EXTRAÇÃO DE DADOS SOBRE "SAÚDE MENTAL" PARA ADICIONAR À BASE DE DADOS DO PSIACADEMIC ANALYTICS

FELIPE SILVA E SILVA[1]; MARIANNE LACERDA DUTRA THEODORO[2]; PABLO HENRIQUE DE SOUSA[3]; LEONARDO SETUBAL MAGGIO[4]; RAYANNA QUIRINO DE SOUSA[5]; PARCILENE FERNANDES DE BRITO[6]

[1]Acadêmico do curso de Ciência da Computação no Centro Universitário Luterano de Palmas – CEULP/ULBRA. Voluntário no PROICT do CEULP/ULBRA. E-mail: felipesilva55@rede.ulbra.br.
[2]Acadêmica do curso de Ciência da Computação no Centro Universitário Luterano de Palmas – CEULP/ULBRA. Voluntário no PROICT do CEULP/ULBRA. E-mail: marianniceci@rede.ulbra.br.
[3]Acadêmico do curso de Sistemas de Informação no Centro Universitário Luterano de Palmas – CEULP/ULBRA. Voluntário no PROICT do CEULP/ULBRA. E-mail: ph.info.cont@rede.ulbra.br.
[4]Acadêmico do curso de Engenharia de Software no Centro Universitário Luterano de Palmas – CEULP/ULBRA. Voluntário no PROICT do CEULP/ULBRA. E-mail: setubal@rede.ulbra.br.
[5]Acadêmica do curso de Ciência da Computação no Centro Universitário Luterano de Palmas – CEULP/ULBRA. Voluntário no PROICT do CEULP/ULBRA. E-mail: rayanna20@rede.ulbra.br.
[6]Doutora em Psicologia pela PUC-GO, Mestre em Ciência da Computação pela UFSC. Professora e coordenadora dos cursos de Ciências da Computação, Sistemas de Informação e Engenharia de Software do CEULP/ULBRA. E-mail: parcilene@gmail.com.

RESUMO

Nos últimos anos, ocorreu um acréscimo substancial de notícias, documentários e trabalhos científicos sobre a temática "Saúde Mental". No mesmo sentido, há uma exponencialização de dados na web sobre os mais diversificados contextos, formando o que alguns pesquisadores denominaram Big Data. Nesse contexto, um dos objetivos do projeto Psiacademic Analytics é fazer a coleta de dados referentes à "Saúde Mental" em repositórios de teses e dissertações de instituições de ensino superior no Brasil. Neste trabalho, especificamente, será apresentado uma parte desse processo, que é a extração de dados e uma análise preliminar desses dados para torná-los mais adequados ao processo de consulta que será realizado em uma etapa posterior.

PALAVRAS-CHAVE: Saúde Mental, Extração de Dados, Teses e Dissertações

1 INTRODUÇÃO

Uma pesquisa realizada pelo Ministério da Saúde, entre 23 de abril e 15 de maio de 2020, constatou que houve um severo aumento da percepção da população brasileira acerca dos impactos da pandemia do Covid-19 e do isolamento social no que se refere à saúde mental. A pesquisa, feita através de um questionário online com 17.491 indivíduos, revelou que na grande maioria dos participantes, cerca de 86,5%, foram verificados casos de ansiedade, seguidos por índices elevados ainda de estresse pós-traumático e depressão, sendo esses 45,5% e 16% dos avaliados respectivamente (BRASIL. Ministério da Saúde. com informações do Nucom SAPS). E apesar da gênese do problema não ter ocorrido propriamente durante o período da pandemia, estudos comprovam que, na média global, cerca de 45% de aproximadamente 21.000 entrevistados afirmaram que o período causou uma piora em seus estados mentais (BBC, 2021). A necessidade de se elaborarem mais estudos que busquem entender a causa e as consequências desse quadro geral de agravamento da saúde mental da população, bem como maneiras de remediar seus impactos é cada vez mais importante no cenário mundial, e principalmente no

cenário brasileiro, uma vez que esse foi um dos destaques no ranking de países que declararam piora em seus estados de bem-estar mental, em uma lista com 30 outros países e territórios pesquisados (BBC, 2021).

