Aluno do PPEE - Programa de Mestrado Profissional em Engenharia Elétrica, Brasília, DF

Relatório Final do Produto ”E” do Edital DPI/DPO - UnB n. 01/2020

Lucas Coelho de Almeida - Matrícula 20/0004506

***Resumo*— Documento que explica a metodologia e a forma de uso das ferramentas criadas para visualização de dados referentes à infraestrutura de pesquisa da UnB e do impacto de produção associada através da extração de dados do Sistema de Currículos Lattes e dos dados disponibilizados pelo departamento.**

***Palavras-Chave*— Infraestrutura de Pesquisa; Extração de Dados; Currículo Lattes; Business Intelligence, *Webscrapping*.**

I. Introdução

A Universidade de Brasília não dispunha, até a data deste documento, de base de dados estruturados relacionados com sua infraestrutura de pesquisa. O objetivo deste documento será explicar detalhes e formas de uso das ferramentas que foram desenvolvidas para vencer as dificuldades de processamento de dados não estruturados e também de geração de informações de valor através de dados disponíveis publicamente, de forma a atender o disposto na descrição do produto ”E” do edital DPI/DPO - UnB n. 01/2020.

II. Entregáveis

Foi entregue uma pasta digital de arquivos compactada em formato ”.zip” denominada ”Entregavel-Edital-DPI-Lucas-Coelho”. Para visualizar os dados e projetos (e/ou editá-los), é necessário descompactá-la localmente. Dentro dela, existem 5 pastas:

1) **”0-Relatorios”:** Pasta na qual se encontra este relatório e os demais com as visualizações geradas e exportadas da ferramenta Tableau.

2) **”1-Extrator-Dados”:** Pasta na qual se encontram os programas que realizam tratamento e extração de dados.

3) **”2-Projeto-Tableau”:** Pasta na qual se encontram os projetos executados utilizando o software Tableau.

4) **”3-Dados-Obtidos”:** Pasta na qual foram disponibilizados os dados relevantes obtidos com o projeto e os quais são a base dos projetos executados no software Tableau. Também dentro dessa se verifica uma pasta com todos os currículos em formato ”.html” extraídos.

5) **”Pre-Requisito-Driver-Extrator-Dados”:** Pasta na qual se verifica o principal e menos trivial pré-requisito para utilização do programa de extração de dados da web. Importante notar que para que os programas sejam usados, é necessário, além de observar essa pasta, instalar a linguagem Python versão 3.6 ou superior e todas as bibliotecas que estiverem no cabeçalho de cada *script*.

Departamento de Pesquisa e Inovação (DPI), Universidade de Brasília (UnB), E-mails: luccoelhoo@gmail.com

III. Metodologia e Uso das Ferramentas

Em primeiro lugar, é importante notar que o departamento dispunha de um documento com tabelas as quais continham um compilado (feito manualmente) de informações sobre os centros de pesquisa da Universidade. Ainda que não estivessem preparados para processamento e contivessem, de forma geral, apenas dados de identificação, foi possível aplicar técnicas arrojadas de comparação de texto e filtragem para obter visualizações interessantes a respeito de toda a infraestrutura de pesquisa da Universidade (visualizações essas obtidas através do uso do software Tableau com uma licença para estudante).

Além dessa tarefa, era necessário trazer uma medida válida para comparar as produções de cada laboratório. Primeiramente, foi sugerido obter artigos de bases internacionais e garimpar aqueles que citassem laboratórios da UnB, entretanto, essa abordagem privilegiaria os resultados das áreas de Ciências Exatas em detrimento das de Ciências Humanas, que nem sempre realizam essas citações e cuja produção mais significante é verificada em outras fontes não tão simples de serem acessadas, como livros e revistas. Assim, foi decidido, junto ao coordenador do projeto, que a produção dos coordenadores dos laboratórios seria uma métrica válida para o contexto atual em que não se tem nenhum critério de comparação. Além disso, verificou-se que a única fonte confiável e padronizada de dados referentes a produção seria verificada através do Sistema de Currículos Lattes, da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). Portanto, seria necessário encontrar uma forma de extrair dados deste portal para processá-los e associá-los às infraestruturas da Universidade.

A seguir, são explicados os dois objetivos escolhidos para cumprir os entregáveis do projeto.

