti tecnolog	10.81	139	139
atua informa ti	10.81	139	139
ti tecnolog localiza	10.81	139	139
escolar ensin medi	10.46	130	129
salari combin carg	9.57	158	158
formaca academ nao	8.72	138	138
nao inform salari	8.72	138	138
academ nao inform	8.72	138	138
inform salari combin	8.	120	120
banc dad relac	7.12	212	199
html cs javascript	6.63	147	134
tecnolog localiza atribu	6.19	56	56
banc dad sql	5.62	119	110
ram recurs human	5.51	68	63
combin carg program	5.41	74	74
ram informa tecnolog	5.16	69	69
recurs human recrut	5.13	63	63
human recrut seleca	5.13	63	63
cienc computaca engenh	4.74	116	113

Fonte: Elaborado pelo autor (2025).

Os trigramas apresentados na Tabela 5 como "rea atua informa" e "atua informa ti" sugerem uma categorização mais específica dentro do mercado de tecnologia. Além disso, a recorrência do trigrama "ti tecnolog localiza" (10,81 TF-IDF) sugere que as vagas também abordam questões de localização, possivelmente indicando formatos de trabalho como presencial, remoto ou híbrido. Esse detalhe aponta para uma adaptação do setor de tecnologia às novas formas de trabalho.

Do mesmo modo, os trigramas reforçam a exigência de ensino médio ("escolar ensin medi" – 10,46), mas também revelam nuances, como a falta de clareza sobre requisitos acadêmicos ("formaca academ nao" – 8,73) e a valorização de cursos específicos ("cienc computaca engenh" – 4,74).

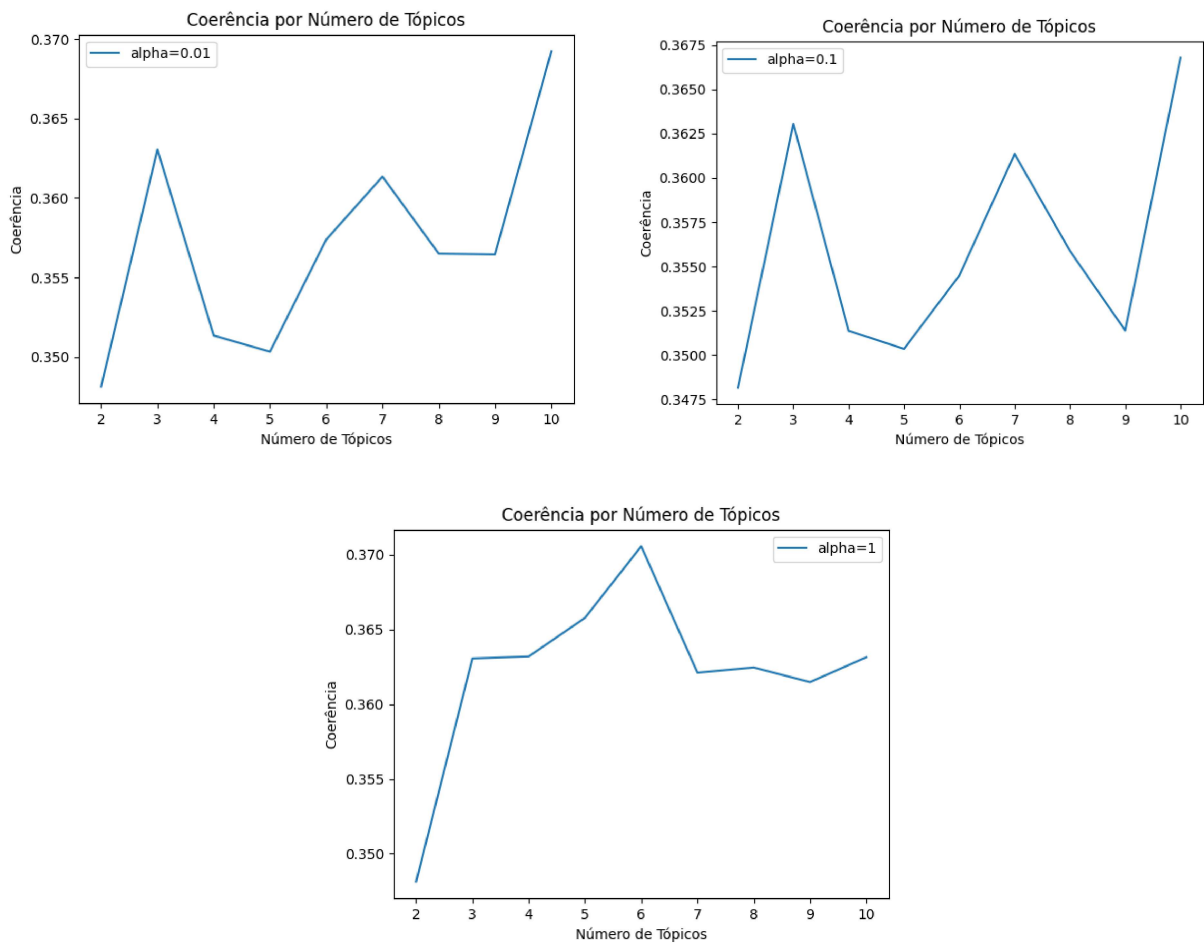
Ademais, trigramas aprofundam a exigência por bancos de dados ao detalhar tecnologias e metodologias específicas. "Banc dad relac" (7,12) indica um foco em bancos de dados relacionais, enquanto "banc dad sql" (5,62) reforça a necessidade de domínio em SQL para manipulação e consulta de dados. No desenvolvimento web, "html cs javascript" (6,63) evidencia a valorização de profissionais que dominam o conjunto dessas três tecnologias, fundamentais para a criação de interfaces dinâmicas e interativas.

Por fim, surgem termos ligados a setores específicos de atuação, como "ram recurs human", "recurs human recrut" e "human recrut seleca" (todos próximos de 5,13), sinalizando que algumas vagas envolvem atuação em áreas de Recursos Humanos, especialmente nos processos de recrutamento e seleção.

4.2 MODELAGEM DE TÓPICOS

A modelagem de tópicos com LDA foi aplicada para diferentes números de tópicos, seguida pela avaliação da coerência dos resultados obtidos. A Figura 18 apresenta a variação da coerência em função do número de tópicos para três valores distintos do parâmetro α (0,01; 0,1; e 1). Cada gráfico ilustra a relação entre coerência e quantidade de tópicos, permitindo comparar o impacto de α no desempenho do modelo.

