ANÁLISE DE TENDÊNCIAS DE VAGAS DE TI NO LINKEDIN UTILIZANDO WEB SCRAPING

Abner Willian Mioti Manha[1], Guilherme Konishi Yoshihara[1], Rodrigo de Oliveira Plotze[1] , Anna Patricia Zakem China [1]

[1]Faculdade de Tecnologia de Ribeirão Preto (FATEC)

Ribeirão Preto, SP – Brasil

abner.manha@fatec.sp.gov.br,
guilherme.yoshihara@fatec.sp.gov.br,
rodrigo.plotze@fatec.sp.gov.br
anna.china@fatec.sp.gov.br

Resumo. *O projeto Análise de Tendências de Vagas de TI no LinkedIn Utilizando o Web Scraping tem como objetivo analisar as vagas no mercado de trabalho por meio de um painel de controle com dados extraídos de uma rede social, englobando todo o Brasil. Com isso, a instituição Fatec pode utilizar como referência as tecnologias para futuras atualizações nas grades curriculares. O projeto foi estruturado em um painel de controle dentro do Power BI, no qual os dados são extraídos por meio de um algoritmo de Web Scraping, desenvolvido em Python. Com os resultados adquiridos, podemos obter algumas percepções sobre as tecnologias que se destacam na área de Desenvolvimento e Análise de Dados.*

Abstract. *The project "Trend Analysis of IT Jobs on LinkedIn Using Web Scraping" aims to analyze vacancies in the job market through a dashboard with data extracted from a social network, encompassing all of Brazil. With this, the institution Fatec can use the technologies as a reference for future updates in the curriculum. The project was structured in a dashboard within Power BI where data is extracted through a Web Scraping algorithm developed in Python. With the results obtained, some insights can be extracted into technologies that stand out in the areas of Development and Data Analysis.*

1. Introdução

A Tecnologia da Informação (TI) é uma área do conhecimento, que engloba um conjunto de atividades e soluções, as quais envolvem o uso de recursos tecnológicos para processar, armazenar e transmitir informações. A informática, por sua vez, é uma área específica da TI que se dedica ao estudo e desenvolvimento de computadores e sistemas de informação. Em resumo, TI e informática, ambas, são dedicadas à utilização da tecnologia para lidar com informações de diversas naturezas, dentre estas informações

dados pessoais, informações de negócios, documentos, imagens, entre outros. Portanto, as áreas de TI e informática tonam-se necessárias para a gestão e organização de informações, aplicando no funcionamento de empresas, assim como em diversos outros setores da sociedade. De acordo com Carvalho e Souza (2020, p. 25),

> A Tecnologia da Informação (TI) é um conjunto de atividades e soluções que envolvem o uso de recursos tecnológicos para processar, armazenar e transmitir informações. A informática, por sua vez, é uma área específica da TI que se dedica ao estudo e desenvolvimento de computadores e sistemas de informação.

Ambos os autores também constam que: "TI e informática estão relacionadas à utilização de tecnologia para lidar com informações de diversas naturezas, sejam elas dados pessoais, informações de negócios, documentos, imagens, entre outros" (CARVALHO; SOUZA, 2020, p. 25).

Atualmente, o mercado de Desenvolvimento e Análise de Dados está em constante crescimento, esse crescimento é evidenciado pelo aumento de vagas que as empresas estão oferecendo a profissionais dessas. A evolução acelerada do uso da tecnologia, com a popularização da internet, biometria, reconhecimento facial, internet das coisas e inteligência artificial, tem levado empresas a analisar dados gerados pelos usuários, obtendo insights para fornecer melhores serviços e vantagens competitivas.

Na atuação do profissional de TI, é necessário estar constantemente atualizado sobre as novidades em áreas como, linguagens, *frameworks*, bibliotecas e plataformas de serviços. O objetivo deste trabalho é realizar uma análise de vagas no LinkedIn na área de Tecnologia da Informação (TI) para os cargos de Desenvolvimento e Análise de dados, diagnosticando quais são as tecnologias e *frameworks* que estão sendo mais solicitados aos candidatos, utilizando a técnica de *Web Scraping* para adquirir os dados para análise. Para isso, serão utilizados códigos em *Python* com as bibliotecas *BeautifulSoup* e *Selenium* para coletar informações sobre vagas de emprego em TI no LinkedIn. Os dados coletados serão processados e analisados, visando identificar quais são as tecnologias e *frameworks* mais requisitados pelos empregadores. O objetivo da desta análise, é que seus resultados possam auxiliar os profissionais de TI em sua busca por emprego, assim como contribuir para uma melhor compreensão das tendências do mercado de trabalho na área de tecnologia da informação. O código e os dados obtidos serão armazenados em um repositório no GitHub https://github.com/GkonishiC4/TCCv1 para que possam ser acessados e replicados por outros pesquisadores interessados no assunto.