1) **Visualização de dados:** As visualizações foram geradas usando o software Tableau e uma licença de estudante. Os dados são consumidos dos arquivos disponíveis na pasta ”3-Dados-Obtidos”. Ao abrir um projeto da pasta, é necessário conectar com a base de dados correspondente. No caso, tanto o nome dos projetos quanto o nome dos arquivos ”.xls”estão relacionados, portanto, para o projeto que contém a palavra ”Identificacao” no nome dentro da pasta

”2-Projeto-Tableau” , deve-se conectar à base/tabela denominada ”Identificacao.xls” na pasta ”3-Dados-Obtidos”, para aquele que contém ”ListaGeral”, deve-se conectar à base/tabela ”ListaGeral.xls”, e para aquele que contém ”Producao” no nome, deve-se conectar à base/tabela ”Producao.xls”. Depois de abrir os projetos e conectar-se à respectiva base, será possível editar as visualizações. Os dados das tabelas ”Identificacao.xls” e ”ListaGeral.xls” correspondem, com pequenas mudanças, aos dados nas tabelas de mesmo nome do arquivo ”1.1 Infraestrutura_Pesquisa_UnB_DPI_Dirpe.xls” que fora fornecido pelo departamento. A tabela ”Producao.xls” é o resultado das extrações, tratamentos e conferências, associando dados das duas tabelas citadas.

2) **Produção Acadêmica Associada:** Conforme dito anteriormente, a produção de cada laboratório foi medida de acordo com o currículo Lattes de cada coordenador registrado conforme o documento ”1.1 Infraestrutura_Pesquisa_UnB_DPI_Dirpe.xls”, o qual contém os dados de identificação disponibilizados pelo departamento. Entretanto, uma quantidade relevante dessas infraestruturas continha mais de um coordenador na mesma linha, e o número total de centros de pesquisa era grande o suficiente para justificar a prospecção de formas automatizadas de tratamento dos dados. Portanto, para essa tarefa, foram criados três (3) scripts de automação escritos na linguagem Python, os quais estão disponíveis na pasta ”1-Extrator-Dados”. Para usá-los, é necessário instalar a linguagem Python versão 3.6 ou superior, bem como as bibliotecas listadas no cabeçalho de cada arquivo. Esses *scripts* devem ser executados numa ordem predefinida e seu uso contém detalhes para cada caso conforme descrito a seguir.

   a) **”0-GeraLista-Profs.py”:** Esse é o primeiro arquivo a ser executado e irá extrair os nomes dos coordenadores listados no documento ”1.1 Infraestrutura_Pesquisa_UnB_DPI_Dirpe.xls” dentro da pasta ”0-dados-laboratorios” e separar aqueles que tiverem mais de um coordenador, gerando, como saída, um arquivo em tabela com a listagem que associa cada coordenador a cada laboratório (lembrando que os laboratórios irão se repetir). Esse arquivo é denominado ”Lista-Professores-Laboratorios.csv” e estará dentro da pasta ”2-nomes-extraidos-coordenadores”.

   b) **”1-Compara-Lista-Com-Nomes-Profs.py”:** Esse é o segundo arquivo a ser executado. Irá comparar os nomes extraídos no arquivo ”Lista-Professores-Laboratorios.csv” dentro da pasta ”2-nomes-extraidos-coordenadores” com os nomes oficiais dos docentes da Universidade de Brasília disponíveis no arquivo ”ListaNomeProfessoresOficial.xlsx” na pasta ”1-nomes-oficiais-professores-unb”. Nessa comparação, o programa irá adequar os nomes, retirando termos e caracteres que atrapalham a pesquisa na base de currículos Lattes, trocando letras maiúsculas por minúsculas e procurando o nome oficial que mais provavelmente corresponde ao nome do coordenador extraído pelo primeiro programa. Para essa comparação, é gerada uma lista preliminar, e se o nome do coordenador não é encontrado nessa lista, utiliza-se um algoritmo baseado no ”Método de Levenshtein” de comparação de textos. O nome só é aceito se a semelhança é igual ou superior a 85%. Se for menor, o algoritmo escolhe da mesma forma, porém é colocado um termo ”SIM” na coluna ”PROVAVELMENTE ERRADO” do arquivo de saída. O arquivo de saída conterá, portanto, os nomes dos laboratórios, os nomes ”normalizados” para pesquisa após comparação e a indicação da probabilidade de erro quando aplicável. É importante destacar que mesmo após esse processamento, é necessária uma checagem manual para verificação da integridade do documento e das decisões tomadas pelo programa. Após várias tentativas, verificou-se uma taxa de sucesso de pouco mais de 80%. O restante dos nomes não parecem estar na lista oficial de nomes dos docentes ou estão demasiadamente incompletos e mesmo manualmente não foram passíveis de identificação (uma possível causa é o fato de que nem todos os nomes de coordenadores de laboratórios são docentes da UnB). O arquivo de saída é denominado ”Lista-Comparada-Professores-Laboratorios.csv” e estará dentro da pasta ”3-nomes-coordenadores-apos-comparacao”.

