Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Suporte computacional para gestão de programas de Pós-Graduação da Universidade de Brasília

Amanda Aline Figueiredo Carvalho

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia de Computação

Orientador
Prof. Dr. Rafael Timóteo de Sousa Júnior

Brasília
2019

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Suporte computacional para gestão de programas de Pós-Graduação da Universidade de Brasília

Amanda Aline Figueiredo Carvalho

Monografia apresentada como requisito parcial
para conclusão do Curso de Engenharia de Computação

Prof. Dr. Rafael Timóteo de Sousa Júnior (Orientador)
ENE/UnB

Prof. Dr. Edna Dias Canedo
CIC/UnB

Dr. Fábio Lúcio Lopes de Mendonça
ENE/UnB

Prof. Dr. José Edil Guimarães de Medeiros
Coordenador do Curso de Engenharia de Computação

Brasília, 13 de dezembro de 2019

Dedicatória

Dedico este trabalho a todos, que me incentivaram e incentivam até hoje, a superar os desafios impostos pela vida e a seguir em frente mesmo com todas as adversidades e contratempos percebidos durante essa trajetória.

Agradecimentos

Em primeiro lugar, agradeço a Deus por ter me dado a vida e principalmente permitido que eu tivesse uma melhora na minha saúde para continuar o meu trabalho, mesmo diante de todas as dificuldades que enfrentei e continuo enfrentando. Agradeço também a Ele, por ter colocado em meu caminho pessoas incríveis auxiliando nesse momento árduo e crítico, que é a escrita da monografia.

Agradeço aos meus pais, Lígia e William por todo o investimento em minha educação, desde a infância até o fim da minha graduação, por toda confiança e crédito ao meu trabalho, além de incentivo e motivação. Agradeço também aos meus avós, que cuidaram de mim durante a infância e zelaram para que eu pudesse completar meus estudos com tranquilidade e desenvolver um bom caráter para vencer as dificuldades da vida.

Agradeço também ao meu grande amigo e companheiro de vida, Gutierres, por todo carinho que tem por mim e paciência para lidar com a falta de tempo e correria que tem sido este último ano. Obrigada por toda a ajuda com aqueles problemas intermináveis durante o tratamento dos dados deste trabalho, por todo incentivo e confiança ofertada a mim, mesmo quando tudo parecia não dar certo. Agradeço ainda mais por me acompanhar ao hospital tantas vezes e me ensinar o quanto é importante viver todos os dias, sem desistir de nenhum deles, e enfrentar todos os problemas sempre com um sorriso no rosto e vontade de vencer.

Ao meu irmão, Alexandre, agradeço por toda ajuda durante a graduação e auxílio neste trabalho de conclusão de curso.

Agradeço imensamente ao Bruno, por me ajudar tanto em toda minha jornada acadêmica, por estudar junto comigo nas matérias que eu tive dificuldade, pela amizade, por todas as caronas para que eu pudesse chegar cedo e segura à faculdade, e por todo seu incentivo, puxões de orelha e preocupação para que eu alcançasse com êxito todos os meus objetivos.

Agradeço ao meu orientador da primeira fase do Trabalho de Conclusão de Curso, João Paulo Lustosa, por ter ensinado que sempre devemos dar o melhor de nós e buscar sempre a excelência em tudo que fizermos.

Ao meu orientador atual, Rafael Timóteo, por todo o auxílio e sugestões para melhoria do meu trabalho, além do apoio dado durante o tempo que trabalhei em projetos junto

ao Latitude.

Às minhas amigas Cecilia, Heloísa e Daniela, que me ajudam desde o Ensino Médio a percorrer o caminho crítico da vida, que compreendem minhas longas ausências e momentos de preparação para as provas da faculdade e concursos públicos.

Aos meus chefes, Norma, Luciano, Jamilla e Rogério, que me acolheram tão bem quando entrei na Secretaria de Saúde, um ambiente tão novo para mim, que sempre compreenderam meus horários complicados por causa da faculdade e permitiram que eu flexibilizasse minha escala tantas vezes, e também pelas várias palavras de apoio, naqueles dias difíceis, em que eu chegava exausta no trabalho no auge dos fins de semestre.

Aos meus amigos e colegas de faculdade, que me acompanharam durante estes 5 longos anos, em meio aos obstáculos e alegrias, agradeço pelas inúmeras sugestões para melhorias em meu trabalho de conclusão de curso e todo o auxílio e parceria durante os trabalhos em grupo, que por vezes eram tão exaustivos que chegávamos a achar que não daríamos conta.

Agradeço também ao Fábio Lúcio por ter me dado oportunidade de trabalhar no laboratório e por ter dado tantas dicas para melhoria do meu trabalho de conclusão de curso.

A todos os professores que se dedicaram em minha formação, aos meus colegas de trabalho que sempre me fazem rir e amenizam as dores do dia a dia e a todos aqueles que não mencionei, mas que fizeram parte dessa trajetória árdua, porém gratificante, recebam o meu "Muito Obrigada" e saibam que parte desse momento também é de vocês.

Este projeto contou com apoio da Agência Brasileira de Pesquisa, Desenvolvimento e Inovação CNPq (Projeto INCT em Segurança Cibernética 465741/2014-2), bem como do Ministério da Economia (TEDs 011/2016 SEST, 005/2016 DIPLA e 083/2016 ENAP) e do Laboratório LATITUDE/UnB (Projeto SDN 23106.099441/2016-43), o que foi de extrema importância para o desenvolvimento do trabalho.

Resumo

A Mineração de Dados é uma área que estuda o comportamento dos dados e visa extrair padrões para construção de uma informação consistente. Uma vertente é a mineração de textos, que busca compreender o comportamento daquilo que está disposto em palavras, associado aos demais elementos a serem analisados. Este trabalho tem como objetivo propor uma plataforma de suporte computacional para gerenciar os programas de pós-graduação e auxiliar a tomada de decisão para solucionar problemas que demandam muita ação humana, de maneira eficiente. A plataforma utiliza ferramentas de Business Intelligence, técnicas de mineração de dados e mineração de textos, bem como aplica os conhecimentos de cientometria e bibliometria, para apresentação dos dados. Os resultados foram validados por meio de comparação junto aos dados originais, extraídos da Plataforma Lattes (Currículo Lattes) e Google Scholar.

Palavras-chave: Lattes, Análise, Inteligência de Negócios, Mineração de Dados

Abstract

Data Mining is an area that studies data behavior and aims to extract patterns for building consistent information. One aspect is text mining, which seeks to understand the behavior of what is arranged in words, associated with the other elements to be analyzed. This paper aims to propose a computer support platform to manage postgraduate programs and assist decision making to solve problems that demand a lot of human action, efficiently. The platform uses Business Intelligence tools, Data Mining and Text Mining techniques, as well as applying the knowledge of scientometry and bibliometrics for data presentation. The results were validated by comparison with the original data, extracted from Lattes Platform (Lattes Curriculum) and Google Scholar.