Durante a primeira seção, a Introdução (1), aborda como a Tecnologia da Informação (TI) e a informática são importantes para gerenciar e organizar informações, como é notável em empresas e em outras esferas da sociedade. No decorrer do tópico subsequente, o Referencial Teórico (2), explica alguns conceitos relevantes, como análise de dados, Web Scraping, LinkedIn como fonte de dados e a importância do *Power BI* para visualizar dados. A partir da terceira parte, Material e Métodos (3), é detalhado as tecnologias que foram usadas, como Python, as bibliotecas BeautifulSoup, Requests e Pandas, além do processo de ETL. Em sequência, na quarta seção, Resultados e Discussões (4), é enriquecida com figuras ilustrativas, como o pipeline do algoritmo, o pipeline do ETL e os Dashboards das análises, para facilitar a compreensão dos processos e dos resultados obtidos. Na seção seguinte, Considerações Finais (5), é explicado como os resultados obtidos e as limitações do estudo das vagas no LinkedIn na área de TI podem

ajudar os profissionais do setor e melhorar a compreensão das tendências do mercado de trabalho em TI. Assim como, é reforçada a importância dos resultados e mencionam o repositório no GitHub onde o código e os dados estão disponíveis para acesso e replicação. Por último, a sexta parte, Referências (6), lista as fontes que foram usadas para escrever o artigo.

2. Referencial Teórico

O processo de análise de dados é fundamental para várias áreas do conhecimento, como negócios, finanças, marketing, saúde, ciência e tecnologia. Através da análise de grandes quantidades de dados, é possível descobrir padrões, tendências, insights e informações relevantes que podem ser utilizadas para tomar decisões estratégicas. Existem diversas técnicas de análise de dados disponíveis, como análise estatística, mineração de dados, aprendizado de máquina e visualização de dados. (SILVESTRE,2007)

Uma técnica importante para a coleta de dados é o *Web Scraping*, que permite a extração de informações de diversas fontes na internet de forma automatizada e rápida. Ferramentas e bibliotecas como *Beautiful Soup*, *Scrapy* e *Selenium* são amplamente utilizadas para a implementação de *Web Scraping* em diversos campos, incluindo negócios, finanças, marketing e tecnologia.

O LinkedIn é uma rede social de negócios que se tornou uma importante fonte de dados para a análise de mercado de trabalho. Através desta rede, é possível coletar informações sobre as habilidades, experiências e formações de profissionais de diversas áreas, além de ter acesso a vagas de emprego postadas por empresas.

O mercado de trabalho em tecnologia é altamente dinâmico e está em constante evolução, como consequência as empresas sempre buscam profissionais qualificados em áreas como programação, banco de dados, redes, segurança da informação, inteligência artificial, entre outras. Analisando os dados pode-se obter *insights* valiosos sobre as tendências e padrões desse mercado, auxiliando na tomada de decisão sobre quais tecnologias e habilidades são mais demandadas.

O *Power BI* é uma plataforma de análise de dados da Microsoft que permite a visualização e compartilhamento de insights de dados de maneira interativa e intuitiva. Essa ferramenta pode ser integrada a diversas fontes de dados, incluindo bancos de dados, planilhas e arquivos em nuvem, e é útil para a criação de dashboards e relatórios. A utilização de ferramentas como o *Power BI* pode auxiliar na análise e interpretação de dados em diversas áreas, contribuindo para a tomada de decisões mais informadas e assertivas.

3. Materiais e Métodos

No desenvolvimento deste trabalho, alguns recursos na linguagem *Python* podem ser bastante úteis. O *Web Scraping* é uma técnica que pode ser usada para coletar dados relevantes em sites específicos para análise. Para isso, pode-se utilizar a biblioteca *BeautifulSoup* para fazer a análise e extração de dados em documentos HTML. Por sua vez, *ETL* é uma tecnologia que permite integrar dados de diferentes fontes para análise e processamento. Pode-se utilizar a biblioteca Pandas para trabalhar com esses dados, que

permite manipular e analisar dados de forma simples e eficiente. No intuito de realizar as requisições necessárias na web, pode-se utilizar a biblioteca *Requests*, que permite fazer requisições HTTP de forma simples e eficiente.

