CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA

FACULDADE DE TECNOLOGIA DE INDAIATUBA

DR. ARCHIMEDES LAMOGLIA

CURSO DE TECNOLOGIA EM

ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

Yuri Rafael Cavalcanti Queriquelli

Utilizando Big Data e Web Scraping para busca de vagas de emprego em Campinas e região

INDAIATUBA
2020

CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA

FACULDADE DE TECNOLOGIA DE INDAIATUBA

DR. ARCHIMEDES LAMOGLIA

CURSO DE TECNOLOGIA EM

ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

Yuri Rafael Cavalcanti Queriquelli

Utilizando Big Data e Web Scraping para busca de vagas de emprego em Campinas e região

Projeto de Trabalho de Graduação apresentado por Yuri Rafael Cavalcanti Queriquelli como pré-requisito parcial para a conclusão do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas, da Faculdade de Tecnologia de Indaiatuba, elaborado sob a orientação do Profa. Dra. Maria das Graças Junqueira Machado Tomazela

INDAIATUBA
2020

RESUMO

Hoje a internet é a principal ferramenta para busca de vagas de emprego na região de Campinas, porém as oportunidades estão disseminadas em diversos canais e sites tornando a tarefa de buscar complicada e exaustiva. Para centralizar uma quantidade massiva de informações existe o conceito de *Big Data*, e para realizar a coleta desses dados existem ferramentas como *web scraping* e *web Crawling*. Assim, este trabalho tem como objetivo criar um *Big Data* para centralizar as informações referentes a vagas de emprego na região metropolitana de Campinas, utilizando *Web Scraping* e *Web Crawler*.

. Na fundamentação teórica são apresentados os principais conceitos sobre *Web Scraping* e *Big Data* que sustentam a pesquisa os quais são ancorados em práticas de pesquisas apresentadas em um conjunto de trabalhos relacionados. Para alcançar os objetivos propostos, será realizada uma pesquisa experimental, que consiste na determinação de um objeto de estudo, na seleção de variáveis que sejam capazes de influenciá-lo, e na definição dos meios para controlar e observar os efeitos que esta variável manipulada possa produzir nesse objeto. Assim, com a documentação, apresentação e análise dos dados, espera-se que a ferramenta de *Web Scraping* seja capaz de coletar dados de diversos sites, organizar um *Big Data* e apresentar em um site com as vagas de emprego de Campinas e região.

Palavras-chave: Web Scraping; Big Data; Web Crawling;

SUMÁRIO

INTRODUÇÃO

De acordo com o IBGE, Instituto Brasileiro de Geografia e Estatística, no ano de 2019 o Brasil possuía mais de 12,8 milhões de trabalhadores em busca de uma oportunidade no mercado de trabalho[1]. Na região metropolitana de Campinas esse número está próximo dos 220 mil trabalhadores.

Em 2020 a economia mundial enfrenta um grande desafio por causa da pandemia do COVID-19 que possivelmente aumentará a quantidade de pessoas em situação de desemprego. Segundo o FMI[2], Fundo Monetário Internacional, os impactos da recessão econômica deste ano serão comparados aos da "Grande Depressão" também conhecida como "Crise de 1929" aonde o sistema capitalista passou pelo pior e o mais longo período de recessão econômica com queda de 15% no PIB mundial. No Brasil as perspectivas são de queda 5,3% na economia e queda de 11% no comércio mundial.

Com o aumento eminente de desempregados existem diversas e novas ferramentas que os trabalhadores possuem para auxiliar na procura de um emprego, por muitas vez a internet é a mais rápida fonte de informações e oportunidades, porém quase nunca as vagas de emprego estão centralizadas, assim o trabalho de busca é exaustivo e muito mais complicado do que o necessário.

Diante desse cenário as pessoas que conseguem encontrar vagas, oportunidades e informações confiáveis, conseguem aumentar sua competitividade e reduzir o tempo de busca nos diversos portais de emprego.

Sobre esse contexto o *Big Data* provê a tecnologia destinada a extrair valor de uma imensa variedade de dados, o que permite alta velocidade com objetivo de capturar, descobrir e analisar estas informações e dados, de forma a transformá-los em informações importantes e valiosas (MACHADO, 2018).

Para realizar a captura de dados a ferramenta de *Web Scraping* fornece as funções essenciais de extrair e reunir conjuntos de dados da web, dados esses que são fundamentais para o *Big Data Analytics*. Outras tecnologias como *Machine Learning* e Inteligência Artificial também podem utilizar dessa ferramenta para coleta de material para análise. (MITCHEL, 2015)

---

[1] Disponível em: <https://exame.com/economia/brasil-tem-desemprego-de-126-no-trimestre-ate-abril-diz-ibge/ > Acesso em: 11 mai. 2020.