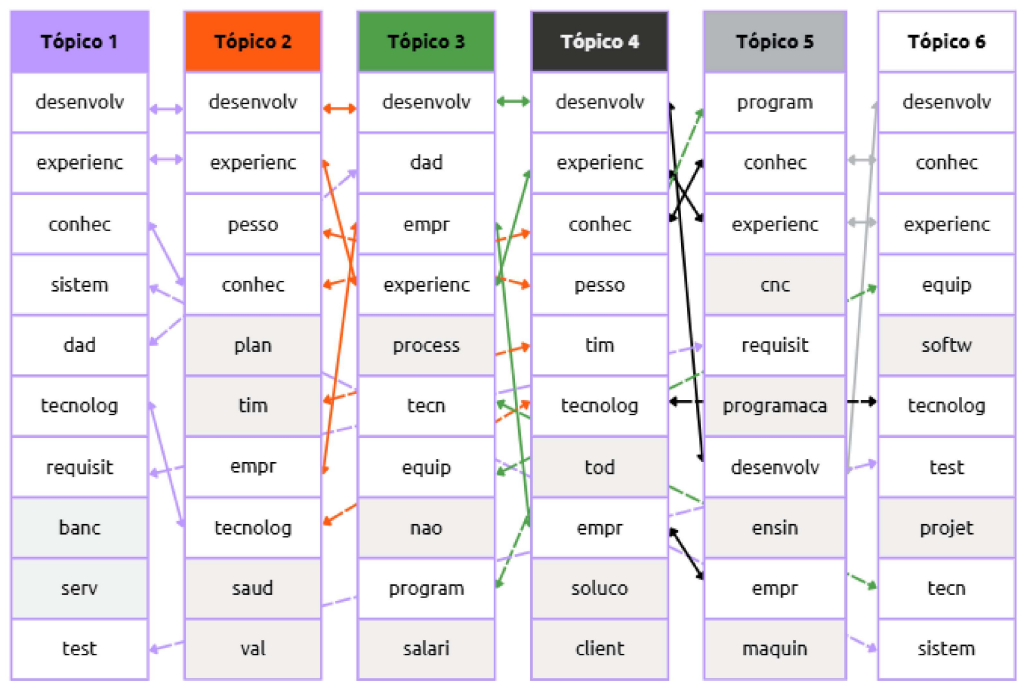
Figura 18 – Gráficos com valores de coerência para diferentes números de tópicos e variação do parâmetro α



Fonte: Elaborado pelo autor (2025).

Conforme demonstrado nos gráficos da Figura 18, o maior valor de coerência foi alcançado com a configuração $\alpha = 1$ e 6 tópicos, resultando em uma coerência aproximada de 0,37. Como o valor de coerência em modelos LDA varia de 0 a 1, sendo valores próximos a 0,7 geralmente considerados ideais, a coerência obtida neste estudo é classificada como baixa. Contudo, isso não impede a utilização do modelo, conforme destacado em estudos anteriores, como o de GUIMARÃES, JÚNIOR & FREITAS (2022). Após aplicar esses parâmetros ao LDA, os resultados estão representados na Figura 19.

Figura 19 – Visualização dos Tópicos Gerados pelo Modelo de Mineração de Texto

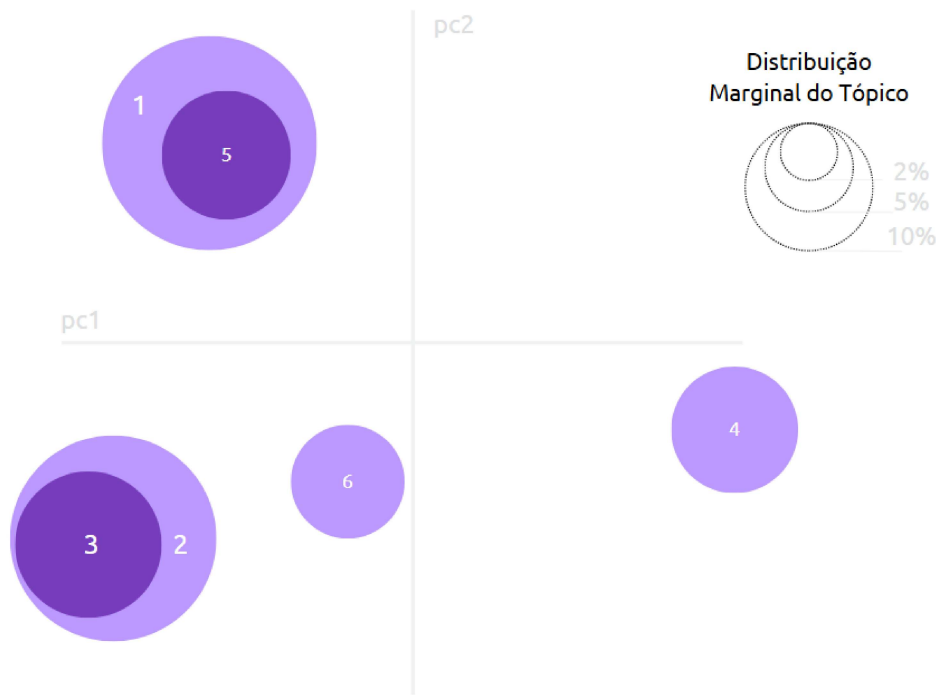


Fonte: Elaborado pelo autor (2025).

A Figura 20 ilustra a relação entre os tópicos por meio de um Mapa de Distância Intertópico. Esse gráfico permite visualizar a proximidade e a sobreposição dos tópicos extraídos.

Figura 20 – Mapa de distância intertrópico

Mapa de Distância entre Tópicos (via escala multidimensional)



Fonte: Elaborado pelo autor (2025).

Os dados das Figuras 20 e 19 revelam uma predominância de temas relacionados ao desenvolvimento tecnológico e à gestão de projetos dentro do *corpus* analisado. O tema com maior representatividade foi o de Desenvolvimento de Sistemas e Tecnologia, abrangendo 29,2% dos dados. Os termos mais recorrentes, como desenvolv, experienc, conheç, sistem, dad e tecnolog, indicam um forte foco em desenvolvimento de sistemas, tecnologia da informação e gestão de dados, evidenciando a centralidade desses assuntos nas discussões presentes.

Outro aspecto relevante identificado foi o foco em Experiência Pessoal e Planejamento, representando 27,7% do *corpus*. Esse tópico destacou termos como experienc, pesso, plan e tim, sugerindo discussões voltadas ao desenvolvimento pessoal, planejamento estratégico e gestão de equipes. Esse resultado demonstra a importância atribuída às dinâmicas interpessoais e ao gerenciamento eficiente de projetos.

As discussões sobre Processos Empresariais e Gestão de Dados também tiveram destaque, com 14,1% de representatividade. Os termos process, empr, dad e salari sugerem um interesse em práticas corporativas, gestão de dados e questões organizacionais, complementando o foco tecnológico com uma perspectiva mais orientada aos negócios.

Outro tema de grande relevância foi a Entrega de Soluções Tecnológicas e Atendimento ao Cliente, que compreendeu 10,6% do *corpus*. Os termos soluco, client, tim e tecnolog indicam um alinhamento entre inovação tecnológica e a busca por um atendimento ao cliente mais eficaz, reforçando a importância de soluções inovadoras para melhorar a experiência do consumidor.

A Programação e Automação Industrial também se destacou, abrangendo 10,4% dos dados. Palavras como program, cnc, requisit, programaca e maquin apontam para a relevância de processos automatizados e tecnologias de controle industrial, refletindo a crescente demanda por eficiência operacional e inovação no setor produtivo.

Por fim, a análise identificou um foco significativo em Projetos de Software e Testes, com 8% de representatividade. Termos como softw, test, projet e sistem mostram a relevância de práticas relacionadas ao desenvolvimento de software, gerenciamento de projetos e validação de sistemas, destacando a preocupação com a qualidade e a eficácia das soluções desenvolvidas.