Keywords: Lattes, Analysis, Business Intelligence, Data Mining

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Ciência é o método utilizado para estudar os fenômenos que acontecem no universo, buscando compreender e até mesmo prever comportamentos. Para que esse objetivo seja alcançado, o método científico deve ser aplicado e pode receber auxílio da tecnologia para melhoria do processamento da informação e criação de novas análises [6]. No vasto campo científico, existe uma vertente que se apoia em meios computacionais para atingir o objetivo de processar um conjunto de dados e torná-los uma informação clara e consistente. A Ciência de Dados pode ser vista como uma ciência multidisciplinar que trata da compreensão do que os dados apresentam, sejam eles elaborados de forma estruturada ou não [7].

1.1 Motivação

Os Programas de Pós-Graduação no Brasil são divididos em segmentos conforme a Figura 1.1. O foco deste trabalho é a Pós-Graduação Stricto Sensu, cujo objetivo principal é formar uma categoria de profissionais qualificados para o atendimento da educação básica e superior, bem como incentivo à pesquisa científica e desenvolvimento nacional a partir do treinamento de mão-de-obra técnica e de alta qualidade. O objetivo é mapear, aplicando técnicas de mineração de dados, o perfil destes programas, o comportamento dos pesquisadores que os compõem e suas relações com a qualidade avaliada e perfil de produção científica.

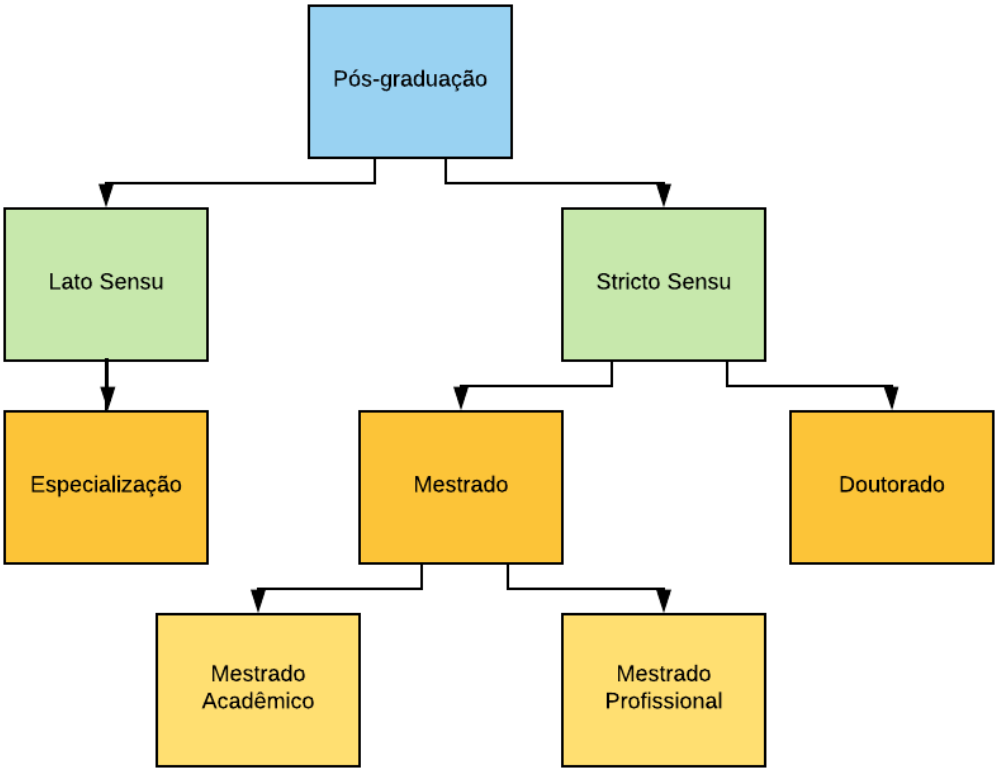

Figura 1.1: Estrutura dos Programas de Pós-Graduação no Brasil

A produção científica brasileira está em constante processo de expansão, de modo acelerado e num ambiente altamente colaborativo, o que motiva a análise das redes de cooperação entre os pesquisadores [6]. Outro ponto importante é entender como a pesquisa científica está distribuída dentro da universidade, bem como o nível de centralização das publicações entre os pesquisadores de uma mesma área. Isso serve de incentivo para descrever os processos de transmissão do conhecimento, produção de tecnologia em âmbito nacional e evolução das publicações em nível internacional [8]. Os padrões geográficos da colaboração científica tendem a seguir uma linha econômica, como é o caso dos países emergentes, cujo os fatores culturais demonstram um baixo grau de internacionalização da colaboração científica. Para validar estas afirmações, é possível utilizar da mineração de dados, para visualização de aspectos espaciais e analíticos sobre o comportamento da pesquisa no Brasil [6].

Cada vez mais se buscam informações de alta relevância nos meios informatizados, que aproximam as pessoas interessadas num mesmo assunto ou até mesmo com intuito de aprender sobre determinada área e aprofundar um campo de pesquisa [9]. A interdisciplinaridade está em presente atualmente dentro do contexto universitário e com este trabalho busca-se verificar o grau de diversidade na Universidade de Brasília, demonstrando

a importância do Currículo Lattes e demais bases do conhecimento para consolidação de respostas sobre o tema [10]. O avanço computacional nesta área científica, até então focada em análises exploratórias e qualitativas, permitiu que uma abordagem teórica pudesse ser validada e controlada por meio de cruzamento de dados dessas bases, e pudesse ser feita a construção do conhecimento estruturado e de fácil acesso e visualização [11].

1.2 Plataforma Lattes

Vivemos numa sociedade cujas as decisões estão pautadas na informação e no conhecimento acerca desta informação. O avanço da tecnologia da informação fez surgir diversos sistemas para levar dados para a realidade das pessoas, de forma cada vez mais próxima. Com isso, plataformas Web foram desenvolvidas para compartilhamento de informações e disseminação do conhecimento [12]. É notável que a ciência aliada à tecnologia agregam valor de cunho técnico, político, social, econômico e cultural, ampliando o leque de soluções para diversos problemas. Um marco importante na década de 50 foi a criação do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), que possuía outro nome na época de sua institucionalização, com a função de formar um elo entre o Estado e o desenvolvimento científico e tecnológico do Brasil [13].