[2] Disponível em: <https://oglobo.globo.com/economia/economia-mundial-deve-ter-maior-recessao-desde-grande-depressao-preve-fmi-24369966> Acesso em: 02 jun. 2020.

A partir desse cenário, a pergunta de pesquisa que orienta este trabalho é:

Como as ferramentas de *Big Data* podem permitir a descoberta de novos conhecimentos referentes a ofertas de trabalho na região metropolitana de Campinas?

Assim, para responder essa pergunta de pesquisa o objetivo deste trabalho é criar um *Big Data* para centralizar as informações referentes a vagas de emprego na região metropolitana de Campinas, utilizando *Web Scraping* e *Web Crawler*.

A hipótese é que se for realizado um processo de análise de *Big Data* com *Web Scraping* como fonte de coleta dos dados será possível a centralização de vagas de emprego na região de Campinas, o que pode auxiliar os cidadãos na busca de novos postos de trabalho.
A metodologia para o desenvolvimento de trabalho será a pesquisa experimental, que tem como finalidade testar hipóteses que dizem respeito a convicção do pesquisador, envolvendo grupos de controle, seleção aleatória e manipulação de variáveis. Selecionado as variáveis que têm a capacidade de influenciá-lo e então os efeitos causados (GIL, 2007).

Para organização do trabalho a estrutura de capítulos será utilizada na sequência descrita:

No Capítulo I, serão apresentados os conceitos chaves que apoiam o desenvolvimento deste trabalho: *Big Data*, *Web Scraping* e *Web Crawling* e as pesquisas relacionadas a essa.

Para o Capítulo II, serão apresentados os caminhos percorridos para realizar este trabalho e o cronograma de entregas para a próxima etapa.

CAPÍTULO I

Fundamentação Teórica

1.1. Conceitos chave

No embasamento desta pesquisa, optou-se por organizar este capítulo em duas partes. Primeiramente apresentam-se os conceitos chave que referenciam o trabalho, sendo eles: *Big Data*, *Web Scraping* e *Web Crawling*. Na segunda parte, apresenta-se um conjunto de trabalhos relacionados a esta pesquisa, decorrentes de estudos realizados nos últimos anos.

1.1.1 – Big Data

O conceito de *Big Data* pode ser definido como grandes volumes de dados com origem em diversas fontes, como aplicativos da web, sites, portais, redes sociais e em formatos diversificados como texto, áudio, vídeo, imagens, diagramas entre outros (MUTHUKUMAR (2014); SIN e MUTHU (2015)).

Big Data também é identificado por um conjunto de características fundamentais, sendo elas:

- Volume – Quantidade massiva de dados, demandando uma grande estrutura para armazenamento, processamento, transferência e análise.
- Velocidade – Taxa crescente que as informações são armazenadas.
- Veracidade – Diz respeito a como os dados serão armazenados e extraídos. Com fina relação a confiança e incerteza.
- Variedade – Corresponde aos diversos formatos, tamanhos e fontes dos dados, sendo eles estruturados ou não.
- Verificação – Relaciona-se à segurança e verificação de dados.

- Valor – Considerada a mais importante, pois diz respeito aos resultados do processo de *Big Data*, essa característica tem pode ser definida como o valor gerado após o sistema ser implementado.

Além dessas propriedades, há três etapas necessárias para que o Big Data agregue valor às atividades de gestão (DANIEL, 2015).

- Coleção – A coleta de dados é primeiro passo para obter valor a partir do *Big Data*. Isto exige a identificação de dados que podem revelar informações úteis e valiosas. Os dados devem filtrados e só depois armazenados de forma que seja útil para a tomada de decisão.
- Análise – Uma vez que os dados foram organizados em uma forma utilizável, eles devem ser analisados. No entanto, com a crescente diversidade na natureza dos dados, o gerenciamento e a análise desses conjuntos de dados diversificados está se tornando um processo muito complexo. A análise precisa incluir vinculação, correlacionando diferentes conjuntos de dados para que seja possível entender a informação que deve ser transmitida por esses dados.
- Visualização e aplicação - Nessa etapa, os dados analisados são disponibilizados aos usuários em uma forma que é interpretável e integrado nos processos existentes e, em última instância, usados para orientar nas tomadas de decisão.

Apresentado assim as características de um *Big Data* é importe salientar que existem duas entidades técnicas. Primeiro a quantidade maciça de informações detalhadas. Em segundo lugar, análises avançadas, utilizando diferentes tipos de ferramentas.