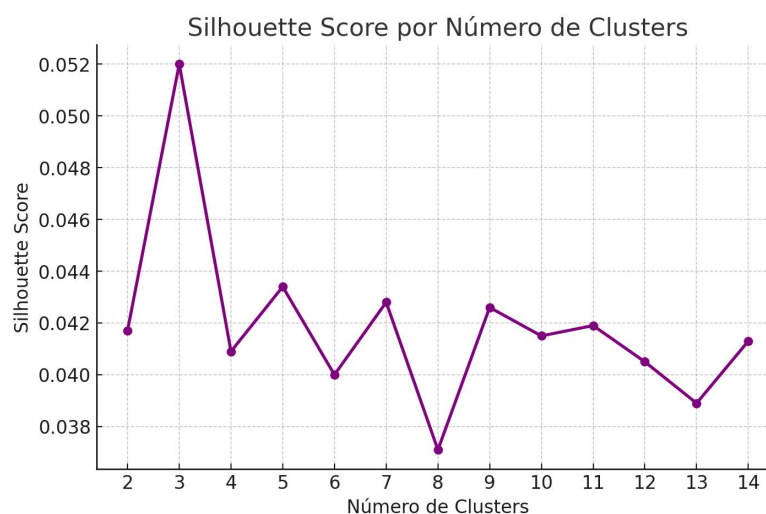
De maneira geral, a análise evidencia que os temas relacionados ao desenvolvimento de sistemas, inovação tecnológica e gestão empresarial são amplamente discutidos no *corpus*. A sobreposição e a proximidade entre alguns tópicos sugerem a existência de conexões relevantes entre as áreas de tecnologia e gestão de processos. Além disso, a presença de temas relacionados à experiência pessoal e automação industrial demonstra a diversidade de abordagens, refletindo a complexidade e a amplitude do contexto organizacional analisado.

Esses resultados são fundamentais para embasar estratégias de desenvolvimento e inovação, oferecendo uma visão clara das áreas de maior interesse e das relações entre diferentes temas. Esse entendimento contribui diretamente para a formulação de ações mais assertivas no contexto estudado, favorecendo a evolução tecnológica e o aprimoramento da gestão organizacional.

4.3 CLASSIFICAÇÃO

A aplicação do algoritmo K-Means foi conduzida para segmentar os dados textuais relacionados às vagas de emprego, resultando em uma clusterização eficiente após a remoção de 88 outliers. O conjunto final analisado compreendeu 1667 documentos. Para definir o número ideal de clusters, foi utilizada a métrica *Silhouette Score*, que avalia a coesão e separação entre os grupos formados. No gráfico da Figura 21 mostra a variação do *Silhouette Score* com relação ao número de *clusters*.

Figura 21 – *Silhouette Score* por número de *clusters*



Fonte: Elaborado pelo autor (2025).

O resultado indicou que a configuração com 3 *clusters* apresentou um *Silhouette Score* de 0,0520. Valores superiores a 0,5 são considerados satisfatórios, refletindo uma segmentação adequada e coesa entre os grupos (ROUSSEEUEW, 1987). No entanto, neste caso, a métrica reflete que não foi possível obter uma segmentação clara, indicando sobreposição entre os *clusters* ou baixa coesão dos grupos formados.

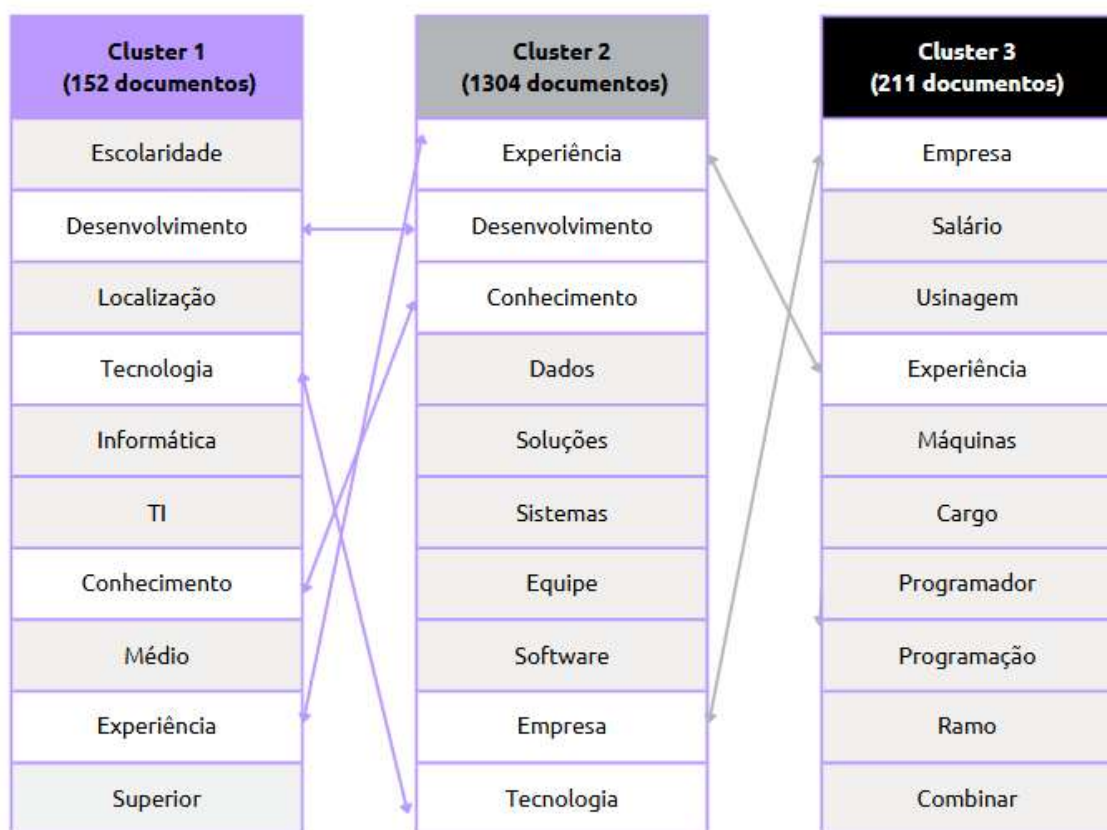
Para representar os textos numericamente, foi utilizado o *TfidfVectorizer*, configurado para priorizar termos mais relevantes e ricos em contexto. Foram considerados até 900 termos mais significativos (*max_features*=900), removendo-se as stopwords e incluindo unigramas, bigramas e trigramas (*ngram_range*=(1, 3)). Embora essas configurações visassem reduzir a dimensionalidade e o ruído, os resultados evidenciaram dificuldades em capturar nuances semânticas mais profundas, impactando negativamente as etapas posteriores.

Em seguida, para melhorar a qualidade dos dados, foi aplicada a detecção de *outliers* utilizando o *IsolationForest*, com uma taxa de contaminação ajustada para 5% (*contamination*=0.05). Essa etapa identificou e removeu registros atípicos. No entanto, mesmo após essa exclusão, os agrupamentos não apresentaram uma melhora perceptível, sugerindo que outros fatores estavam dificultando a formação de agrupamentos coerentes.

O agrupamento inicial foi realizado com o KMeans, testando uma faixa de 2 a 14 *clusters*, utilizando o Silhouette Score para determinar o número ideal de grupos. Apesar da configuração de múltiplas inicializações ($n_{init}=50$) para garantir maior consistência, os resultados não mostraram uma separação clara ou representativa da estrutura dos dados.

Posteriormente, a redução de dimensionalidade com PCA foi utilizada para gerar uma visualização em 2D dos agrupamentos formados. Contudo, as visualizações reforçaram a ausência de uma separação consistente entre os grupos, indicando que as abordagens testadas, mesmo com ajustes nos parâmetros, não foram suficientes para segmentar os dados de maneira satisfatória.

Figura 22 – Palavras e volume de documentos por *cluster*.



Fonte: Elaborado pelo autor (2025).

A Figura 22 apresenta as palavras-chave representativas identificadas para cada *cluster*, destacando os principais termos que caracterizam os documentos agrupados. Essa segmentação revela a diversidade das vagas analisadas, abrangendo desde qualificações acadêmicas até demandas técnicas e operacionais específicas.