Para automatizar a interação com sites, pode-se utilizar a biblioteca *Selenium*, que permite simular interações do usuário com a página, como clicar em botões, preencher formulários, entre outras ações. Por fim, o *GitHub* pode ser utilizado para compartilhar e gerenciar o código do projeto em desenvolvimento, permitindo que múltiplos colaboradores trabalhem juntos e facilitem a gestão do versionamento do código.

3.1. Tecnologia Escolhida

O Python é uma das linguagens de programação mais populares e eficientes para realizar *Web Scraping*. De acordo com Kharade e Kharade (2021), "*Python* é uma das linguagens de programação mais populares usadas para *Web Scraping*, e isso se deve principalmente à facilidade de aprendizado, à grande variedade de bibliotecas e ferramentas disponíveis e à sua flexibilidade" (p. 12). Isso se deve ao fato de possuir diversas bibliotecas que auxiliam no desenvolvimento e ser uma linguagem que permite um aprendizado rápido comparado a outras linguagens, como o *Java*, por exemplo. Além disso, o *Python* é uma linguagem de programação interpretada, isso significa que o código é executado linha por linha e pode ser facilmente modificado e depurado, esse processo é relevante no processo de *Web Scraping*. A sua sintaxe é simples e acessível, permitindo que os programadores se concentram na lógica do algoritmo, focando menos na implementação. Por fim, o *Python* possui uma grande comunidade de desenvolvedores e usuários, o que significa que existem muitos recursos e suporte disponíveis para quem está começando a trabalhar com *Web Scraping*.

3.2. *Web Scraping* (Raspagem de Dados)

O processo de *Web Scraping* em *Python* envolve a utilização de bibliotecas específicas para realizar a extração de dados de uma página da web. Existem diversas bibliotecas em Python para realizar o *Web Scraping*, mas as mais populares são o *BeautifulSoup, Requests* e *Selenium*. O processo básico de *Web Scraping* em *Python* envolve os seguintes passos:

a) Enviar uma solicitação HTTP para a página que deseja extrair dados usando a biblioteca *Requests*.
b) Receber a resposta da página com o conteúdo HTML usando a biblioteca *Requests*.
c) Analisar o conteúdo HTML da página com o *BeautifulSoup* para extrair os dados necessários.
d) Salvar os dados extraídos em um arquivo ou banco de dados para posterior análise ou processamento.

O processo pode ser mais complexo dependendo da página que se utiliza para extrair dados, como por exemplo, quando a página possui dados dinâmicos que são

carregados somente após a interação com o usuário. Nesse caso, é necessário utilizar a biblioteca *Selenium* para simular a interação do usuário e obter os dados desejados.

O *Web Scraping* em *Python* é uma técnica essencial para extrair dados de páginas da web quando não há uma *API* disponível, assim como em casos que os dados precisam ser obtidos de várias fontes diferentes. No entanto, é importante respeitar as políticas de privacidade e termos de uso das páginas da web para evitar problemas legais (ÁLVAREZ,2007).

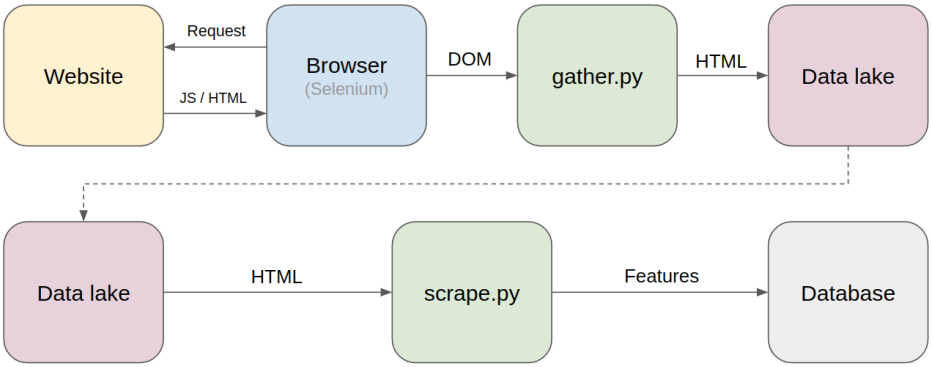

Figura 1. PipeLine Algoritmo

Fonte: (Juha Kiili, 2022)