1.1.2 – Web Scraping

A ferramenta *Web Scraping* permite o *download* automatizado de dados de uma página web e extrai informações específicas (MITCHELL, 2015). Os dados coletados podem ser exportados para arquivos ou armazenados em banco de dados. A literatura especializada também referencia a ferramenta como "Extração de Dados da Web" e os programas que realizam a tarefa são conhecidos como scrapers. Sua característica principal é a capacidade de trazer informações detalhadas das páginas web. Para construção de um *Big Data* essa ferramenta permite a coleta de textos, imagens e dados específicos para desenvolvimento análise futuras. A figura 1 apresenta o fluxo de funcionamento de um *Web Scraping*

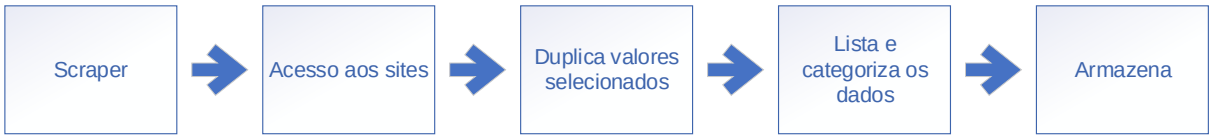

Figura 1: Esquema de funcionamento do *Web Scraping*

Fonte: Adaptado de https://www.edureka.co/blog/web-scraping-with-python/

1.1.3 – Web Crawling

A ferramenta *Web Crawling* permite o *download* automático de dados de uma página web, extraindo hiperlinks contidos (MITCHELL, 2015). Os dados baixados são, em grande parte, armazenados em um banco de dados para facilitar a busca. *Web Crawling* também é conhecido como indexação e tem amplo uso em sites de busca como Google, Yahoo e Bing.

Em geral os *Web Crawlers,* como são chamados os *bots* que realizam a captura, trazem informações genéricas sobre as páginas, essa é principal diferença entre a ferramenta e o *Web Scraping* (GABARDO, 2018). Utilizando as coletas realizadas por esses sistemas é possível mapear todos os sites que possuem links entre si e montar um quadro geral da internet inteira.

Em uma análise de *Big Data* essa ferramenta tem como função criar referência entre as fontes de dados e mitigar as duplicidades. A figura 2 demonstra o esquema de funcionamento de um *Web Crawling.*

Figura 2: Esquema de funcionamento do *Web Crawling*

Fonte: Adaptado de https://www.edureka.co/blog/web-scraping-with-python/

1.2 – Trabalhos relacionados

Nesta seção são apresentados estudos e pesquisas relacionados ao tema deste trabalho.

O levantamento realizado foi orientado pela busca de pesquisas científicas ou tecnológicas que têm em seus objetivos o desenvolvimento ou análise de ferramentas e tecnologias para a aplicação de *Web Scraping* e análise de *Big Data.*

A ferramenta que serviu de referência para isso foi o Google Acadêmico por meio do qual se buscou mapear as pesquisas dessa natureza circunscritas nos últimos anos.

Boeing e Waddell (2016) utilizaram Web Scraping nas ofertas de imóveis para aluguel nos Estados Unidos da América para analisar padrões dos mercados imobiliários metropolitanos de pequeno porte. Como parte do desenvolvimento foram coletados mais de onze milhões de ofertas de aluguel utilizando Web Scraping e técnicas de Big Data para classificar e filtrar os dados. Ao final foi apresentado um mapa demográfico nacional com custo médio mensal do aluguel e quantidade em oferta, permitindo comparações precisas entre estados e cidades. Concluindo o projeto foi possível compreender melhor a atividade do mercado de aluguéis em escala nacional, regional e local.

Haddaway (2016) utilizou *web scraping* para compilar, de maneira centralizada, informações públicas governamentais na Suécia, conhecidas como *Grey Literature*. As tecnologias aplicadas para a coleta de dos dados foram *Helium Scraper* para simular um usuário, *WebSundew* e *FMiner* para *scraping*. Ao finalizar o sistema os dados antes invisíveis para o grande público foram reunidos e classificados para compartilhamento democrático das informações. O autor também disponibilizou os códigos para que outros interessados possam fazer o mesmo em seus países.

Utilizando *web scraping* Polidoro *et al.* (2016) analisaram os preços ao consumidor com referência específica a produtos eletrônicos de consumo (mercadorias) e tarifas aéreas (serviços). O trabalho utilizou como ponto de partida o levantamento realizado pelo *Italian National Statistical Institute* (Istat), com contexto europeu, e seu objetivo foi modernizar o processo de coleta para tornar o levantamento de dados mais preciso, dinâmico e eficiente. Nos testes realizados a ferramenta de *web scraping* desenvolvida em *Python e* apresentou dados mais confiáveis que o modelo atual.