As competências técnicas e interpessoais identificadas nos anúncios de emprego corroboram com o quadro adaptado de Assyne, Ghanbari & Pulkkinen (2022), apresentado na fundamentação teórica, que destaca a importância de habilidades técnicas, como programação e design de *software*, aliadas a competências interpessoais, como comunicação eficaz e trabalho em equipe.

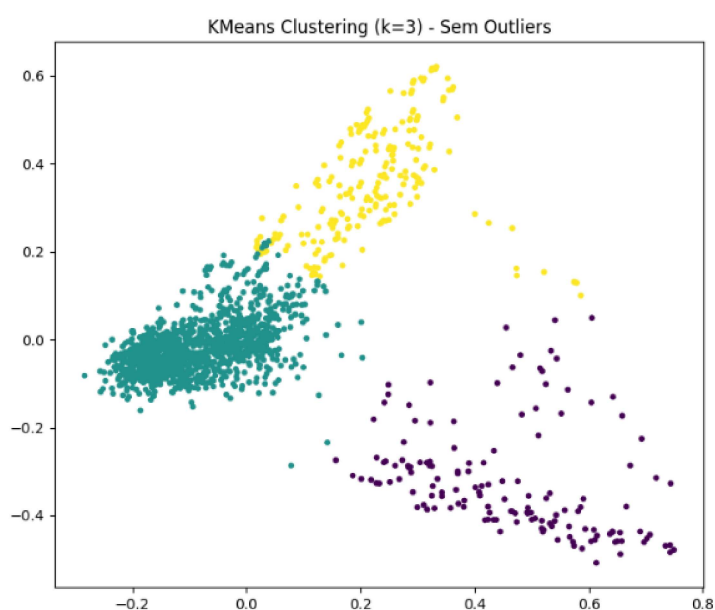
O *Cluster 1*, com 152 documentos, agrupa termos como "escolaridade", "desenvolvimento", "localização", "tecnologia", "informática", "TI", "conhecimento", "médio", "experiência" e "superior". Como mostrado na tabela, esse cluster reflete vagas que priorizam requisitos acadêmicos e competências técnicas básicas. Ele evidencia a valorização da formação educacional como critério essencial de seleção para algumas oportunidades de emprego.

Já o *Cluster 2*, o maior grupo identificado com 1.304 documentos, reúne palavras-chave como "experiência", "desenvolvimento", "conhecimento", "dados", "soluções", "sistemas", "equipe", "software", "empresa" e "tecnologia". A alta concentração de termos relacionados ao desenvolvimento de sistemas e integração em ambientes corporativos destaca a demanda crescente por profissionais qualificados em desenvolvimento de software, análise de dados e projetos empresariais.

Por fim, o *Cluster 3*, com 211 documentos, apresenta palavras-chave como "empresa", "salário", "usinagem", "máquinas", "cargo", "programador", "programação", "ramo" e "combinar". Esse cluster reflete a segmentação de vagas voltadas para o setor industrial, com foco em usinagem, operação de máquinas e negociação salarial flexível, evidenciada pela recorrência do termo "combinar" associado a "salário".

A análise visual da clusterização, apresentada no gráfico da Figura 23, ilustra a distribuição dos agrupamentos no espaço projetado. Embora os *clusters* sejam distinguíveis no gráfico, é importante ressaltar que essa separação visual pode não refletir completamente a estrutura dos dados originais, especialmente considerando que a projeção bidimensional pode distorcer relações em alta dimensionalidade. Assim, o baixo *Silhouette Score* corrobora que os *clusters* formados apresentam baixa coesão interna e separação entre si.

Figura 23 – Distribuição dos *Clusters* Formados pelo Algoritmo K-Means (k=3) sem *Outliers*



Fonte: Elaborado pelo autor (2025).

Nesse contexto, resultados do K-Means demonstram sua capacidade de segmentar dados

textuais, revelando três perfis distintos de vagas de emprego: aquelas que enfatizam qualificações acadêmicas, as que demandam competências técnicas avançadas e as que priorizam aspectos salariais e funções industriais. Apesar do *Silhouette Score* indicar uma segmentação com baixa coesão e separação limitada, o algoritmo foi eficaz em destacar padrões relevantes nos dados. A análise visual reflete esses padrões, evidenciando a diversidade das demandas do mercado de trabalho. Esses *insights* são valiosos para orientar candidatos e instituições sobre as competências mais requisitadas e as áreas com maior concentração de oportunidades, auxiliando na definição de estratégias alinhadas às tendências do mercado.

4.4 COMPETÊNCIAS TÉCNICAS E SENIORIDADE

A análise dos termos mais relevantes nos títulos das vagas de emprego, com base no cálculo do TF-IDF, permite identificar padrões relacionados à senioridade exigida nas oportunidades analisadas. Os resultados revelam a distribuição de níveis de experiência buscados pelas empresas e destacam as principais tendências do mercado.

Tabela 6 – Valores de TF-IDF e Ocorrências de Termos Relacionados a Níveis de Experiência

Termo	TF-IDF	Número de documentos
seni	70,71	175
sr	29,74	54
plen	77,74	183
pl	22,85	39
juni	49,63	101
jr	30,39	51
estagi	18,99	37

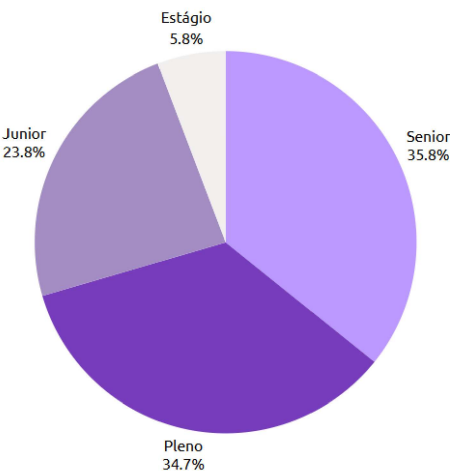
Fonte: Elaborado pelo autor (2025).

Os resultados da Tabela 6 indicam que há uma demanda considerável por profissionais de nível pleno e sênior, sugerindo que as empresas priorizam profissionais com mais experiência para ocuparem cargos estratégicos. A predominância do termo "plen" reforça que há um equilíbrio entre a busca por profissionais que possuem certa bagagem profissional, mas ainda têm espaço para crescimento dentro das organizações.

Embora o nível júnior também tenha representatividade nos dados, com termos como "juni" e "jr", o número de ocorrências foi inferior aos níveis pleno e sênior. Isso pode indicar que as empresas buscam profissionais mais experientes para lidar com projetos mais complexos ou que há uma menor oferta de vagas destinadas a profissionais iniciantes.

Por fim, a presença de termos como "estagi" revela oportunidades voltadas para estágios, indicando abertura para a formação de novos profissionais, mesmo que em menor escala. No gráfico 24, os dados foram agrupados por nível de senioridade e o percentual correspondente foi calculado. Observa-se que, à medida que a senioridade aumenta, o volume de vagas abertas também cresce.

Figura 24 – Percentual de Vagas por Nível de Experiência.



Fonte: Elaborado pelo autor (2025).

Os resultados obtidos com o uso das técnicas de mineração de texto, como BoW, TF-IDF e LDA, são consistentes com as aplicações propostas por Feldman & Sanger (2007) e Blei, Ng & Jordan (2003), que destacam a eficácia dessas técnicas para a extração de padrões em grandes volumes de dados textuais.