Nos anos 90, surge a Plataforma Lattes juntamente com os Diretórios dos Grupos de Pesquisa, com o papel de monitorar as políticas e mecanismo de manutenção e incentivo à pesquisa. Em 1999, a plataforma traz o sistema Currículo Lattes, uma proposta de sistema e-gov para prestar serviços à população e conectar o campo governamental ao cidadão, transmitindo conhecimento e acesso à informação [14]. O Currículo Lattes é um banco de dados com informações curriculares, que serve de referência para a pesquisa e profissionais de diversas áreas. É de alta relevância no contexto da Ciência e Tecnologia para auxiliar na tomada de decisão, seja para escolha de um profissional para o mercado de trabalho, ingresso em programas universitários ou concorrência em bolsas de pesquisa, avaliando as competências associadas a cada estudante ou profissional [15].

Ao explorar a base, é possível encontrar informações como: artigos publicados, propriedade intelectual, formação acadêmica, autoria em livros, orientações em dissertações e teses, participação em congressos, além de outras informações relevantes para elaboração de um ciclo metodológico para mapeamento do comportamento da pesquisa científica [16]. É de suma importância validar as informações constantes no Currículo Lattes, visto que é um canal de análise para avaliar as produções, os programas de pós-graduação, méritos para usufruir de financiamentos à pesquisa e a extensão do potencial produtivo do país. Dessa forma, a comunidade acadêmica tem se inserido na plataforma de currículos e

contribuído para a geração de dados e consequentemente, produção de informação acerca do cenário intelectual e tecnológico brasileiro [17].

Com a possibilidade de modelagem e caracterização dos perfis de publicações dos pesquisadores e de redes de colaboração, é possível fornecer condições para diversos estudos como análise preditiva de publicações entre pesquisadores, possibilitar a aplicação de técnicas, como gamificação, para ranquear os pesquisadores em suas áreas de pesquisa e estimular a produtividade, bem como gerenciar estratégias para maximizar pontuação dos programas em avaliações educacionais, com a perspectiva de melhoria do ensino [18]. Cabe ressaltar que esta tarefa possui diversos obstáculos, que torna necessário o uso de técnicas para tratamento dos dados, logo após a extração. Essas dificuldades surgem devido à ambiguidade nos nomes de autores ao serem informados, a incompatibilidade entre os autores das publicações que, por vezes, não informam os outros componentes que também tiveram autoria na publicação, significativos erros nos dados, durante o processo de extração, dentre outros.

Este trabalho propõe uma plataforma de suporte computacional para visualização de dados, formada estruturalmente com mecanismos de extração de todo o conjunto de dados dos currículos Lattes, capaz de manter o escopo de dados científicos atualizados, a cada execução de código, e facilitar a composição dos painéis de visualização. Além disso, foram utilizadas técnicas para análises bibliométricas dos dados, com a possibilidade de identificar redes de colaboração científica entre os pesquisadores dos programas, auxiliando no aumento da colaboração e pontuação em avaliações, como da Capes, por exemplo.

A primeira parte do processo é realizar a extração dos dados curriculares da Plataforma Lattes e, após, fazer a modelagem dos bancos de dados que abrigarão o conjunto de dados extraído. A extração pode ser feita de forma automatizada por meio de scrapping, caso a Plataforma Lattes não utilize nenhum mecanismo de segurança, como ReCaptcha, ou por meio de convênio da instituição com o CNPq, que será responsável por manter a base de dados atualizada, podendo os dados serem usados apenas com a finalidade de pesquisa científica. Feita a extração, é necessário fazer o tratamento dos dados para que possam se adequar ao banco de dados modelado e às necessidades de análise. Com os dados consolidados, é possível criar as análises bibliométricas, para serem exibidas de forma a facilitar o consumo das informações pelos usuários, focando em quantificar os dados extraídos e tratados, e torná-los informações precisas. Essas análises são de suma importância e devem apresentar corretude e precisão, pois realçam diversos aspectos de como as pesquisas científicas estão sendo conduzidas nos programas analisados.

Outro ponto importante, é que vários editais de financiamento de projetos feitos por instituições de amparo à pesquisa, como o próprio CNPq, CAPES e FAP, utilizam os currículos Lattes como um método de avaliação das propostas que receberão recursos para

sua execução, bem como auxílio financeiro para viagens a congressos e bolsas de estudo de diversas modalidades. Tal fato motiva os pesquisadores a manterem seus currículos atualizados, tornando a Plataforma Lattes uma fonte rica e confiável para analisar a produção científica no Brasil.

É fato que a universidade deve propiciar um ambiente que vise e incentive a melhoria contínua de seus pesquisadores e a tecnologia da informação (TI) pode ser uma grande aliada nesse processo. A TI permitiu que a informação chegasse a pessoas que, anteriormente, por não terem conhecimentos técnicos suficientes, não poderiam acessar determinados tipos de conteúdo [9]. Hoje em dia, existem métodos para manipular dados e filtrar da maneira que se desejar, sem que haja qualquer conhecimento sobre programação. Isso é possível por meio da utilização de painéis interativos, mais conhecidos como dashboards. Os dados da Platafoma Lattes, que antes demandavam consulta individual ao currículo de cada pesquisador, podem ser vistos utilizando painéis interativos, de fácil manuseio. Para atingir tal fim, é necessário desenvolver previamente recursos para a organização, tratamento, manutenção e disseminação dos dados, para atingir a finalidade de tomada de decisão a que se deseja.

1.3 Google Scholar

O Google Scholar é uma plataforma que age por meio de um mecanismo de pesquisa de dados acadêmicos, baseado em dados presentes na Internet, sendo capaz de catalogar entre 2 e 100 milhões de registros de literatura acadêmica e não-acadêmica. O Google Scholar coleta resultados sobre publicações na Internet de forma gratuita. Diante disso, têm crescido bastante as pesquisas que tratam deste tema. No último ano, a plataforma Google Scholar recebeu uma considerável atenção e alavancou como um método para quantificar os dados literários e fornecer uma visão detalhada do comportamento da pesquisa em relação aos pesquisadores envolvidos [19].

A plataforma Google Scholar Citations foi lançada em julho de 2011, que trouxe a possibilidade de criar um perfil pessoal predefinido para cada pesquisador, com seus indicadores bibliométricos, bem como sua lista de publicações e citações. O produto, feito pelo Google, permite também a criação e edição do perfil pessoal, atribuindo corretamente suas próprias publicações e excluindo duplicatas e outros erros, sendo o pesquisador responsável por manter seus dados atualizados. Dessa forma, ele buscou resolver uma das limitações mais importantes desde o lançamento do Google Scholar, em novembro de 2004: a duplicação e a aplicação de autoria duvidosa, baseado apenas no nome citado, dos registros bibliográficos associados a um pesquisador. Existem diversos homônimos,

bem como formas diversas de citar um mesmo pesquisador ou mesmo, a mesma maneira de citar diferentes pesquisadores, o que ocasiona em divergências no resultado final.