Dimitrov (2016) publicou um estudo sobre como *Big Data* e *IOT, Internet of Things*, podem auxiliar no combate de doenças e possíveis pandemias. Os dados a serem analisados passariam entre especialistas de tecnologia, médicos, pacientes e o público interessado. Os dados podem ser coletados de equipamentos médicos como um monitor multiparâmetro e de equipamentos de uso diário como celulares e relógios. O autor abre precedentes para a utilização das mesmas tecnologias em ambiente farmacêutico, que poderia facilitar o desenvolvimento de medicamentos, antibióticos, vacinas, sedativos entre outros.

Najada e Mahgoub (2016) utilizaram *Big Data* para analisar o trânsito do estado de Flórida nos Estados Unidos da América. O trabalho utilizou as bases de dados disponíveis ao público pelo governo local e ferramentas de análise estatísticas como *Naive Bayes, Random Foreste e AdaBoost.* O principal objetivo dos autores foi detectar e ajudar na prevenção de acidentes com automóveis nas rodovias estaduais. A conclusão do trabalho foi compilada em

pontos estratégicos de monitoramento e ações para redução efetiva de acidentes, o relatório final foi enviado aos governantes do estado como sugestão para melhorias e aprimoramento da ferramenta.

Chaulagain *et al.* (2017) desenvolveram uma ferramenta de Web Scraping em aplicações de Big Data baseadas em nuvem. As tecnologias utilizadas foram Elastic Compute Cloud, DynamoDB e selenium. O ambiente de execução foi o Amazon Web Services. Como conclusão os desenvolvedores abordaram o fato de ser mais fácil utilizar *Scraping* em nuvem por se tratar de um ambiente com melhor distribuição de dados. O projeto provou-se eficiente e de fácil adaptação em outras plataformas de aplicação em nuvem.

Utilizando *Web Scraping* e aprendizado de máquina, Slamet *et al.* (2018) melhoraram as classificações de empregos em sites de buscas. Durante o desenvolvimento foram utilizados Phyton para a ferramenta de *Web Scraping* e *Naïve Bayes* para classificação das vagas coletadas. Ao final foram criadas duas aplicações, sendo um portal para a visualização das vagas e um painel de controle para melhorias e correções administrativas. Para finalizar foram realizados testes de caixa preta que comprovaram a precisão das classificações das vagas.

Jen (2019) criou uma ferramenta automatizada de *Web Scraping* com o objetivo de facilitar o processo de coleta de dados turísticos na Malásia e atrair o foco das indústrias de turismo e do setor governamental para os dados públicos de turismo, assim melhorando o setor turístico no país. *Python* foi a linguagem de programação escolhida para escrever o programa, o resultado foi compilado em arquivos de extensão CSV *comma-separated-values*. Segundo o autor, com a realização desse projeto, os analistas de dados podem obter facilmente os dados necessários para o processo de análise de dados e, portanto, melhorar o turismo na Malásia. Como melhoria o autor sugeriu implementar outras maneiras de compilar os dados e exportá-los para outras extensões.

Para estudar como as alterações de leis locais que impactam os estudantes do estado de Virginal nos Estados Unidos da América, o trabalho da Anglin (2019) criou um Framework utilizando *Web Scraping* e processamento de linguagem natural. O sistema foi programado utilizando as técnicas de *deep learning* PNL, Processamento de Linguagem Natural, e linguagem de programação em *Python*. Ao concluir o projeto a autora criou uma base atualizada com políticas, manuais e atas de reunião do estado disseminando as informações e fornecendo aos formuladores de políticas evidências fortes e oportunas para ajudar na tomada de decisões.

Araujo (2019) usou a análise de *Big Data* para comprar o desempenho acadêmico dos discentes do curso de Análise e Desenvolvimento de Sistemas das FATECs Indaiatuba e Sorocaba. O autor utilizou os conceitos de *MapReduce* e *Big Data Analytics* para realizar sua análise e as ferramentas selecionadas foram *o Google Cloud Dataproc, Google BigQuery*, e o *Google Data Studio*. Os resultados mostraram que a média de notas dos vestibulares e a porcentagem de alunos provenientes de escolas públicas ou particulares são similares nas duas instituições

Conclui-se que dentre os dez trabalhos relacionados utilizados, três utilizam os conceitos de *Big Data* e sete fazem uso de *Web Scraping* para desenvolvimento de análises. Dentre os quais, os que mais se destacaram para auxiliar na elaboração deste trabalho foram os artigos de Slamet *et al.* (2018) e a aplicação Chaulagain *et al.* (2017). No Quadro 1 são referenciados o objetivo dos trabalhos utilizados e sua relação com essa pesquisa.