Os resultados da Tabela 7 destacam a predominância do termo "Desenvolvimento" com um valor de TF-IDF de 239,28 e 1.065 ocorrências, confirmando o foco central do *corpus* na criação e implementação de sistemas e softwares. O termo "Programação" também apresentou alta relevância (TF-IDF de 141,61 e 496 ocorrências), evidenciando a ênfase nas práticas de codificação e na aplicação de linguagens de programação.

Tabela 7 – Principais Termos Técnicos Identificados no *Corpus*

Termo Técnico	TF-IDF Total	Ocorrências
Desenvolvimento (desenvolv)	239,28	1.065
Programação (program)	141,41	496
Java (jav)	77,06	153
Backend (backend)	67,53	141
Software (softw)	58,80	140
Fullstack (fullstack)	57,32	103
Frontend (frontend)	48,37	82
PHP (php)	49,02	89
Python (python)	26,47	41
React (react)	25,37	52
Node.js (nodej)	19,48	34
Angular (angul)	18,29	30

Fonte: Elaborado pelo autor (2025).

Linguagens como Java (TF-IDF de 77,06; 153 ocorrências), PHP (TF-IDF de 49,02; 89 ocorrências) e Python (TF-IDF de 26,47; 41 ocorrências) foram amplamente mencionadas,

evidenciando que o *corpus* contempla diferentes paradigmas e arquiteturas de desenvolvimento de software. Além disso, a presença expressiva de termos como Backend (TF-IDF de 67,53; 141 ocorrências), Frontend (TF-IDF de 48,37; 82 ocorrências) e Fullstack (TF-IDF de 57,32; 103 ocorrências) reforça a abordagem abrangente do estudo, cobrindo desde aspectos conceituais até a aplicação prática das tecnologias analisadas.

No contexto de desenvolvimento de interfaces, termos como React (TF-IDF de 25,37; 52 ocorrências), PHP (TF-IDF de 49,02; 89 ocorrências), Node.js (TF-IDF de 19,48; 34 ocorrências) e Angular (TF-IDF de 18,29; 30 ocorrências) evidenciam a importância de competências técnicas relacionadas ao front-end e full stack. Esses dados estão alinhados com o estudo de Montandon *et al.* (2021), que também destacou Java, Python e JavaScript como linguagens altamente demandadas, além dos frameworks React e Angular como diferenciais competitivos no mercado.

A análise foi ampliada para incluir os *stems* das descrições das vagas, conforme apresentado na Tabela 8. O termo "dad" (banco de dados) registrou o maior TF-IDF (198,33), com 1.716 ocorrências em 804 documentos, evidenciando sua centralidade nos temas analisados. Outros termos relevantes incluem "serv" (servidores, TF-IDF de 168,41), "process" (processos, 163,31) e "test" (testes, 162,31), sugerindo um forte enfoque em infraestrutura, validação e operações tecnológicas. Esse padrão também foi identificado por Montandon *et al.* (2021), que apontou a relevância de OS & Infrastructure e Process & Methods para desenvolvedores *backend*, DevOps e QA Testers. Assim, além das habilidades de codificação, o mercado valoriza conhecimentos em gestão de servidores, automação de processos e qualidade de software.

Tabela 8 – Resumo de N-gramas (TF-IDF, Ocorrências e Documentos)

Termo	TF-IDF	Ocorrências	Documentos
dad	198.33	1716	804
serv	168.41	1206	657
process	163.31	1069	605
test	162.31	1206	686
banc	132.13	987	679
jav	111.54	625	361
ti	105.87	363	310
cnc	101.83	329	117
backend	99.29	552	354
sql	94.42	564	458
design	91.11	609	379
frontend	90.29	491	292
react	89.03	495	297
net	83.01	366	193
framework	82.59	520	384
javascript	81.49	493	413
computaca	78.71	373	272
git	76.59	505	457
php	74.23	304	193

Fonte: Elaborado pelo autor (2025).

Embora JavaScript não tenha sido um dos termos mais destacados na Tabela 7, sua relevância foi confirmada na análise das descrições das vagas, conforme demonstrado na Tabela 8. O termo apresentou um TF-IDF de 81,49, com 493 menções em 413 documentos, reforçando sua importância no desenvolvimento *frontend* e *full stack*. Esse resultado indica que, enquanto Java e PHP aparecem mais frequentemente nos títulos dos anúncios de emprego, JavaScript é mais citado nas descrições detalhadas das vagas, possivelmente associado a frameworks e bibliotecas específicas, como React e Angular, que também foram identificados nesta análise.

No contexto de tecnologias essenciais, termos especializados como "framework" (TF-IDF de 82,59) e "git" (TF-IDF de 76,59) reforçam a importância de conhecimentos relacionados à implementação, versionamento e colaboração no desenvolvimento de *software*.

Além disso, a análise revelou que o termo "Net" (TF-IDF 83,01; 366 ocorrências) está possivelmente relacionado ao *framework* .NET, indicando uma demanda relevante por essa tecnologia no mercado. Já o termo "Design" (TF-IDF 91,11; 609 ocorrências) sugere a importância de aspectos visuais e estruturais, podendo estar ligado tanto à arquitetura de *software* quanto ao design de interfaces (UI/UX).

A comparação entre as tabelas 7 e 8 destaca os termos em comum, como mostrado na Tabela 9. Esses dados reforçam a relevância dos temas tanto em frequência geral quanto em sua importância contextual. Por exemplo:

- "Java": Com TF-IDF de 77,06 na Tabela 7 e 111,54 na Tabela 8, é amplamente mencionado e particularmente relevante no backend e desenvolvimento de sistemas, com 625 ocorrências em 361 documentos;
- "Backend": Aparece com TF-IDF de 67,53 na Tabela 7 e 99,29 na Tabela 8, indicando seu papel fundamental em arquiteturas e APIs;
- "Frontend": Com TF-IDF de 48,37 (Tabela 7) e 90,29 (8), está associado ao design e desenvolvimento de interfaces, destacando-se na experiência do usuário.
- "PHP": Embora tradicional, apresenta TF-IDF de 49,02 e 74,23, reforçando sua presença tanto em sistemas legados quanto em soluções backend.

A Tabela 9 sintetiza os principais termos e suas interpretações:

Tabela 9 – Comparação entre a frequência dos principais termos dos títulos e das descrições

Termo	TF-IDF Títulos	TF-IDF Descrições	Interpretação
Java	77,06	111,54	Ampla relevância no desenvolvimento backend.
Backend	67,53	99,29	Importância na arquitetura e construção de sistemas.
Frontend	48,37	90,29	Foco em interfaces e experiência do usuário.
React	25,37	89,03	Tecnologia moderna, destaque em contextos estratégicos.
PHP	49,02	74,23	Consolidação em backend e sistemas legados.

Fonte: Elaborado pelo autor (2025).

Os resultados destacam um enfoque significativo em temas centrais, como banco de dados, servidores e processos, alinhando-se às demandas práticas do mercado. Tecnologias amplamente difundidas, como Java, PHP e SQL, coexistem com ferramentas modernas, como React, refletindo a integração de práticas consolidadas e inovação tecnológica.

4.5 RECOMENDAÇÕES PARA ALINHAMENTO DE COMPETÊNCIAS E DEMANDAS DO MERCADO

Os resultados obtidos na análise dos anúncios de vagas para desenvolvedores de software evidenciam padrões claros em relação às competências técnicas mais demandadas, às exigências de experiência profissional e às expectativas do mercado de trabalho. A partir dessas informações, é possível apontar recomendações que possam contribuir para uma melhor conexão entre candidatos, instituições de ensino e empregadores, promovendo um alinhamento mais eficaz entre oferta e demanda de talentos na área de tecnologia.