Existem diversas possibilidades de análise para exploração dos dados do Google Scholar, pois, como se trata de uma grande base de dados científicos, é possível ter uma visão clara da Ciência e das relações interdisciplinares fazendo cruzamentos entre as diferentes classes de dados presentes em seu conjunto. Para estudos bibliométricos, a base do Google Scholar tem sido usada de modo complementar, devido ao fato de ser aberta e não necessita de qualquer autenticação ou acesso especial, apesar de toda limitação existente em seu conteúdo [20].

Portanto, considerando os fatores que necessitam ser investigados, o Google Scholar é utilizado como um suporte ao banco de dados principal: currículo Lattes. Nele, é analisado o impacto das publicações dos pesquisadores, baseado no indicador fornecido, o índice-H, unido aos indicadores levantados via Plataforma Lattes, estimado como um pesquisador atua na sua área de formação e qual seu impacto no programa de pós-graduação que atua. Além disso, é possível validar a confiabilidade do conjunto de dados do Currículo Lattes em relação aos dados oferecidos pelo Google Scholar, e vice-versa.

1.4 Outras bases

Para criar um conjunto de dados enriquecido, para expandir o alcance da informação, outras bases serviram de auxiliares para composição de outros dados, bem como validação da informação. Dessa forma, se propõe avaliar a visibilidade da Universidade de Brasília, na área de conhecimento de Engenharia Elétrica, levando em conta fatores nacionais e internacionais, que impactam diretamente nas avaliações dos programas de pós-graduação. Serão utilizadas como fontes de pesquisa auxiliares a Web Of Science, Portal Periódicos Capes e Qualis Capes.

A Web of Science é uma base de dados multidisciplinar, que abrange diversas análises e em 2011 já indexava mais de 11 mil publicações. Essa base é responsável por oferecer alguns dados bibliométricos, como citações, artigos relacionados ao pesquisador, quais publicações foram citadas e por quem, se há interdisciplinaridade e quais redes de pesquisadores são formadas, por meio de suas publicações. A Web of Science também oferece um Relatório de Citações, que contém várias informações sobre o quantos artigos foram citados, o número de citações por artigo, H-index, que é um indicador muito importante para a pesquisa científica, que mede a produtividade e a qualidade da produção científica de um autor [21].

O Portal Periódicos Capes é um portal, de origem brasileira, que consolida a informação científica. Por meio dele, é possível que alunos, professores, pesquisadores e demais

vinculados à instituição conveniada, tenham acesso à produção científica de todo o mundo e de modo atualizado. No Portal se encontram artigos completos de mais de 12 mil revistas nacionais e internacionais, 126 bases de dados com resumos de documentos em todas as áreas do conhecimento [22].

O Qualis Capes é uma métrica formada por indicadores, que visam estratificar a qualidade da produção intelectual dos programas de pós-graduação, levando em conta a publicação em periódicos científicos, que visam avaliar o quão impactante é determinado artigo dentro de uma revista. Essa estratificação é importante para atender às necessidades de padronizar a avaliação dos programas e se baseia em informações fornecidas pelas instituições de ensino superior [23]. A cada quadriênio, uma lista é disponibilizada com a classificação dos meios utilizados pelos programas de pós-graduação para a divulgar sua produção científica. Desta forma, o sistema consegue medir a qualidade dos artigos e de outros tipos de produção propondo uma análise acerca da qualidade dos periódicos em que os pesquisadores publicam.

1.5 Problema

Os bancos de dados formados pelo Currículo Lattes e Google Scholar formam um grande conjunto de dados, porém os dados são separados por pesquisador e não por programa de pós. Diante da necessidade do programa criar estratégias para melhorar seu desempenho nas avaliações, é importante ter uma plataforma que una esses dados e traga uma informação consistente e clara sobre o comportamento do programa.

1.6 Justificativa

Os programas de pós-graduação são constantemente avaliados, gerando métricas a partir de sua produção científica, para verificar a eficiência do programa. Na Capes, um dos critérios para atuar na pós-graduação é que o docente ou orientador seja pesquisador, visto que o conceito da inovação é de suma importância para a avaliação da qualidade do programa, pois o orientador que faz pesquisa, permite que o pós-graduando esteja imerso neste contexto e traga inovações. A qualidade da pesquisa científica é percebida no momento em que o autor e co-autores fazem uma publicação, sendo altamente recomendado agir na ponta, no momento em que é feita uma publicação, para que o impacto do programa seja cada vez maior. É possível que um trabalho seja muito bom e não represente qualquer inovação, e por isso não seja citado, o que não é bom para o programa. Por isso, é importante que os programas tenham essas informações consolidadas, em forma de

indicadores, que indiquem quais artigos recebem mais citações e quais deveriam ser mais citados.

Outro aspecto é que para melhorar o conceito do programa é importante que haja um equílibrio entre os pesquisadores, de modo que, mesmo que haja pesquisadores mais experientes e outros que estão iniciando a jornada acadêmica, estes possam contribuir coletivamente. A distribuição das publicações deve ser feita de forma que corpo docente, em sua maioria, siga com um volume de publicações de determinada qualidade. Não é interessante que apenas uma parte dos docentes publiquem em alta qualidade, enquanto outros são regulares. O programa deve buscar sempre a homogeneidade.

Por fim, outra métrica importante são as orientações do programa. Os índices de orientação não devem ficar concentrados apenas em alguns docentes, enquanto outros se dedicam apenas à pesquisa. É importante que os alunos do programa publiquem junto com seus orientadores, bem como os docentes façam orientações de vários tipos, como em cursos de especialização, mestrado e doutorado, não focando apenas em formar apenas mestres ou doutores, por exemplo.

1.7 Objetivos

O objetivo deste trabalho é propor um *dashboard* consolidado utilizando conceitos de mineração de dados, desde o processo de extração de dados até a visualização final para o usuário, para promover a integração das bases de dados do Currículo Lattes, Google Scholar, bem como de outras bases auxiliares, a fim de evidenciar o comportamento da pesquisa científica, de forma clara e detalhada, nos programas de pós-graduação em Engenharia Elétrica e Sistemas Mecatrônicos. Tal objetivo permite a melhoria dos programas, frente às avaliações da Capes e também, maior integração entre os pesquisadores entre programas distintos.

1.8 Trabalhos Relacionados

Em [1] os autores analisaram o perfil dos pesquisadores de Administração, existentes no Brasil, cuja formação acadêmica era de doutorado. Como base de análise, foram utilizados modelos estatísticos e estatística descritiva pura. Como base de dados, a plataforma Lattes e Qualis Capes foram as únicas fontes a serem consumidas. Foi utilizado um *software* proprietário para tabular e explorar os dados do currículo Lattes, mas não houve desenvolvimento de tecnologia que gerassem as análises automaticamente, apenas algumas análises pré-definidas estiveram disponíveis. Um exemplo pode ser visto na Figura 1.2.