3.3. Processos de ETL

O processo ETL (*Extract*, *Transform* e *Load*) possui como objetivo extrair dados de uma ou mais fontes de dados, transformá-los para uma forma adequada e carregá-los em um destino, como um banco de dados ou um *data Warehouse*. Conforme Kimball et al. (2013, p. 76),

> O processo ETL, ou *Extract, Transform e Load*, é uma técnica amplamente utilizada para mover dados de uma ou mais fontes de dados para um destino, como um banco de dados ou um data Warehouse. O processo ETL é composto por três etapas principais: extrair os dados da(s) fonte(s), transformá-los para uma forma adequada e carregá-los no destino.

Cada uma das três fases desse processo tem uma função específica:

a) *Extract* (extração): nesta fase os dados são extraídos da(s) fonte(s) de dados de origem, como arquivos, bancos de dados, APIs, entre outros.
b) *Transform* (transformação): nesta fase os dados extraídos são transformados para um formato adequado para a análise, incluindo limpeza, agregação, cálculos, formatação, entre outras operações.

c) *Load* (carga): nesta fase os dados transformados são carregados em um destino, como um banco de dados ou um *data Warehouse*, onde ficarão disponíveis para consulta e análise.

O processo de ETL é relevante para garantir a qualidade e a integridade dos dados utilizados para análise e tomada de decisões. Tal processo permite que os dados sejam limpos, padronizados e organizados de forma apropriada antes de serem armazenados em um *data Warehouse* ou usado para análise. Isso ajuda a garantir a precisão e a confiabilidade das análises e relatórios gerados a partir desses dados.

Além disso, o processo de ETL pode ajudar a integrar dados de diferentes fontes, permitindo que informações de várias áreas da organização sejam combinadas e analisadas juntas. Isso pode levar a insights mais profundos e a uma melhor compreensão das operações da organização como um todo.

Em resumo, o processo de ETL é fundamental para garantir a qualidade e a integridade dos dados utilizados para análise e tomada de decisões, permitindo que informações de diferentes fontes sejam combinadas e analisadas juntas para insights mais profundos.

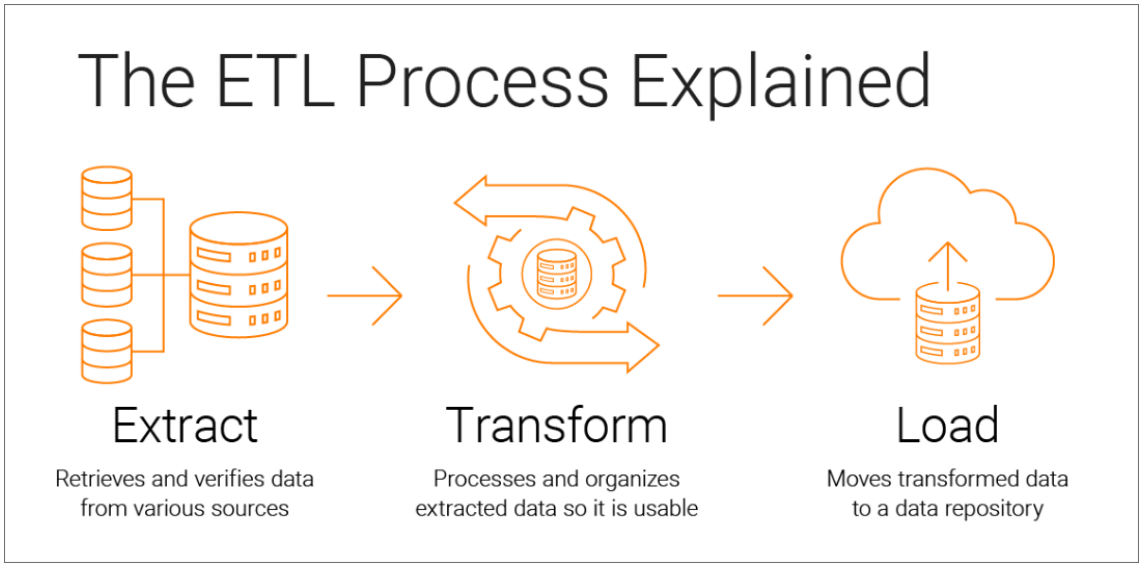

Figura 2. PipeLine ETL

Fonte: (Vijay Sharma, 2022)