Quadro 1**:** *Trabalhos Relacionados*

Autor (Ano)	**Objetivo da pesquisa**	**Relação com esta pesquisa**	**Ferramentas utilizadas**	**Tipo**
Boeing e Waddell (2016)	Desenvolver e analisar de ofertas de aluguel utilizando Web Scraping e Big Data.	Utilização de técnicas de gerenciamento que possuem grande importância para o projeto.	• *Web Scraping* • *Big Data.*	Aplicação
Haddaway (2016)	Desenvolver uma ferramenta para centralizar *Grey Literature* na Suécia.	Amplia as tecnologias de extração ao utilizar *Helium Scraper.*	• *Helium Scraper* • *Web Scraping*	Aplicação
Polidoro *et al.* (2016)	Analisar e comprar preços de produtos e serviços na Itália.	Essa aplicação possui uma vasta variedade de ferramentas, procedimentos e métodos utilizados.	• *Python* • *Web Scraping*	Aplicação
Dimitrov (2016)	Falar sobre as técnicas de gerenciamento de *Big Data e IOT* no combate de doenças e possíveis pandemias.	Utilização de diversas e fontes de coletas de dados ao abranger a possibilidade de IOT.	• *IOT* • *Big Data*	Conceito
Najada e Mahgoub (2016)	Analisar grandes quantidades de informações para prevenção de acidentes de trânsito no estado da Flórida.	O trabalho explora os conceitos de *Big Data* em uma aplicação de grande e massiva quantidade de informações.	• *Naïve Bayes,* • *Random Foreste* • *AdaBoost*	Aplicação
Chaulagain *et al.* (2017)	Desenvolver uma ferramenta de Web Scraping em aplicações de Big Data baseadas em nuvem.	Este trabalho traz conceitos e ferramentas que servem de base para a aplicação.	• *Phyton* • *DynamoDB* • *Selenium*	Aplicação

Autor (Ano)	**Objetivo da pesquisa**	**Relação com esta pesquisa**	**Ferramentas utilizadas**	**Tipo**
Slamet *et al.* (2018)	Melhorar as classificações de empregos em sites de buscas.	Utiliza e amplia as possibilidades do *Web Scraping* adicionado análise com *Naïve Bayes.*	• *Web Scraping* • *Phyton* • *Naïve Bayes*	Aplicação
Jen (2019)	Facilitar o processo de coleta de dados turísticos na Malásia utilizando Web Scraping.	A ferramenta utilizada nesse trabalho é semelhante como a da proposta apresentada.	• *Phyton* • *CSV*	Aplicação
Anglin (2019)	Utilizar um Framework com Web Scraping para criar uma base atualizada com políticas, manuais e atas de reunião do estado.	Essa aplicação possui uma vasta explicação dos procedimentos e métodos utilizados, tornando-a significativa para a elaboração deste trabalho.	• Phyton • *Deep Learning* • Processamento de Linguagem Natural	Aplicação
Araujo (2019)	Comparar o desempenho acadêmico dos discentes do curso de Análise e Desenvolvimento de Sistemas das FATECs Indaiatuba e Sorocaba.	O artigo possui conceitos importantes que serão utilizados neste trabalho, e exibe uma aplicação com as ferramentas de *Big Data Analytics.*	• Google Cloud • Dataproc • Google BigQuery	Conceito

Fonte: Elaborado pelo autor

CAPÍTULO II

Metodologia

2.1– Natureza da Pesquisa

A metodologia a ser utilizada durante o desenvolvimento deste estudo será a **Pesquisa Experimental** que, segundo Gil (2002), consiste em determinar um objeto de estudo, selecionar as variáveis capazes de influenciá-lo e definir meios de controle e observação dos efeitos que esta variável produz neste objeto.

Para que tal método seja utilizada faz-se necessário a execução de algumas atividades, como descritas abaixo:

- Definição da hipótese;
- Concepção das regras ambientais e comportamentais para a execução do projeto;
- Execução do experimento;
- Análise dos resultados obtidos.

2.2– Variáveis de análise

Para que o desenvolvimento desse trabalho fosse possível, três variáveis foram definidas para possibilitar a coleta e análise de dados:

1. Sites e portais de referência
2. Dados de vagas de emprego na região de Campinas
3. Região das vagas de emprego

2.3– Padrões para pesquisa experimental

Nesta seção são apresentadas as principais referências de mercado relacionados ao tema deste trabalho.

- **Busca de vagas do Glassdoor**[3]

A *start-up* brasileira *Love Mondays* foi reformulada e renomeada como de *Glassadoor*. Novas ferramentas foram adicionadas ao sistema, em suma o site permite que os funcionários atuais e ex-funcionários avaliem anonimamente as empresas. O sistema também permite que os usuários enviem e visualizem anonimamente salários, além de pesquisarem e se candidatarem a empregos em sua plataforma.

Sua principal característica é a possibilidade de o candidato ter uma perspectiva de terceiros sobre as empresas além de permitir a comparação salarial de uma mesma vaga em diversas corporações.