Cidade	# Pesquisadores	# de publicações	% de bolsistas de produtividade	Média de SJR das publicações com SJR	Média de pontos Qualis
São Paulo, SP	404	7007	8,91%	0,094	52,97
Rio de Janeiro, RJ	189	3067	10,05%	0,197	49,90
Brasília, DF	118	1552	4,24%	0,091	48,43
Porto Alegre, RS	109	1922	9,17%	0,074	48,49
Florianópolis, SC	107	2726	6,54%	0,075	43,30
Belo Horizonte, MG	101	2374	11,88%	0,112	47,42
Curitiba, PR	96	1804	6,25%	0,058	40,82
Salvador, BA	71	1067	8,45%	0,046	48,32
Ribeirão Preto, SP	63	1009	9,52%	0,037	45,65
Recife, PE	53	986	5,66%	0,046	39,24

Figura 1.2: Estatísticas de Publicação Científica da Área de Administração em Relação ao Domicílio do Pesquisador. Fonte: [1]

No artigo [24], foi feita uma análise da produtividade científica dos pesquisadores docentes da Universidade Estadual Paulista (UNESP), levando em consideração os dados da Plataforma Lattes, ISI Web e SCOPUS. Foi feita a análise dos programas de pós-graduação de alguns departamentos da instituição, aplicando métodos estatísticos e todo o processo de extração e tratamento de dados foi feito de forma manual. Também não há nenhum processo automatizado para visualização de dados ou que permita que o usuário faça suas próprias consultas. O estudo permitiu concluir que existem dados muito sensíveis, como o H-index, e que este não deve ser usado como único fator de avaliação para mensurar a qualidade da produção científica.

Em [25] os autores utilizaram como fonte de dados a plataforma Sucupira, que segundo o artigo, é o principal meio de comunicação entre os Programas de Pós-graduação Stricto sensu e a CAPES, e a plataforma Lattes. Foi desenvolvida um suporte para prestação de contas utilizando uma ferramenta denominada *Scripsucupira*, que permite extrair diversos tipos de produções científicas, gerar gráficos de colaboração, mapas de geolocalização e relatórios de diversas naturezas, que podem ser acessados por qualquer pessoa e de maneira interativa. Os resultados obtidos mostram que o suporte computacional para gestão em programa de pós-graduação é extremamente útil e permite dar maior dinamismo em atividades que antes se concentravam na mão de certas pessoas.

Em [26] os autores visam demonstrar a eficácia do Scriptlattes, para realizar a construção e apresentação de redes de colaboração entre pesquisadores de Programas de Pós-Graduação Stricto Sensu na área de Administração, Engenharia de Produção e Direito, por meio de um suporte computacional. Os resultados apresentados confirmam que o Scriptlattes é bastante útil para analisar redes de colaboração, bem como identifica a

importância das análises serem facilmente acessadas na internet, como um componente auxiliar ao traçar metas para melhoria dos programas de pós-graduação.

Em [27], foi utilizado o conceito de data mining para extrair dados de pesquisa científica da Universidade Federal de Lavras (UFLA). Como resultado, percebeu que o processo de tratamento de dados é extremamente laboroso, devido à falta de padronização de preenchimento no Currículo Lattes e necessidade de minucioso refino para obtenção de indicadores confiáveis. De modo geral, a Plataforma Lattes se mostrou uma boa fonte de dados para apontar o comportamento da produção científica do universo analisado.

1.9 Descrição dos capítulos

O presente trabalho é apresentado com a seguinte estrutura:

- Capítulo 2: Conceitos em Mineração de Dados, Mineração de Texto e Business Intelligence. Apresenta o conjunto de técnicas e metologias que foram necessárias para o desenvolvimento deste trabalho.
- Capítulo 3: Plataforma de Visualização de Dados Proposta. Discorre sobre a metodologia empregada no trabalho e exemplifica os passos seguidos e necessários para entendimento do trabalho desenvolvido.
- Capítulo 4: Resultados. Retrata as principais análises que podem ser feitas a partir da plataforma desenvolvida.
- Capítulo 5: Conclusão e Trabalhos Futuros. Trata das conclusões retiradas do trabalho e o que pode ser compreendido das análises.

Capítulo 2

Conceitos em Mineração de Dados, Mineração de Texto e Business Intelligence

O presente capítulo apresenta conceitos relacionados a Mineração de Dados, Business Intelligence e Mineração de Textos, que são os temas principais desta monografia. O capítulo está dividido da seguinte forma: Na seção 2.1 são introduzidos os conceitos básicos acerca de mineração de dados e modelos utilizados no desenvolvimento do trabalho, bem como conhecimentos acerca de mineração de textos. E, por fim, a seção 2.2 apresenta os conceitos principais sobre o tema Business Intelligence, trazendo também informações constantes na documentação da ferramenta utilizada na geração da visualização dos dados, neste trabalho.

2.1 Mineração de Dados

2.1.1 Conceitos Básicos

Com o avanço da tecnologia, muitos dados passaram a trafegar pela rede mundial de computadores e um dos principais dilemas é como armazená-los e organizá-los, de modo que seja possível produzir uma informação consumível. Nos últimos anos essa tendência ficou ainda mais evidente, visto que muito se investe para aquisição de hardwares, softwares capazes de transformar os dados e esboçá-los em forma de gráficos, textos, indicadores, tornando possível processar mais e mais dados [27]. Com isso, surge outro conceito atrelado, o de Big Data.

Big Data é um termo utilizado para denominar práticas que envolvem análise e manipulação de um grande volume de dados e de variedade ampla [28]. Devido a isto, es-

truturas mais robustas tiveram de ser desenvolvidas, como os Data Warehouses, surgindo o conceito, posteriormente, de Business Intelligence, que permitia criar uma inteligência aplicada às regras de negócio específicas a serem empregadas caso a caso. Logo, para realizar qualquer atividade, é necessário entender o problema, possuir um volume de dados razoável para identificar padrões e construir um modelo adequado.

O modelo mais antigo existente é o modelo tradicional para transformação dos dados em informação, que consiste em processar manualmente um alto volume de dados, o que pode levar dias, além de demandar a mão de obra especializada, sendo oneroso financeira e temporalmente. Estes especialistas teriam a função essencial de produzir relatórios com várias análises, que seriam estáticas e, caso o consumidor da informação necessitasse de novas, deveria demandar ao especialista, que faria a geração delas. Além disso, deveria ter uma grande infraestrutura formada por profissionais de TI que tratariam os dados e os entregaria aos especialistas após feita toda lapidação. Todo o processo geraria um alto custo de tempo e de profissionais, além de não ser nada prático e aumentar o risco de contaminação da amostra por erro humano.

Na maior parte das situações, dado o volume alto de dados, esse tratamento manual torna-se inviável. Dessa forma, para solucionar o problema da sobrecarga de dados, surgiu o KDD que trata de todo o processo desde a obtenção do dado até a produção do conhecimento. Este processo deve ser feito de modo iterativo e dividido em fases [28]. A Figura 2.1 relata passo a passo o funcionamento do KDD.