3.4. Bibliotecas Utilizadas

3.4.1 *BeautifulSoup*

A biblioteca *BeautifulSoup* é uma das principais ferramentas utilizadas para fazer *Web Scraping* em *Python*. Esta biblioteca permite extrair informações de documentos HTML e XML de forma simples e eficiente, através de métodos que permitem navegar na estrutura do documento e buscar por elementos específicos, como *tags* e classes *CSS*. Além disso, a *BeautifulSoup* possui recursos para manipulação e formatação de texto, tornando-a uma biblioteca bastante versátil para processamento de dados web. (*Beautiful Soup Documentation*,2023).

3.4.2. *Requests*

A biblioteca *Requests* é uma biblioteca *Python* que permite enviar solicitações HTTP/1.1 com facilidade. A *Requests* simplifica o processo de fazer solicitações web e oferece uma série de recursos úteis, como gerenciamento de sessões, cookies e autenticação básica. Com a biblioteca *Requests*, os desenvolvedores podem facilmente recuperar dados de APIs e sites da web sem precisar lidar com a complexidade do protocolo HTTP. Além disso, a biblioteca possui uma documentação clara e bem-organizada, o que a torna fácil de usar e aprender. (Requests: HTTP for Humans™, 2022)

3.4.3. Pandas

A biblioteca Pandas é uma das mais populares em *Python* para análise de dados, fornecendo estruturas de dados flexíveis e eficientes para lidar com tabelas e séries temporais. O Pandas permite que os usuários manipulem dados em uma variedade de formatos, como CSV, *Excel* e *SQL*. Ele também é útil para tarefas como limpeza de dados, agregação, filtragem, fusão e transformação. Com o Pandas, os usuários podem realizar operações complexas de forma fácil e eficiente, tornando-o uma ferramenta essencial para análise de dados em *Python*. (*Pandas Guide*,2023).

3.4.4. Selenium

A biblioteca *Selenium* é uma ferramenta que permite automatizar a interação de um programa com um navegador web. Ela pode ser utilizada para testes automatizados, preenchimento de formulários, navegação em sites complexos, entre outras aplicações. Através da API do *Selenium* em *Python*, é possível acessar todas as funcionalidades do navegador de forma intuitiva e realizar ações como clicar em botões, preencher campos, navegar em páginas, além de extrair informações de elementos *web*. A biblioteca é bastante útil para projetos que envolvem *Web Scraping* em sites que utilizam tecnologias dinâmicas como *AJAX* e *JavaScript*.

3.5. *GitHub*

O repositório https://github.com/GkonishiC4/TCCv1 contém os códigos e dados utilizados neste trabalho de conclusão de curso (TCC) de *Web Scraping*, o qual teve como objetivo realizar uma análise de vagas na área de tecnologia da informação no LinkedIn. Neste repositório, é possível encontrar os códigos em Python que foram desenvolvidos utilizando as bibliotecas *BeautifulSoup* e *Selenium* para coletar informações sobre vagas de emprego na plataforma LinkedIn. Além disso, o repositório contém os dados brutos e processados em formato CSV que foram utilizados para a análise dos resultados. O código foi desenvolvido de forma clara e bem estruturada, permitindo que outros pesquisadores possam facilmente replicar a metodologia utilizada neste trabalho. O repositório também inclui um arquivo *README* com instruções detalhadas sobre como utilizar os códigos e dados, assim como as políticas de privacidade e ética na coleta de dados que foram adotadas durante todo o processo de *Web Scraping*. Dessa forma, o repositório se torna uma importante fonte de informação para pesquisadores interessados no assunto, permitindo a replicação e validação dos resultados obtidos neste trabalho.

4. Resultados e Discussões

Como apresentado na figura 3 pode-se visualizar que a área de Desenvolvimento possui uma quantidade maior de vagas ofertadas pelo fato que possui profissionais que trabalham apenas com um tipo de linguagem para realizar o Desenvolvimento de sistemas, softwares, *front e back end*. Sendo a vaga para assistente que tem posse de 58,67% (1944) das vagas analisadas (3313) sendo a região sudeste e sul que mais possuem ofertas de vagas, o setor de Serviços e Consultoria de TI possui a maiores quantidades de vagas. Destacam-se as tecnologias para a área de Desenvolvimento (*SQL, REST, .NET, PYTHON, HTML, NBSP, JAVA TYPESCRIPT, JAVASCRIPT, JAVA, CSS, C e PHP*).