O sistema é acessível a qualquer usuário, porém é necessário cadastro para visualizar e inserir informações. O sistema também permite o acesso das empresas ao sistema com um acesso especial, o modelo comercial possui vantagens como gerenciamento de marca, respostas a usuários e divulgação de vagas. A figura 3 apresenta a versão disponível para celulares Android, o aplicativo também está disponível para os candidatos nas nos celulares IOS.

[3] Disponível em: < **www.glassdoor.com.br** > Acesso em: 25 mai. 2020.

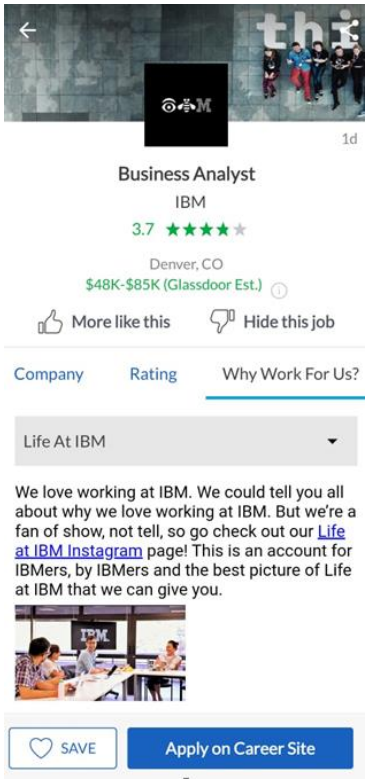

Figura 3: Página de uma vaga na versão android do site

Fonte: https://bit.ly/3dvGN3P

- **Vagas de Emprego e Oportunidades de Trabalho**[4]

O site oferece aos usuários um currículo virtual que pode ser alterado sem custos a qualquer momento. Para empresas o sistema possui uma página personalizada com informações de vagas, curso. Grandes companhias como IBM, Ypê, Unilever e Itaú, são parceiras do vagas.com e utilizam as soluções por ela oferecida.

O sistema também oferece provas para as empresas selecionarem previamente os candidatos antes das entrevistas.

Para os candidatos existem versões do sistema nos dispositivos móveis com Android e IOS. Na figura 4 apresenta-se a página de *e-partners* do vagas.com.

[4] Disponível em: < **www.vagas.com.br** > Acesso em: 25 mai. 2020.

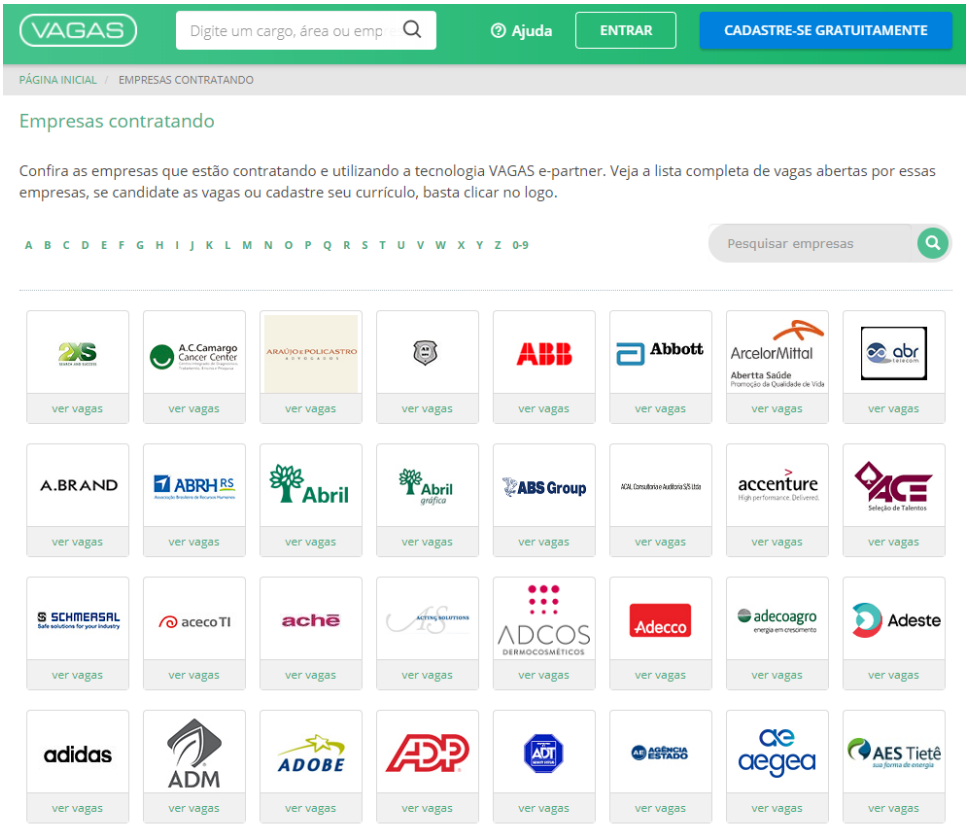

Figura 4: Página de e-partner do vagas.com

Fonte: https://www.vagas.com.br/empresas-contratando

- **SmartRecruiters | Recruiting Software & Talent Acquisition Suite**[5]

SmartRecuiters é o único aqui apresentado que possui de forma centralizada vagas nacionais e internacionais, além de oferecer um software para facilitar a seleção de candidatos e estagiários.