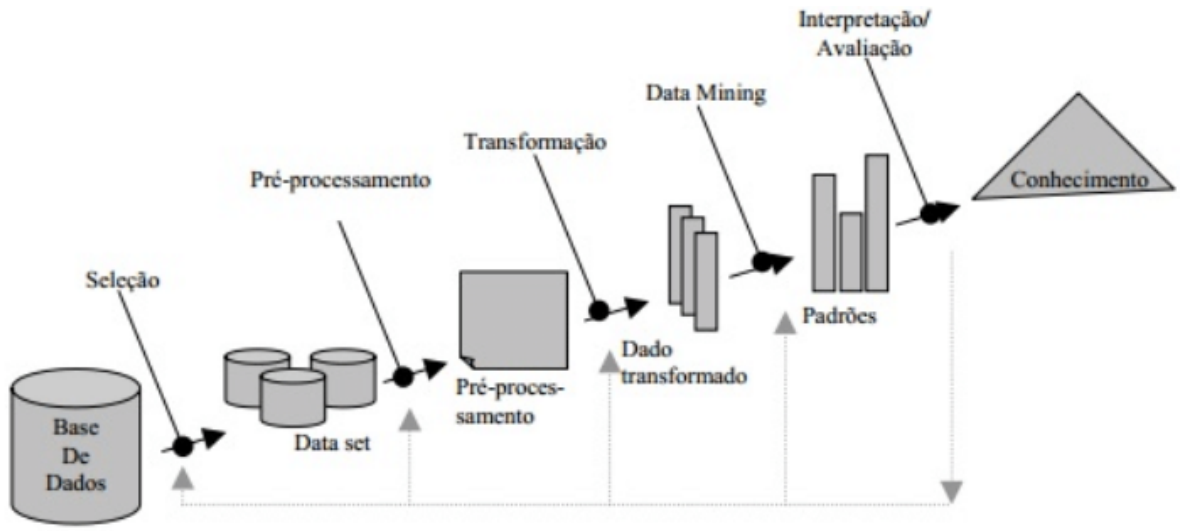

Figura 2.1: Etapas sequenciais do modelo KDD. Fonte: [2]

Além disso, existe o modelo CRISP-DM, modelo que foi escolhido para nortear este trabalho e permitir as análises bibliométricas. Este processo possui 6 fases, que formam um ciclo flexível, dando a oportunidade de ir e voltar várias vezes numa mesma fase, até que se obtenha êxito.

2.1.2 CRISP-DM

Em seu livro, Larose [29] cita detalhadamente as seis fases do modelo. O ciclo de etapas pode ser visto de acordo com a Figura 2.2.

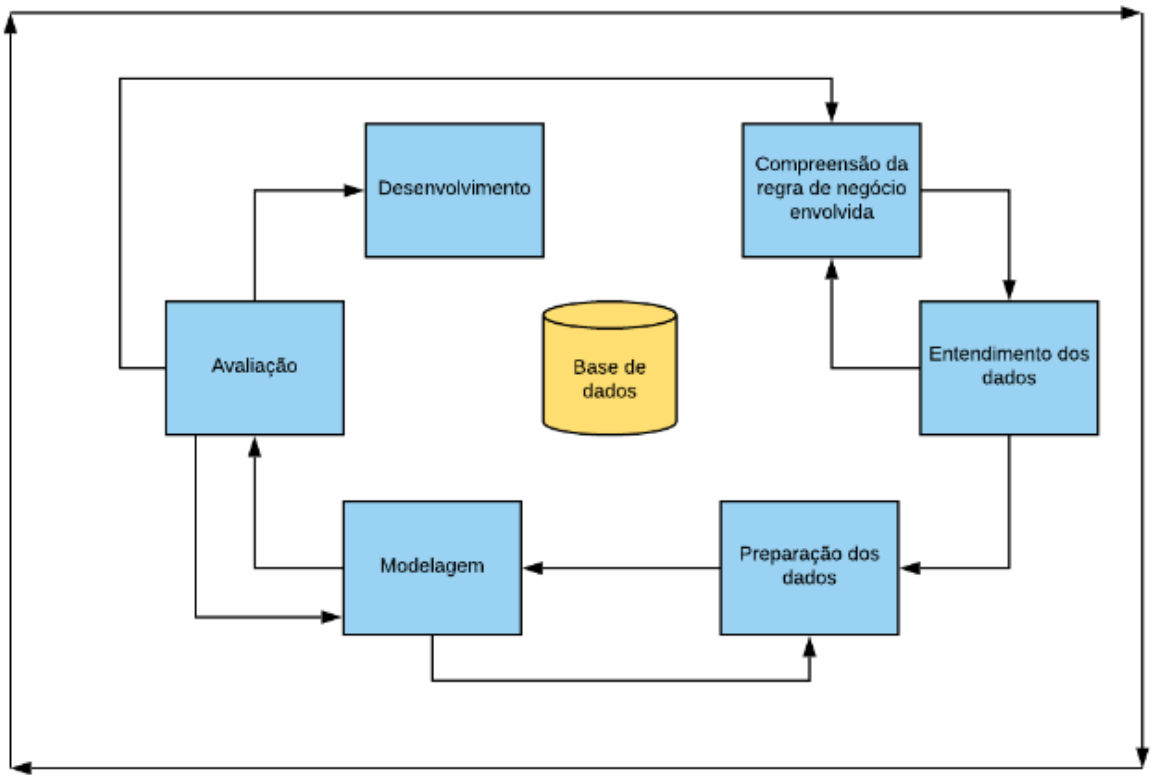

Figura 2.2: Ciclo de atividades do modelo CRISP-DM.

Este modelo, por permitir que algumas etapas sejam retomadas diversas vezes, é bastante flexível quando se deseja trabalhar com dados bibliométricos, como é o caso. De modo geral, suas etapas são:

1. Compreensão da regra de negócio envolvida
2. Entendimento dos dados
3. Preparação dos dados
4. Modelagem
5. Avaliação
6. Desenvolvimento/implementação da aplicação

Compreensão da regra de negócio envolvida

O principal foco dessa etapa é compreender o objetivo a ser alcançado a partir do processo de mineração de dados. As principais perguntas que foram levantadas devem ser respondidas, a fim de obter o máximo de proveito e entendimento das necessidades.

Entendimento dos dados

O foco dessa etapa é compreender o que as fontes de dados são capazes de prover. Não adianta levantar problemas que não possuam solução palpável, pois é a partir do leque de dados disponíveis, mesmo que ainda necessite de um processo de extração e tratamento, ou utilizar modelos de predição para gerar novos dados, que será possível estimar a amplitude que a informação poderá tomar.