Para os profissionais que buscam ingressar ou se posicionar no mercado, os dados indicam a necessidade de um foco estratégico no desenvolvimento de habilidades técnicas específicas. Linguagens de programação como Java, JavaScript, PHP e Python, além do domínio de bancos de dados e frameworks como React e Angular, aparecem recorrentemente nas descrições das vagas, demonstrando a importância dessas tecnologias para quem deseja ampliar sua empregabilidade.

No entanto, as exigências não se limitam às competências técnicas, mas a experiência prática também se destaca como um fator decisivo no processo seletivo. As empresas valorizam candidatos que demonstram aplicação real do conhecimento, seja por meio de projetos desenvolvidos, participação em comunidades de código aberto ou contribuições para iniciativas reais. Isso evidencia que a formação acadêmica, por si só, pode não ser suficiente para garantir um diferencial competitivo. Além disso, as habilidades interpessoais têm um papel relevante na seleção dos profissionais. A análise dos dados mostrou que comunicação eficaz, trabalho em equipe e colaboração são aspectos frequentemente mencionados, indicando que o perfil ideal para as empresas vai além do conhecimento técnico e requer a capacidade de interagir de forma produtiva dentro das equipes de desenvolvimento.

As instituições de ensino desempenham um importante papel na preparação desses profissionais, e os resultados deste estudo reforçam a necessidade de atualização contínua dos currículos acadêmicos para acompanhar a evolução do mercado. A inclusão de disciplinas voltadas para metodologias de gestão de projetos, versionamento e testes automatizados podem aumentar a aderência dos cursos às demandas do setor. Além disso, iniciativas que incentivem o aprendizado baseado em projetos e a interação com o mercado, como parcerias com empresas, *hackathons* e *bootcamps*, tendem a proporcionar uma experiência mais prática demandada pelo setor. O incentivo ao aprendizado contínuo também se apresenta como um fator essencial, uma vez que a área de tecnologia evolui rapidamente e demanda que os profissionais se mantenham atualizados ao longo de suas carreiras.

No que se refere às empresas e recrutadores, os resultados apontam para algumas lacunas

que podem ser ajustadas para melhorar a eficiência dos processos seletivos e atrair candidatos mais qualificados. A falta de clareza em muitos anúncios de emprego foi um ponto identificado na análise textual, sugerindo a necessidade de descrições mais objetivas e detalhadas sobre requisitos técnicos, expectativas do cargo e informações salariais. A ausência desses dados pode afastar talentos qualificados ou prolongar etapas do processo seletivo devido a desalinhamentos entre as expectativas do candidato e da empresa.

Diante dessas observações, as recomendações apresentadas visam não apenas contribuir para o aperfeiçoamento das estratégias de contratação e formação profissional, mas também incentivar práticas que favoreçam um ecossistema mais equilibrado entre as necessidades das empresas e as expectativas dos profissionais. O alinhamento entre as competências ofertadas e as demandas do mercado pode ser aprimorado por meio de uma comunicação mais transparente nos processos seletivos, da adaptação dos currículos acadêmicos às exigências atuais do setor e da valorização do aprendizado prático e contínuo por parte dos candidatos. Esse conjunto de ações pode potencializar a eficiência dos processos de recrutamento, reduzir gaps de qualificação e tornar o setor de tecnologia mais acessível e alinhado às exigências do futuro.

4.6 LIMITAÇÕES DA PESQUISA

Apesar dos resultados relevantes obtidos, esta pesquisa apresenta algumas limitações que devem ser consideradas na interpretação dos dados. A coleta de dados foi restrita a três plataformas de emprego – LinkedIn, Indeed e Gupy – o que pode não refletir a totalidade das oportunidades disponíveis no mercado de tecnologia. Outras plataformas de recrutamento ou redes sociais, como Glassdoor, Vagas.com ou grupos especializados em tecnologia, não foram incluídas, o que pode ter impactado a abrangência da amostra.

Além disso, a análise de salários foi prejudicada pela ausência de informações claras em grande parte dos anúncios. Apenas uma fração das vagas apresentava dados completos sobre remuneração, o que limitou a profundidade da análise salarial. Essa limitação pode ter influenciado a identificação de padrões de remuneração entre diferentes níveis de senioridade.

Outro aspecto relevante é a dependência de técnicas de mineração de texto, como Bag-of-Words (BoW) e TF-IDF, que, embora eficientes, não capturam nuances semânticas mais profundas presentes nas descrições das vagas. O uso de modelos mais avançados de Processamento de Linguagem Natural (PLN), como Word2Vec ou BERT, poderia oferecer resultados mais detalhados e precisos.

Destaca-se também o desempenho insatisfatório obtido na aplicação do algoritmo K-means durante a análise dos dados. Embora a técnica seja amplamente utilizada para agrupamento em estudos semelhantes, os resultados apresentaram baixa separação entre os clusters formados, dificultando a identificação de grupos bem definidos de competências ou categorias de vagas. Esse problema pode ter sido causado pela alta dimensionalidade dos dados após a vetorização por TF-IDF, resultando em clusters sobrepostos e com baixa interpretabilidade.

Por fim, o período de coleta de dados, limitado a poucos dias, pode não ter captado

variações sazonais ou tendências de longo prazo no mercado de trabalho. Estudos com coletas contínuas ou em intervalos mais amplos poderiam oferecer uma visão mais completa sobre a evolução das demandas por competências em desenvolvimento de software.

4.7 BASE DE DADOS E CÓDIGO UTILIZADO

Os resultados obtidos neste estudo foram gerados a partir da análise dos dados disponíveis no repositório do GitHub: <https://github.com/adeliamara/job-analysis>. O código desenvolvido para o processamento e interpretação dos dados também está acessível, permitindo que os métodos empregados sejam reproduzidos e validados por outros pesquisadores.

Além disso, disponibilização desses recursos reforça a transparência e a reprodutibilidade da pesquisa, possibilitando que futuras investigações aprofundem as análises ou adaptem as abordagens para diferentes contextos. Além disso, os resultados apresentados neste capítulo foram obtidos a partir de um rigoroso processo de análise, garantindo que as conclusões refletem de forma precisa os padrões observados nos dados.

Nesse contexto, o repositório GitHub permanece aberto a contribuições, permitindo que pesquisadores e profissionais expandam as análises e testem novas metodologias. No próximo capítulo, serão discutidas as principais implicações dos resultados, bem como as limitações do estudo e sugestões para trabalhos futuros, ampliando a compreensão sobre os impactos e desdobramentos desta pesquisa.

5 CONCLUSÃO OU CONSIDERAÇÕES FINAIS

Este estudo teve como propósito analisar as demandas do mercado de trabalho para desenvolvedores de software, identificando as competências mais requisitadas por meio da aplicação de técnicas de mineração de texto e aprendizado de máquina. Com o avanço tecnológico e a crescente digitalização dos processos produtivos, a qualificação profissional tornou-se um fator determinante para a inserção no mercado de trabalho. A análise automatizada de anúncios de emprego permitiu identificar tendências e padrões que auxiliaram profissionais, empresas e instituições de ensino a alinharem suas expectativas às demandas reais do setor.

Para atingir esse objetivo, foi desenvolvido um *pipeline* de análise baseado na metodologia CRISP-DM, que compreendeu as etapas de coleta, preparação, modelagem, avaliação e interpretação dos dados. A extração das informações foi realizada a partir de plataformas de emprego, utilizando técnicas de web scraping e APIs para capturar um conjunto representativo de anúncios. A análise exploratória foi conduzida com métodos de Processamento de Linguagem Natural (PLN), como tokenização, remoção de *stopwords* e lematização, enquanto a modelagem empregou técnicas como *Bag-of-Words* (BoW), *Term Frequency-Inverse Document Frequency* (TF-IDF), *Latent Dirichlet Allocation* (LDA) e K-Means.