   c) **”2-Download-CV-HTML-Usando-Lista.py”:** Esse é o terceiro e último programa a ser executado. Para usá-lo, é necessário instalar, além das bibliotecas no seu cabeçalho, o programa ”GeckoDriver”, o qual irá permitir o controle remoto do navegador através da ferramenta ”Selenium”. Mais informações sobre essa necessidade estão disponíveis na pasta ”Pre-Requisito-Driver-Extrator-Dados” entre os entregáveis do projeto. Sua execução é lenta devido o fato de lidar com extração de dados de páginas web, portanto, para extrair todos os currículos, esse pode levar diversas horas e indica-se o uso de um servidor específico para tal ou de uma máquina virtual em ambiente de regime contínuo de atividade. Entretanto, antes de proceder, é importante salientar, conforme comentado no item anterior, que a saída do *script* ”1-Compara-Lista-Com-Nomes-Profs.py”, o qual faz a comparação e tenta ”limpar” os nomes para obterem mais sucesso na pesquisa, deve ser conferida manualmente. Após essa conferência, ela deve estar na mesma formatação

e com o mesmo nome do arquivo ”Professores-Laboratorios-Manual.xlsx” disponível dentro da pasta ”4-nomes-coordenadores-apos-conferencia-manual”. Esse arquivo foi deixado propositalmente na pasta, e também está disponível na pasta ”3-Dados-Obtidos” entre os entregáveis do projeto. Esse é, provavelmente, o arquivo mais confiável e estruturado com os nomes dos coordenadores e das infraestruturas de pesquisa da UnB. Assim, quando executado, o programa irá buscar pelos nomes no arquivo ”Professores-Laboratorios-Manual.xlsx” dentro da pasta ”4-nomes-coordenadores-apos-conferencia-manual”, controlar o navegador ”Firefox” em modo de ”controle remoto” e interagir com páginas web conforme um ser humano faria. Assim, levará cerca de 20 segundos para extrair cada currículo. É importante salientar que o método usado é o único disponível pois o portal utiliza tecnologia de *Captcha* da empresa Google para impedir acessos de programa aos dados. Existe uma forma institucional de acesso, através de uma interface de programa disponibilizada pela CAPES, porém esta exige cadastro da instituição e não foi possível obter o acesso da Universidade até o momento em que este relatório foi escrito. Ao fim da execução, o programa terá salvo os currículos em formato ”.html” na pasta ”5-curriculos-extraidos-html” e também um arquivo em formato ”.csv” de *log* do processo para cada coordenador na pasta ”5-log-extracao-curriculos”. Nesse arquivo, será possível verificar, para cada nome, se o processo de extração foi um sucesso ou não e se foi usado o nome oficial da lista de docentes da Universidade ou se foi usado o nome secundário, que seria o original na listagem fornecida pelo departamento (se o primeiro não obtém sucesso, o programa tenta com o segundo). Também, na última coluna do arquivo de *log*, é possível verificar a quantidade de resultados obtidos após a busca do nome do docente. Isso é importante pois o programa baixa sempre o primeiro resultado. Felizmente, foi possível verificar poucos casos de mais de um resultado retornado pela pequisa.

A seguir, as conclusões sobre o projeto e desenvolvimento executados.

IV. Conclusão

No início, o projeto era um grande desafio, principalmente pela novidade e falta de dados estruturados. No entanto, com as reuniões e decisões de projeto bem acertadas, bem como a base teórica fornecida pelo programa de Pós-Graduação da UnB, foi possível diminuir a complexidade desse e visualizar informações bastante relevantes sobre a infraestrutura de pesquisa da UnB, como a distribuição dos laboratórios por área, por faculdade, a área útil de pesquisa da Universidade, a distribuição dessa área física e a produção associada dos laboratórios (representados por seus coordenadores). Essas informações, em conjunto, são essenciais para o processo de tomada de decisão e podem contribuir bastante para o futuro da pesquisa na instituição. As melhorias só seriam possíveis caso os dados fossem estruturados numa base complexa e atualizados periodicamente, trabalho que exige dispêndio de tempo e técnica bastante relevante. De toda forma, os resultados foram satisfatórios e permitem uma visão de alto nível das distribuições de espaços e contribuições/produções das infraestruturas de pesquisa da Universidade.