Preparação dos dados

Pelo fato dos dados virem de fontes variadas ou mesmo não estruturadas da melhor forma, é nessa etapa que os dados se tornam úteis para cada caso. Ao passar por um processo de extração de dados, os dados podem vir com uma configuração inapropriada. Essa etapa envolve o processo de limpeza dos dados, filtragens, agrupamentos e demais formas de tratamento.

Modelagem dos dados

Feita toda a lapidação do dado bruto, é hora de trabalhar com esses dados e gerar informação. É nessa etapa que os algoritmos de mineração são aplicados. Neste trabalho, foi utilizado um algoritmo de predição, para avaliar qual a probabilidade de um programa evoluir, com base na quantidade de citações.

Avaliação dos dados

Nessa fase, os dados já trabalhados são colocados em gráficos ou outra maneira de visualização, para que seja verificado se as regras de negócio, de fato, estavam corretas. Tanto a etapa que utiliza conceito de Business Intelligence na montagem de dashboards, quanto a etapa do modelo preditivo para dados de citação do Google Scholar, tiveram em sua estrutura análises minuciosas, como avaliação dos indicadores e medidas de acurácia e precisão.

Desenvolvimento/Implementação da aplicação

Nesta etapa, é feita a montagem dos dados em um ambiente interativo com o usuário, para que ele possa ter acesso aos gráficos e aos demais filtros, para assim ter autonomia para fazer as próprias análises. Os resultados são conhecidos em sua totalidade ao fim deste ciclo.

2.1.3 Dados utilizados

É essencial conhecer o tipo de dados com que se irá trabalhar, para determinar a melhor abordagem ou método a ser utilizado. Sabe-se que ao tratar de dados quantitativos, a maioria dos dados serão numéricos, que podem estar disposto de modo contínuo ou discretizado. Existe também a possibilidade de trabalhar com dados qualitativos, que podem conter valores nominais, bem como outros tipos de dados variados. Logo, a mineração de dados será parte elementar para conhecer e ter a capacidade de explorar os dados disponíveis ou mesmo criar novos.

Uma sugestão para aprimorar essa etapa de entendimento dos dados, é utilizar um visualizador ou inseri-los num Sistema Gerenciador de Banco de Dados (SGBD), para que sejam analisados os dados em potencial para responder as perguntas levantadas. No SGBD, é possível ter uma visão geral do que se tem e do que se pode fazer, que pode ser auxiliado por uma plataforma adjacente para plotagem de gráficos e indicadores, para validação dos possíveis resultados.

Nesse ponto, o mais importante é explorar os dados e destacar aquilo que pode ser prejudicial à amostra, como valores com vícios, valores duplicados, valores nulos ou brancos, códigos não pertencentes ao sistema alfabético analisado, entre outros problemas. Tudo isso é passível de contaminar os dados que serão analisados, logo uma busca minuciosa deve ser feita, fazendo diversos filtros e comparações de dados. Essa parte é a mais trabalhosa e demorada de todo o projeto, e deve ser executada com o maior cuidado possível.

Toda essa discussão, pode gerar uma falsa ideia de que o processo de extração de conhecimento e mineração de dados em geral, é totalmente automático e necessita somente de robôs para executar os códigos pré-programados. A verdade é que, apesar de todo o aparato tecnológico e programação disponíveis hoje, ainda são usadas diversas ferramentas como auxílio para executar os algoritmos de mineração. Além do mais, a fase de análise das informações resultantes da mineração de dados, precisa muito da análise humana, para que sejam conferidos os valores e avaliados os parâmetros utilizados nos códigos, que podem ser ou não apropriados para determinados casos.

De modo geral, é inegável que a mineração gera uma contribuição significativa no processo de geração de conhecimento a partir de dados brutos, o que permite que as atividades que antes eram desempenhadas por mão de obra cara e especializada, sejam resumidas a análises mais técnicas e robustas, o que acaba sendo mais produtivo para o próprio especialista contratado, visto que otimiza seu tempo e aumenta sua produtividade.

2.1.4 Tarefas que podem ser realizadas utilizando Mineração de Dados

A Mineração de Dados permite que sejam feitas desde análises descritivas até aplicação de modelos preditivos. Assim, é escolhido um método de acordo com a necessidade do projeto. De modo geral, são tarefas comumente realizadas.

Associações

Essa tarefa consiste basicamente em identificar quais atributos se relacionam e criar associações entre eles, de modo a gerar uma informação parcial ou completa para o caso. É uma das mais comuns e mais utilizadas para tarefas simples do dia a dia. Geralmente, se apresenta na forma SE... ENTÃO, associando atributos entre si. É o famoso exemplo didático que determina onde os alimentos devem ficar dispostos na feira, pois se fulano compra banana, então a probabilidade de comprar maçãs é alta, logo uma estratégia é posicionar as bananas próximas das maçãs, para assim incentivar a compra. A Figura 2.3 esboça um tipo de associação simples.

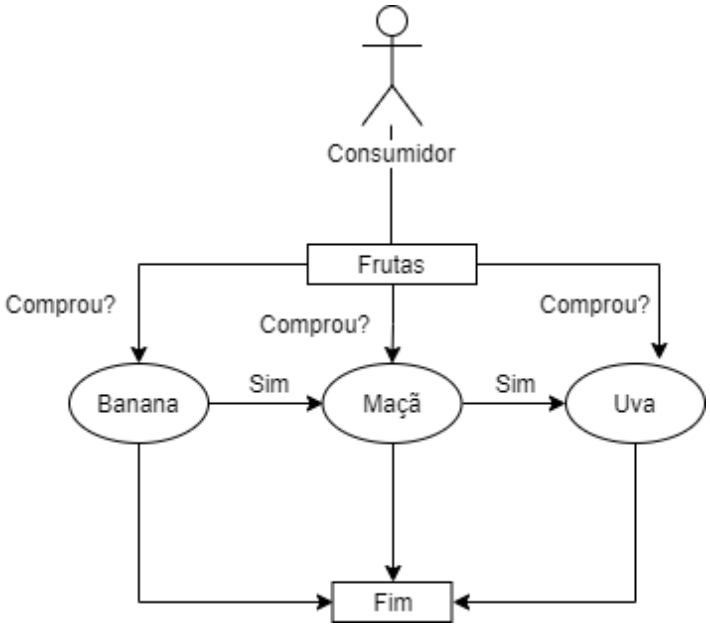

Figura 2.3: Associações entre frutas e desejo do consumidor.

Clusterização

A clusterização, ou também conhecida por agrupamento, é uma tarefa de mineração de dados que visa identificar e unir registros de dados que possuam características similares. Cada cluster é um conjunto de dados semelhantes entre si, mas com conteúdo diferente. É uma técnica útil quando se utiliza aprendizado de máquina não-supervisionado, pois os elementos não precisam ser separados em categorias previamente, como ocorre na classificação para modelos supervisionados. Os clusters podem ser vistos como grandes "bolsas"de dados, como pode ser exemplificado na Figura 2.4 .