A tomada de decisão na área da saúde é um processo que necessita de um embasamento extremamente solidificado e, para isso, a informatização dos processos tem sido grande aliada na gestão da enorme quantidade de volume de dados, os chamados *big data*, que compõem os estudos e registros de saúde. A evolução da *Web* conta com um fenômeno de crescimento acelerado de dados, frutos de uma globalização também acelerada, que incorpora diariamente novas ideias, informações e inovações, tornando a saúde um tema cada vez mais amplo e por consequência, seu fluxo de dados cada vez maior (CÉSAR, 2019). O desafio do *big data* se dá na gestão desses dados, de forma que eles precisam ser analisados e organizados corretamente para que seus conhecimentos sejam disseminados para a população, a fim de atingir o desenvolvimento local. A análise dos dados referentes a serviços de saúde tem o potencial de reduzir custos de tratamentos, evitar epidemias, melhorar o atendimento prestado, bem como a qualidade de vida dos profissionais e pacientes, além de evidenciar processos evitáveis em tratamentos de saúde, por isso, é de suma importância sua aplicação. (MAGALHÃES *et al.*, 2019)

O projeto "Desenvolvimento de um Software para análise inteligente de dados - PsiAcademic Analytics", apoiado pelo Programa de Iniciação Científica e Tecnológica - PROICT - do CEULP/ULBRA, é uma das pesquisas realizadas pelo grupo "Engenharia Inteligente de Dados", a partir da intersecção das áreas de computação e psicologia, e tem como objetivo mapear publicações científicas em repositórios online de instituições de ensino superior de todas as regiões do Brasil, que contenham trabalhos de conclusão de curso, teses e dissertações cujo as temáticas estejam alinhadas com "Saúde Mental". Uma vez mapeados então, esses dados são extraídos de seus repositórios, analisados e sumarizados. Esse é um passo inicial para o entendimento do Big Data sobre "Saúde Mental" e suas mais diversas vertentes, nesse caso, a vertente de produção científica acadêmica.

Nesse sentido, este trabalho apresenta as etapas do desenvolvimento de um *script* de programação em Python, apoiado pela ferramenta Scrapy, para extração de dados de repositórios de universidades brasileiras que seja capaz de fazer a análise dos dados obtidos e seu devido armazenamento no banco de dados MongoDB.

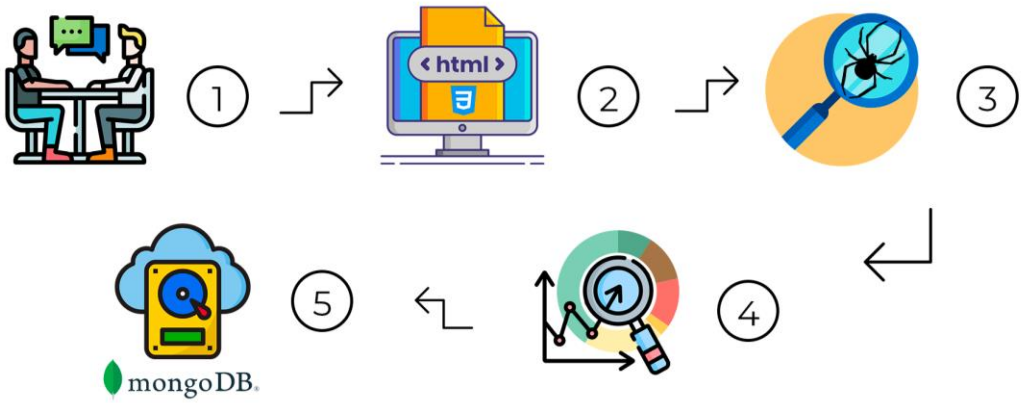
2 MATERIAL E MÉTODOS

O domínio definido para a realização deste trabalho foi repositórios de teses e dissertações de instituições de ensino superior, a saber: UTFPR, UFV, UNICAMP, UNESP, UFPE, USP, UFRJ, UFMG, UFPB, UFRN, UNB, UFRGS, UFSC, UFAM, UFPA e UFG.

Os materiais utilizados no desenvolvimento de Spiders para mineração de dados foram o *framework Scapy* e o banco de dados usado para o armazenamento das informações foi o MongoDB. **Scrapy:** é uma ferramenta de rastreamento e extração de informações da web de alto nível (SCRAPY, 2021). O Scrapy foi utilizado com finalidade de mineração de dados por disponibilizar um conjunto de ferramentas que auxilia a definir maneiras de rastreamento em sites e coleta das informações utilizando Crawlers; **MongoDB:** é um programa de banco de dados não relacional NoSQL de multiplataforma de código aberto. Os bancos de dados NoSQL, como o MongoDB, são úteis para trabalhar com um grande conjunto de dados que não se encaixam bem em um modelo relacional rígido. Por esse motivo, o MongoDB foi utilizado na centralização de todos os dados gerados e manipulados. O MongoDB Atlas é um serviço de banco de dados em nuvem global, em que diferentes desenvolvedores podem ter acesso a um armazenamento flexível que utilize a plataforma de dados de aplicações MongoDB (MONGODB, 2021). A utilização desse tipo de banco de dados em conjuntos de dados que não tem ligações explícitas é uma vantagem para lidar com o problema do Big Data, que necessita armazenar um grande volume de dados com custos reduzidos e que tenha facilidade para adicionar novas informações.