O site possui uma grande quantidade de parceiros comerciais, sendo em grande parte empresas multinacionais como Bosch, Visa, LinkedIn, Shell, entre outras. O sistema oferece diversos produtos e plataformas para empresas e tem nelas seu principal foco comercial. A figura 5 apresenta a página de oportunidades do Grupo Bosch no Brasil.

[5] Disponível em: < **https://www.smartrecruiters.com** > Acesso em: 25 mai. 2020.

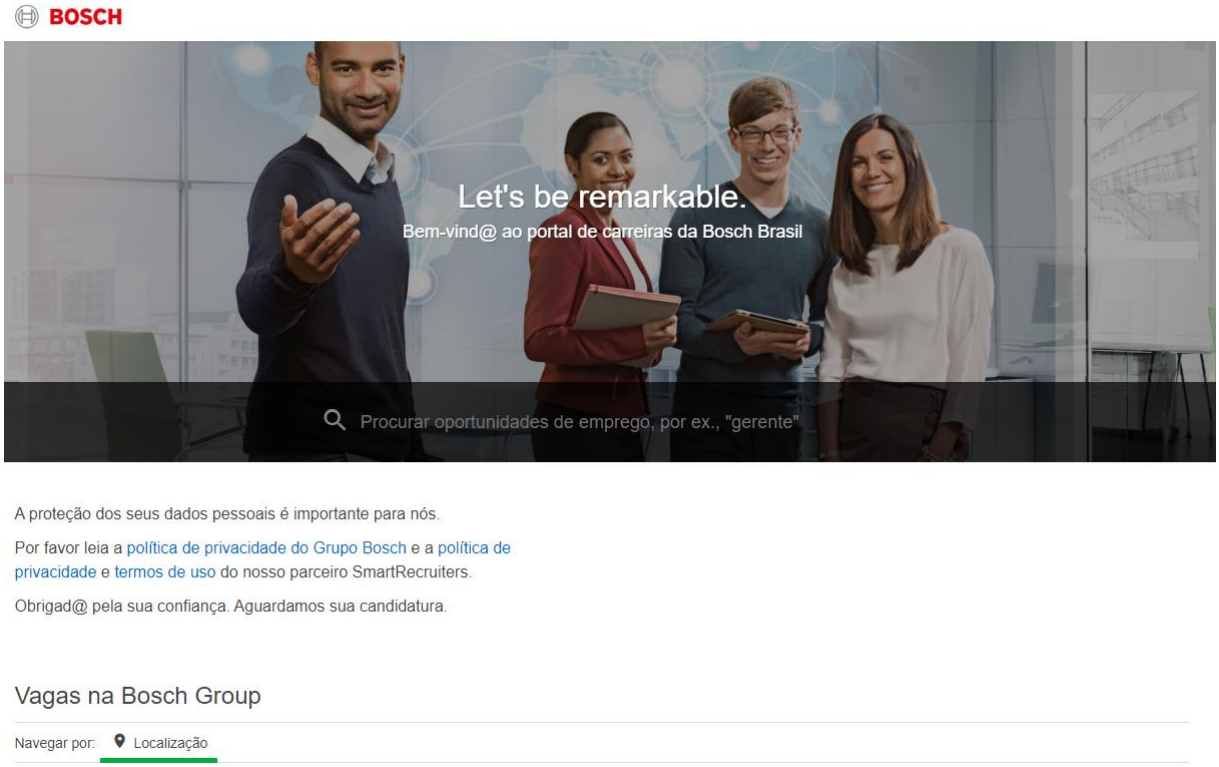

Figura 5: Página personalizada do Grupo Bosch

Fonte: https://careers.smartrecruiters.com/BoschGroup/

2.4– Experimento de Pesquisa

Para a realização deste trabalho serão utilizados uma série de passos que definirão o futuro da ferramenta. O primeiro passo será selecionar os sites e portais de referência, nesta etapa não existe programação, porém é de suma importância para o decorrer do projeto, visto que os dados coletados serão a pedra angular para construção de um sistema confiável.