Os resultados indicam uma demanda consolidada por linguagens de programação, com destaque para Java, PHP e Python. Essas tecnologias apareceram frequentemente nos títulos dos anúncios de emprego, sugerindo seu papel central em muitas vagas de desenvolvedor de software.

Por outro lado, JavaScript, apesar de não estar entre os termos mais destacados nos títulos, apresentou alta recorrência nas descrições das vagas, evidenciando sua relevância no mercado. Sua forte associação com tecnologias front-end, como HTML e CSS, e com frameworks modernos, como React e Angular, ressalta seu papel essencial no desenvolvimento full stack e na construção de aplicações dinâmicas. A análise dos trigramas reforçou essa interdependência, mostrando que essas tecnologias são frequentemente citadas em conjunto nas descrições das vagas.

Além disso, a pesquisa identificou uma valorização crescente de práticas de qualidade no desenvolvimento de software, refletida pela recorrência de menções a testes unitários. Esse resultado sugere que o mercado busca profissionais que não apenas dominem tecnologias de programação, mas também adotem boas práticas de desenvolvimento, garantindo qualidade, manutenibilidade e escalabilidade do código.

Outro aspecto relevante identificado foi a importância de conhecimentos em banco de dados, evidenciada pela presença frequente de tecnologias como SQL e bancos de dados relacionais nas descrições das vagas. Esses conhecimentos são fundamentais para desenvolvedores, independentemente da especialização, uma vez que grande parte das aplicações modernas depende de armazenamento, recuperação e manipulação eficiente de dados.

Os resultados deste estudo possuem implicações significativas para diferentes atores do

mercado de TI. Profissionais podem direcionar seus esforços de qualificação para as competências mais demandadas, aumentando sua competitividade no mercado. Instituições de ensino podem alinhar seus currículos às exigências do setor, incorporando disciplinas voltadas às linguagens de programação mais utilizadas, *frameworks* modernos, testagem de *software* e bancos de dados. Já empregadores podem aprimorar seus processos de recrutamento, tornando as descrições das vagas mais claras e detalhadas, incluindo informações sobre habilidades técnicas exigidas, expectativas salariais e requisitos específicos.

Embora este estudo tenha fornecido uma visão abrangente do mercado, algumas limitações devem ser consideradas. A análise baseou-se em vagas publicadas online, o que pode não refletir integralmente todas as oportunidades disponíveis no mercado, especialmente aquelas divulgadas por redes informais ou processos internos de recrutamento. Além disso, a falta de informações salariais dificultou uma análise mais detalhada sobre a relação entre remuneração e nível de qualificação dos candidatos. Por fim, as técnicas de mineração de texto utilizadas, embora eficazes para identificar padrões gerais, podem não capturar nuances semânticas mais complexas presentes nos anúncios de emprego.

Para trabalhos futuros, recomenda-se a ampliação da base de dados, incorporando fontes mais diversificadas e explorando abordagens complementares, como entrevistas com recrutadores e análise longitudinal das tendências do mercado. A adoção de modelos mais avançados de Processamento de Linguagem Natural (PLN), como redes neurais profundas (ex.: BERT e GPT), pode aprimorar a interpretação semântica das descrições de vagas e oferecer insights ainda mais detalhados.

Em suma, este estudo contribui para uma compreensão mais aprofundada das exigências do mercado de TI, permitindo um alinhamento mais eficiente entre as expectativas das empresas e as competências dos profissionais. Espera-se que os resultados auxiliem no fortalecimento da formação acadêmica, na qualificação profissional e na otimização dos processos de recrutamento, promovendo um setor mais dinâmico, competitivo e alinhado às necessidades do mercado.

REFERÊNCIAS

- ALMEIDA, M. B. **Detecção Automática de Discurso de Ódio em Redes Sociais**. Dissertação (Master's thesis) — Universidade Federal de São Paulo (UNIFESP), São José dos Campos, SP, 2020. Disponível em: <<https://repositorio.unifesp.br/server/api/core/bitstreams/e7541bbd-ce93-4de4-a017-741a988c9e04/content>>. Acesso em: 20 dez. 2024.
- AMAZON WEB SERVICES. **O que é Análise de Sentimentos?** 2024. Disponível em: <<https://aws.amazon.com/pt/what-is/sentiment-analysis/>>. Acesso em: 20 dez. 2024.
- ASSYNE, N.; GHANBARI, H.; PULKKINEN, M. The essential competencies of software professionals: A unified competence framework. **Information and Software Technology**, v. 151, p. 107020, 2022. ISSN 0950-5849. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0950584922001446>>. Acesso em: 20 dez. 2024.
- BAUMEISTER, F.; BARBOSA, M. W.; GOMES, R. R. What is required to be a data scientist?: Analyzing job descriptions with centering resonance analysis. **International Journal of Human Capital and Information Technology Professionals (IJHCITP)**, v. 11, n. 4, p. 21–40, 2020. Disponível em: <<https://doi.org/10.4018/IJHCITP.2020100102>>. Acesso em: 20 dez. 2024.
- BEDREGAL-ALPACA, N.; ARUQUIPA-VELAZCO, D.; CORNEJO-APARICIO, V. Técnicas de data mining para extraer perfiles comportamiento académico y predecir la deserción universitaria. **Revista Ibérica de Sistemas e Tecnologías de Informação**, n. E27, p. 592–604, 2020. Disponível em: <<https://www.proquest.com/scholarly-journals/t%C3%A9cnicas-de-data-mining-para-extraer-perfiles/docview/2385757429/se-2>>. Acesso em: 20 dez. 2024.
- BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Toolkit**. 2009. Disponível em: <<https://www.nltk.org/>>. Acesso em: 10 jan. 2025.
- BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of Machine Learning Research**, v. 3, p. 993–1022, 2003. Disponível em: <<https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>>. Acesso em: 20 dez. 2024.
- BRAUNERHJELM, P.; EKLUND, J.; KÜGLER, M. 1. labor market regulations, innovation and technological change. In: . [s.n.], 2023. Disponível em: <<https://doi.org/10.4337/9781800884472.00010>>. Acesso em: 20 dez. 2024.
- CRIVEANU, M. A. H. e Mumtaz Anwar Chaudhry e Maryam Batool e Huma Ghulam e Abid Raza Khan e Cristi Spulbar e Muhammad Zahid Naeem e Ramona Birau e M. Transformando a crise em uma oportunidade sustentável em relação à demanda por treinamento e novas habilidades no mercado de trabalho: uma análise empírica da pandemia de covid-19 e atualização de habilidades. **Sustainability**, 2022. Disponível em: <<https://doi.org/10.3390/su142416785>>. Acesso em: 20 dez. 2024.
- DEVLIN, J. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics,

2019. p. 4171–4186. Disponível em: <<https://arxiv.org/abs/1810.04805>>. Acesso em: 20 dez. 2024.

DUBEY, P.; KUMAR, R. A study on importance of data processing in data mining practices. out. 2014. Disponível em: <<https://s3-eu-west-1.amazonaws.com/pfigshare-u-files/1705683/4413.pdf>>. Acesso em: 20 dez. 2024.

FELDMAN, R.; SANGER, J. **The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data**. [S.l.]: Cambridge University Press, 2007.