Na figura a seguir são apresentadas as etapas para o desenvolvimento deste trabalho.

Figura 1 - Etapas de Realização do Trabalho

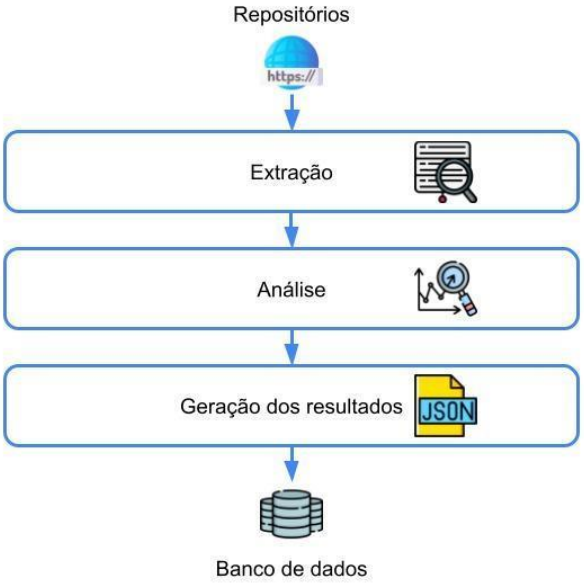


A etapa 1 da Figura 1 ilustra a definição dos critérios de validação dos repositórios escolhidos para compor o processo de extração. Nesta etapa foram definidos os repositórios e instituições de ensino superior válidos: UTFPR, UFV, UNICAMP, UNESP, contabilizando novos quatro repositórios extraídos, totalizando dezesseis quando somados aos doze já extraídos anteriormente, UFPE, USP, UFRJ, UFMG, UFPB, UFRN, UNB, UFRGS, UFSC, UFAM, UFPA e UFG. A etapa 2 demonstra o reconhecimento das estruturas das páginas. Após a verificação das páginas resultantes da pesquisa do termo “Saúde Mental”, foram selecionados os seguintes campos para a extração: título, autores, orientadores, palavras chaves, data de publicação, resumo, repositório e o tipo da publicação. A etapa 3 ilustra o desenvolvimento dos Crawlers. Nesta etapa foi feita a criação de Spiders para cada repositório, utilizando os seletores definidos na análise da estrutura HTML e CSS de cada página, com a utilização do Framework Scrapy. A etapa 4 ilustra a validação das informações extraídas de cada repositório. Nesta etapa o processo de análise dos dados identifica se todos os campos que são requisitos para aprovação estão presentes, verifica e corrige possíveis erros de legibilidade na estrutura da extração final. E na etapa final, 5, os dados coletados e validados provenientes da extração foram armazenados no MongoDB.

3 RESULTADOS E DISCUSSÃO

O processo de entendimento do Big Data sobre a temática de “Saúde Mental” no contexto de trabalhos científicos desenvolvidos nas instituições de ensino perpassa, primeiramente, a extração e organização desses dados. Nesse contexto, para este trabalho, foram construídos quatro Crawlers. Isso foi necessário, pois cada repositório, mesmo tendo estruturas semelhantes, possuem códigos HTML diferentes. A estrutura do HTML presente em cada repositório torna complexo o desenvolvimento de só um programa de extração que se encaixe nos quatros cenários. A figura 2 ilustra os passos seguidos no desenvolvimento do projeto.

Figura 2 - Modelo de desenvolvimento



Repositórios: Inicialmente foi preciso identificar os repositórios válidos, por meio da análise da estrutura de algumas páginas, onde foi possível averiguar se os campos definidos anteriormente pelo especialista de domínio, como tópicos importantes para extração, estavam de fato disponíveis. E, se disponíveis, o próximo passo segue sendo identificar o posicionamento de tais informações, seja encontrando seu endereço XPath, que seria a expressão do caminho em HTML, ou a determinação da classe especificada em sua estilização, em CSS. **Extração:** A extração de dados das páginas dos repositórios ocorreu através de Crawlers, ou Spiders, que por meio de Seletores (mecanismos específicos de busca oferecidos pela própria estrutura do Scrapy) reconhecem o posicionamento dos dados. Assim, de página em página, dentro do repositório da universidade pretendida, são identificados os dados e extraídos para dentro da classe *Items* do Scrapy, que tem como finalidade estruturar esses resultados obtidos. A imagem a seguir exemplifica como esses seletores são criados dentro da Spider.