Após a escolha dos sites, o passo seguinte será o desenvolvimento das ferramentas que terão a finalidade de coletar os dados de forma automática sendo elas o *Web Crawler* e do *Web Scraping*. A linguagem de programação *Python* é amplamente utilizada no desenvolvimento dessas ferramentas, sendo assim foi escolhida para o desenvolvimento no projeto.

O passo seguinte será o processamento dados coletados no serviço do *Google Cloud Dataproc*. Nesta etapa, os dados serão coletados e dimensionados utilizando os processos do *MapReduce* e *Naïve Bayes*.

Com os dados separados, processados e reduzidos, esses dados poderão ser armazenados no *Big Query*. O passo seguinte é a realização da limpeza, transformação e análise dos dados coletados, utilizando.

No último passo, será utilizado o framework *Bootstrap 4.0* para desenvolver um site de busca e apresentação dos resultados obtidos.

2.5– Cronograma

Etapas da pesquisa	**Períodos**			
	Agosto	Setembro	Outubro	Novembro
Seleção dos sites	X			
Elaboração do texto da Monografia	X	X	X	X
Desenvolvimento das ferramentas	X	X		
Análise dos dados coletados		X	X	X
Desenvolvimento do site		X	X	X

REFERÊNCIAS

ANGLIN, K. Gather-Narrow-Extract: A Framework for Studying Local Policy Variation Using Web-Scraping and Natural Language Processing. **Journal of Research on Educational Effectiveness**, Virigina. Disponível em: < https://curry.virginia.edu/sites/default/files/uploads/epw/68_Framework_Local_Policy_Variation_updated_1.pdf> Acesso em: 27 Mai. 2020.

ARAUJO B. Análise de Big Data: Uma comparação do desempenho acadêmico dos discentes do curso de Análise e Desenvolvimento de Sistemas das FATECs Indaiatuba e Sorocaba. 2019. 27 - **Faculdade de tecnologia de Indaiatuba**. Indaiatuba, 2019.

CHAULAGAIN, R. et al. Cloud Based Web Scraping for Big Data Applications. **Conference Paper.** Disponível em: < https://ieeexplore.ieee.org/abstract/document/8118431/> Acesso em: 10 Mai. 2020.

DANIEL, B. Big Data and analytics in higher education: opportunities and challenges. **British journal of educational technology**, v. 46, n. 5, p. 904-920, 2015.

DIMITROV, D. V. Medical Internet of Things and Big Data in Healthcare. **Healthcare Informactis Research**. Varna, 22 Jul 2016. Disponível em: <http://dx.doi.org/10.4258/hir.2016.22.3.15> Acesso em: 10 Mai. 2020.

GABARDO, C. A. LOPES, H. S. **Web crawling e web scrapping: Capturando dados da Internet**. UTFPR. Disponível em: <http://silverio.net.br/heitor/disciplinas/md/aulas/class2b-Webcrawling-Webscrapping.pdf> Acesso em: 22 Mai. 2020

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

HADDAWAY, N. R. The Use of Web-scraping Software in Searching for Grey Literature. **The Grey Journal**. Disponível em: < https://d1wqtxts1xzle7.cloudfront.net/39055185/Haddaway_2015_The_Grey_Journal.pdf> Acesso em: 19 Mai. 2020.

JEN, C. An automated web scraping tool for Malaysia tourism. 2019. 65 - **Faculty of Information and Communication Technology**. Perak, 2019

MACHADO, F. N. R. **Big Data O futuro dos dados e aplicações** 1 ed. São Paulo: Saraiva, 2018

MITCHEL, R. **Web Scraping com Python** 1 ed. São Paulo: Novatec, 2015

NAJADA, H. MAHGOUB, I. Big Vehicular Traffic Data Mining: Towards Accident and Congestion Prevention. **Computer & Electrical Engineering & Computer Science Department**. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7577067/> Acesso em: 10 Mai. 2020.

POLIDORO, F. GIANNIN, R. MOSCA, S. et al. Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. **Statistical Journal of the IAOS**, Roma. Disponível em: <https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji901> Acesso em: 10 Mai. 2020.

SIN, K. MUTHU, L. Application of Big Data in education data mining and learning analytics: a literature review. **ICTACT journal on soft computing**, v. 5, n. 4, 2015.

SLAMET, C. ANDRIAN, R. et al. Web Scraping and Naïve Bayes Classification for Job Search Engine. In: **The 2nd Annual Applied Science and Engineering Conference**, Bandung, Indonesia, 2017. p. 501 – 521.

WADDEL, P. BOEING, G. New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. **Sage journals**, California, 23 Ago 2016. Disponível em: < https://doi.org/10.1177%2F0739456X16664789> Acesso em: 04 Mai. 2020.