Já na imagem 4 pode-se visualizar que a área de Análise de Dados possui uma quantidade menor de vagas ofertadas, devido ao fato de possuir uma maior segmentação de profissionais na área de dados como, analista de dados, engenheiro de dados e cientistas de dados. Sendo a vaga para os cargos Pleno e Sênior que juntos tem posse de 26,73% (250) das vagas analisadas (935) possuindo um equilíbrio maior pelas regiões que possuem ofertas de vagas, o setor de Serviços e Consultoria de TI dominado as quantidades de vagas. Destacam-se as tecnologias para a área de Desenvolvimento (*SQL, R, PYTHON, DAX, AWS e ORACLE*).

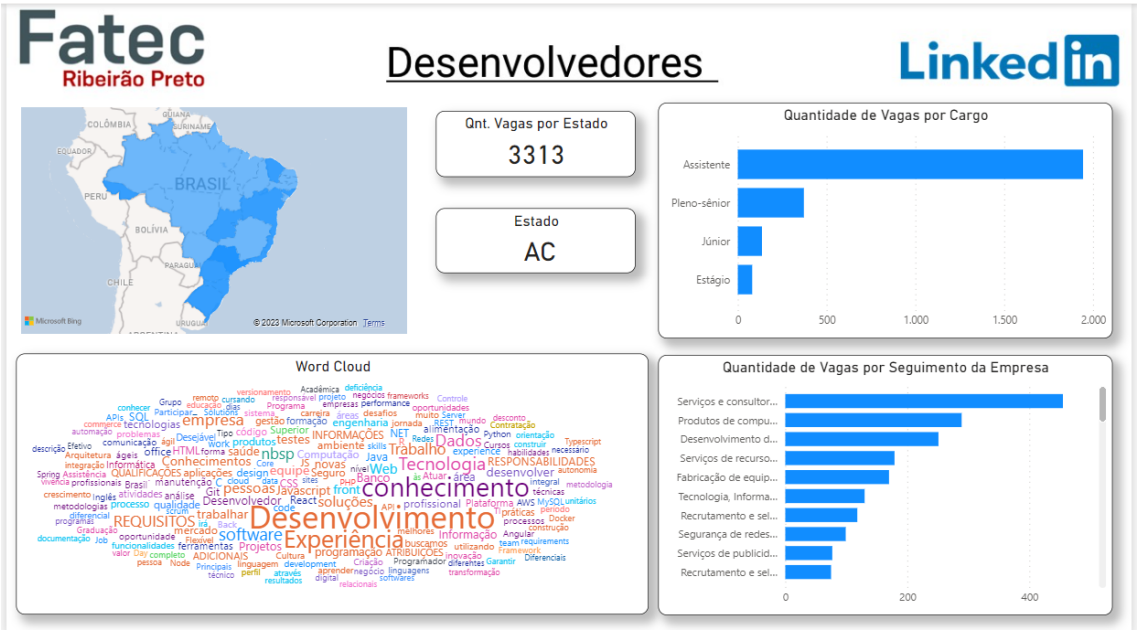

Figura 3. Dashboard Desenvolvedores

Fonte: (Autoria Própria, 2022)