GALL, H. C. *et al.* Data-driven decisions and actions in today's software development. In: . [s.n.], 2017. Disponível em: <https://doi.org/10.1007/978-3-319-73897-0_9>. Acesso em: 20 dez. 2024.

GUIMARÃES, A. J. R.; JÚNIOR, R. M.; FREITAS, M. d. C. D. Requisitos para a ciência de dados: analisando anúncios de vagas de emprego com mineração de texto. **Revista Ibérica de Sistemas e Tecnologias de Informação**, n. 46, p. 54–70, 2022. Disponível em: <<https://scielo.pt/pdf/rist/n46/1646-9895-rist-46-54.pdf>>. Acesso em: 20 dez. 2024.

HADININGRUM, T. R.; MARDIANA, B. D.; ROCHIMAH, S. 1. relationship between features volatility and bug occurrence rate to support software evolution. **International Journal of Electrical and Computer Engineering**, 2024. Disponível em: <<https://doi.org/10.11591/ijece.v14i5.pp5381-5389>>. Acesso em: 20 dez. 2024.

HENSVIK, L.; Le Barbanchon, T.; RATHELOT, R. Job search during the covid-19 crisis. **Journal of Public Economics**, v. 194, p. 104349, 2021. ISSN 0047-2727. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0047272720302139>>. Acesso em: 20 dez. 2024.

HIRSCHEY, J. K. Symbiotic relationships: Pragmatic acceptance of data scraping. **Berkeley Tech. LJ**, HeinOnline, v. 29, p. 897, 2014. Disponível em: <<https://doi.org/10.15779/Z38B39B>>. Acesso em: 20 dez. 2024.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 3rd (draft). ed. [s.n.], 2021. Rascunho. Disponível em: <<https://web.stanford.edu/~jurafsky/slp3/>>. Acesso em: 20 dez. 2024.

KARAGÖL, T.; KARAGÖL, A. 1. benchmarking gromacs on optimized colab processors and the flexibility of cloud computing for molecular dynamics. 2024. Disponível em: <<https://doi.org/10.1101/2024.11.14.623563>>. Acesso em: 20 dez. 2024.

KAUR, P. Sentiment analysis using web scraping for live news data with machine learning algorithms. **Materials today: proceedings**, Elsevier, v. 65, p. 3333–3341, 2022. Disponível em: <<https://www.sciencedirect.com/science/article/abs/pii/S2214785322037075>>. Acesso em: 20 dez. 2024.

KHATTER, H. *et al.* Web scraping based product comparison model for e-commerce websites. In: IEEE. **2022 IEEE International Conference on Data Science and Information System (ICDSIS)**. 2022. p. 1–6. Disponível em: <<https://doi.org/10.1109/ICDSIS55133.2022.9915892>>. Acesso em: 20 dez. 2024.

KOŠTÁLOVÁ, Z.; LYČSA Štefan; ŠTEFÁNIK, M. 2. online job vacancy attractiveness: Increasing views, reactions and conversions. **Electronic Commerce Research and Applications**, 2022. Disponível em: <<https://doi.org/10.1016/j.elerap.2022.101192>>. Acesso em: 20 dez. 2024.

KROTOV, V.; JOHNSON, L.; SILVA, L. Tutorial: Legality and ethics of web scraping. 2020. Disponível em: <<https://digitalcommons.murraystate.edu/cgi/viewcontent.cgi?article=1071&context=faculty>>. Acesso em: 20 dez. 2024.

LAWSON, R. **Web scraping with Python**. [S.l.]: Packt Publishing Ltd, 2015.

MAYER-SCHÖNBERGER, V.; CUKIER, K. **Big Data: die Revolution, die unser Leben verändern wird**. Redline Wirtschaft, 2013. Disponível em: <<https://www.m-vg.de/redline/shop/article/3116-big-data/>>. Acesso em: 20 dez. 2024.

MIKOLOV, T. *et al.* Distributed representations of words and phrases and their compositionality. In: **Advances in Neural Information Processing Systems**. [s.n.], 2013. v. 26, p. 3111–3119. Disponível em: <<https://arxiv.org/abs/1310.4546>>. Acesso em: 20 dez. 2024.

MONTANDON, J. E. *et al.* What skills do it companies look for in new developers? a study with stack overflow jobs. **Information and Software Technology**, Elsevier, v. 129, p. 106429, 2021. Disponível em: <<https://doi.org/10.1016/j.infsof.2020.106429>>. Acesso em: 20 dez. 2024.

PAULA, A. d.; AL. et. Mercado de trabalho formal e pandemia: análise da alocação laboral em diferentes setores econômicos. **Nova Economia**, Redline Wirtschaft, v. 33, n. 01, p. 263–290, 2023. Disponível em: <<https://doi.org/10.1590/0103-6351/7110>>. Acesso em: 20 dez. 2024.

RASCHKA, S. **Python Machine Learning: Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics**. Birmingham, Mumbai: Packt Publishing Ltd., 2015. ISBN 978-1-78355-513-0.

RÖDER, M.; BOTH, A.; HINNEBURG, A. Exploring the space of topic coherence measures. In: ACM. **Proceedings of the eighth ACM international conference on Web search and data mining**. 2015. p. 399–408. Disponível em: <<https://doi.org/10.1145/2684822.2685324>>. Acesso em: 20 dez. 2024.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987. Disponível em: <[https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)>. Acesso em: 20 dez. 2024.

SAVEANU, T. *et al.* Labour market in a time of crisis. what do hr experts say about the importance of digitalization in the postpandemic context. 2023. Disponível em: <[https://doi.org/10.47535/1991auoes32\(2\)008](https://doi.org/10.47535/1991auoes32(2)008)>. Acesso em: 20 dez. 2024.

SILVA, E. M. *et al.* Metodologia para desenvolvimento de sistemas de suporte a decisão crisp-dm utilizando a notação uml: Uma abordagem aplicada à gerência de atendimento hospitalar. In: **Anais do 1 Simpósio Brasileiro de Inteligência Computacional**. Florianópolis, SC: SBRN, 2007. p. 1–5. Disponível em: <<https://sbic.org.br/wp-content/uploads/2016/03/60100046.pdf>>. Acesso em: 20 dez. 2024.

SILVA, N.; LOPES, G. Stopwords in Text Mining: Handling High-Frequency Terms in Natural Language Texts. In: **Text Mining Applications**. [S.l.: s.n.], 2013.

TOMAN, M.; TESAR, R.; JEZEK, K. Influence of word normalization on text classification. In: **Proceedings of InSciT**. [s.n.], 2006. v. 4, n. 354–358, p. 9. Disponível em: <<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=49280008b5a46a6b45f01b58afa8777d91e7ebe2>>. Acesso em: 20 dez. 2024.

VUALA, E. A.; OUTROS. Automatização e seu impacto no mercado de trabalho. **Revista Científica Multidisciplinar**, v. 1, n. 2, 2023. Disponível em: <<https://vistacien.com.br/wp-content/uploads/2023/08/7-AUTOMATIZACAO-E-SEU-IMPACTO-NO-MERCADO-DE-TRABALHO.pdf>>. Acesso em: 20 dez. 2024.

ZHAO, B. Web scraping. **Encyclopedia of big data**, Springer New York, v. 1, 2017. Disponível em: <https://doi.org/10.1007/978-3-319-32001-4_483-1>. Acesso em: 20 dez. 2024.

ŘEHŮŘEK, R.; SOJKA, P. **Gensim - Python framework for vector space modelling**. 2011. Version 3(2). Disponível em: <<https://radimrehurek.com/gensim/index.html>>. Acesso em: 20 dez. 2024.