Figura 3 - Exemplo de Seletores XPath e CSS implementados na criação do Crawler.

```
campos = response.xpath('//*[@id="content"]/div[2]/table//tr')
titulo = response.css('.metadataFielValue.dc_title::text').get()
```

Análise: Essa etapa teve como objetivo analisar os campos obtidos e definir se o conjunto de dados é válido ou descartável. Para isso, foram implementadas funções que verificaram a quantidade de dados obtida, para filtragem inicial de publicações que tivessem todos os dados completos, e também para eliminação de caracteres indesejados no corpo do arquivo da extração, para fins de legibilidade. Essa etapa foi fundamental para preparar os dados para as etapas finais, que tratam de disponibilizá-los para dentro do banco de dados. **Geração dos resultados:** utiliza a classe *Pipeline* que recebe os dados aprovados do passo anterior e gera um arquivo no formato JSON; ao passarem pela classe os dados são armazenados no **Banco de Dados** MongoDB. E o processo se repete até todas as páginas disponíveis serem manipuladas.

4 CONCLUSÃO

O processo de criação do *script* descrito neste trabalho teve como objetivo a extração de dados de 4 novos repositórios de instituições de ensino superior, que obteve como resultado a coleta das informações selecionadas, suas análises e, no final, seu armazenamento. O desenvolvimento do projeto revelou que as práticas, inicialmente planejadas, não seriam suficientes para abranger todos os cenários, devido às diferenças entre os repositórios. O que tornou o processo mais demorado e as soluções de cada *script* mais personalizadas.

O desempenho do processo de mineração de dados pôde ser melhorado utilizando novas técnicas ou aperfeiçoando os métodos empregados na busca de informações dentro das páginas web e no acesso às publicações disponíveis.

Para trabalhos futuros, tem-se que o desenvolvimento de uma ferramenta de extração e análise de dados, que seja capaz de unificar diferentes fontes de informação dentro de um mesmo banco de dados, apoiado pela implementação de outra ferramenta que facilite a visualização de todo esse material obtido, criará uma enorme oportunidade de distribuição de informações válidas e comprovadamente verídicas para a comunidade da área de saúde mental, sendo que essa carece cada vez mais de gestão em seu *big data*.

5 AGRADECIMENTOS

Este projeto é apoiado pelo Programa de Iniciação Científica e Tecnológica - PROICT - do CEULP/ULBRA.

6 REFERÊNCIAS

BRASIL. MARINA PAGNO. . **Ministério da Saúde divulga resultados preliminares de pesquisa sobre saúde mental na pandemia**. 2020. Ministério da Saúde, com informações do Nucom SAPS. Disponível em: https://antigo.saude.gov.br/noticias/agencia-saude/47527-ministerio-da-saude-divulga-resultados-preliminares-de-pesquisa-sobre-saude-mental-na-pandemia. Acesso em: 30 set. 2021.

CÉSAR, Hilton Vicente. **Mineração de Processos para Extração de Indicadores de Sistema de Informação para Construção de Matriz de Saúde Mental**. 2019. 119 f. Tese (Doutorado) - Curso de Bioengenharia, Universidade de São Paulo Escola de Engenharia de São Carlos Faculdade de Medicina de Ribeirão Preto Instituto de Química de São Carlos, São Carlos, 2019. Disponível em: https://www.teses.usp.br/teses/disponiveis/82/82131/tde-27022020-163442/publico/Tese_HiltonVicenteCesar_DO_Corrigida.pdf. Acesso em: 30 set. 2021.

COVID: saúde mental piorou para 53% dos brasileiros sob pandemia, aponta pesquisa. S,L, 14 abr. 2021. Disponível em: https://www.bbc.com/portuguese/geral-56726583. Acesso em: 07 out. 2021.

MAGALHÃES, Jorge *et al*. Gestão do conhecimento em tempos de big data: um olhar dos desafios para os sistemas de saúde. **Sannitatem Quaerens In Tropicos**, Sl, v. 17, n. 2, p. 7-16, fev. 2019. Anual. Disponível em: https://anaisihmt.com/index.php/ihmt/article/view/256/212. Acesso em: 30 set. 2021.

MONGODB. **Atlas**. Disponível em: https://www.mongodb.com/pt-br/cloud/atlas. Acesso em: 30 set. 2021

SCRAPY. Scrapy 2.5.0 documentation. Disponível em: https://docs.scrapy.org/en/master/index.html. Acesso em: 30 set. 2021.