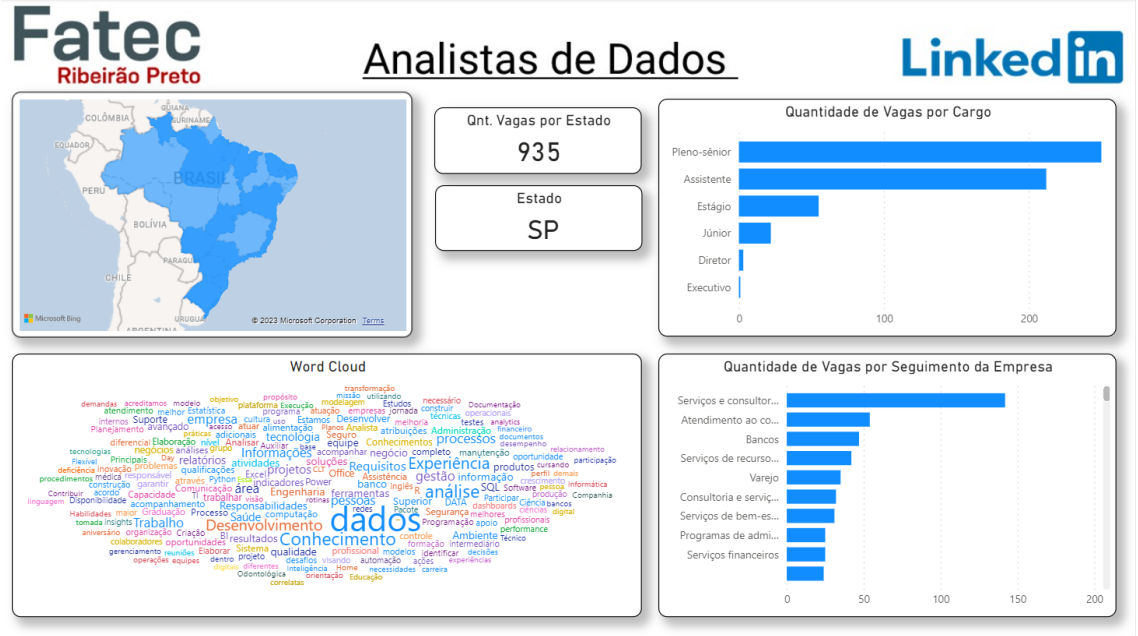

Figura 4. Dashboard Analistas de Dados

Fonte: (Autoria Própria, 2022)

5. Considerações Finais

A pesquisa realizada teve como objetivo obter informações sobre o mercado de trabalho na área de Tecnologia da Informação, identificando as habilidades e tecnologias mais requisitadas, bem como as especificações e oportunidades disponíveis no mercado. Após extensa e detalhada levantamento de dados, foram coletadas aproximadamente 4000 vagas de emprego no LinkedIn para realizar a análise, entre os meses de setembro e outubro de 2022, os dados foram compilados e apresentados em um *dashboard*. A análise dos resultados revelou que as *soft skills* estão se tornando cada vez mais importantes para os cargos de estágio e assistente, pois as empresas buscam moldar os profissionais de acordo com suas necessidades. Para os cargos de pleno e sênior, é fundamental ter um bom domínio técnico, além de certificações e especializações. A pesquisa é útil para os profissionais de Análise e Desenvolvimento de Sistemas, pois fornece informações relevantes para o crescimento de suas carreiras. Durante o processo, encontramos alguns desafios com o algoritmo devido a uma alteração que o LinkedIn fez em seu código fonte, impedindo-nos de continuar a coletar os dados. Tentamos contornar a situação utilizando a *API* do LinkedIn, mas essa opção só está disponível para empresas que assinam um serviço na parte de negócios que ela oferece como um serviço. Uma sugestão de melhoria seria ampliar a pesquisa para outras plataformas, além do LinkedIn, para obter uma visão mais ampla do mercado.

6. Referências

ÁLVAREZ, Alberto Cáceres. **Extração de informação de artigos científicos:** uma abordagem baseada em indução de regras de etiquetagem. 2007. Tese de Doutorado. Universidade de São Paulo.

CARVALHO, A. B. A.; SOUZA, F. R. S. Introdução à Tecnologia da Informação. 1. ed. São Paulo: Novatec Editora, 2020.

SILVESTRE, António. Análise de dados e estatística descritiva. Escolar editora, (2007).

KHARADE, G.; KHARADE, R. Web Scraping using Python: A Beginner's Guide to Data Extraction. Birmingham: Packt Publishing, (2021).

KIMBALL, R.; ROSS, M.; THORNTHWAITE, W.; MAUSSANG, S. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. 3rd ed. Indianapolis: Wiley Publishing, (2013).

OLIVEIRA, F. A. Web Scraping: Coletando Dados na Web com Python. São Paulo: Novatec Editora, (2021).

Bibliotecas

Beautiful Soup Documentation. Documentação BeautifulSoup. Disponível em: https://www.crummy.com/software/BeautifulSoup/bs4/doc/ Acesso em: 03 Mar. 2023.

PANDA GUIDE. Documentação Pandas. Disponível em: https://pandasguide.readthedocs.io/en/latest/ .Acesso em: 03 Mar. 2023.

Requests: HTTP for Humans™. Documentação Requests. Disponível em: https://requests.readthedocs.io/en/latest/ . Acesso em: 03 Mar. 2023.

Selenium with Python. Documentação Selenium. Disponível em: https://selenium-python.readthedocs.io./ Acesso em: 03 Mar. 2023.