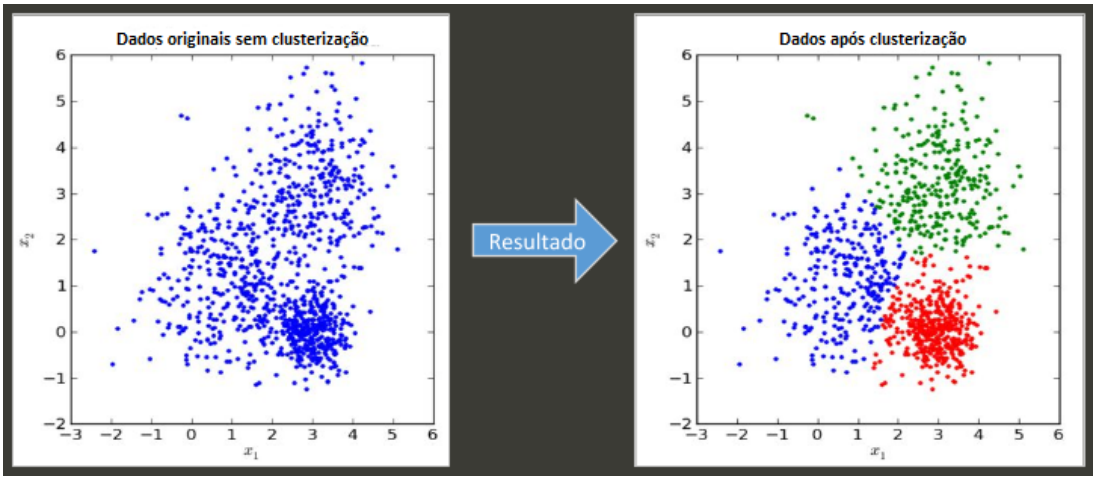

Figura 2.4: Exemplo didático de clusterização. Fonte: [3]

Predição

É uma tarefa cuja ação principal é descobrir um valor que futuramente se adeque a determinado elemento. Um exemplo é uma prática muito comum em mercado financeiro, que analisa o valor de um ativo hoje e diante disso, prediz o valor futuro para uma data fixada, tendo como base uma série histórica de dados. Sua ideia é bastante parecida com a classificação, porém seu foco são valores futuros.

Regressão

É uma técnica, muito relacionada à estatística, que utiliza um registro numérico e estima seu valor a partir do valor das outras variáveis, a cada surgimento de nova informação. Por exemplo, é possível identificar quanto um novo funcionário irá ganhar, com base nos salários e competências de outros funcionários, dispostos numa base de dados. A Figura 2.6 esboça os dados como pontos azuis e a reta central em vermelho indica o valor central a ser encontrado.

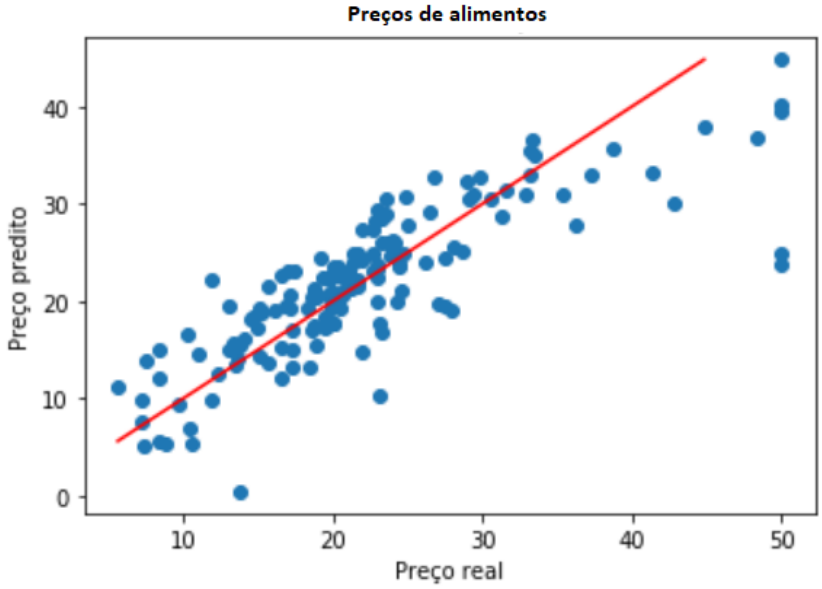

Figura 2.5: Exemplo gráfico do que ocorre com os dados num processo de regressão.

Descrição

É uma técnica bastante simples, relacionada ao estudo de Estatística, que busca relevar padrões e tendências nos dados analisados. A Descrição é uma tarefa muito útil, pois analisa o significado daquela mineração de dados, ou seja, o que os dados, após toda metodologia aplicada, revela de informação relevante.

Classificação

É uma tarefa para determinar a que grupo aquele registro de dado pertence. O modelo de classificação categoriza os conjuntos de dados a fim de elaborar perfis para os dados analisados. Este modelo é bastante utilizado no presente trabalho, bem como a tarefa anterior.

2.1.5 Mineração de Texto

A Mineração de Texto é um estudo na área de Mineração de Dados em constante avanço. Devido à quantidade de dados estruturados ou não, disponíveis para uso, percebeu-se que dados de discursos ou mesmo textos longos fugiam do padrão de análise numérica feito para estatística. Assim, novas áreas começaram a trabalhar com esses dados textuais, como a Informática, Linguística e Ciências Cognitivas [30]. A Estatística utiliza de teorias probabilísticas para explicar eventos e propor estudos e experimentações, logo apresenta seus dados utilizando o modo descritivo. Como necessidade de prever resultados futuros, atrelou-se aos estudos de informática, para utilizar também dados textuais

como objeto de pesquisa. O mesmo caminho seguiu a Linguística, que visa compreender a linguagem humana, que utilizou amparo tecnológico para propor melhores soluções a seus problemas.

Assim, mineração de texto pode ser conceituado como um conjunto de métodos utilizados para garimpar informações em meio a elementos textuais e não categorizados. Existem diversas técnicas para fazer mineração de texto como a indexação, que faz buscas rápidas em textos à procura de palavras-chaves; processamento de linguagem natural ou PNL, que atrelado ao estudo da Linguística, extrai o máximo de conteúdo de um discurso humano, detectando as diversas variações existentes na forma humana de se comunicar; e, simplesmente aplicação de técnicas de mineração de dados, que pode ser usado para processar texto em um banco de dados organizado e pré-processado [30].

Stop Words

Stop Words são as palavras de parada, ou seja, os termos que não possuem relevância para a análise do texto e devem ser retiradas da amostra. Em geral, essas palavras são conectivos e preposições, que representam apenas as ligações entre as palavras sem qualquer significado que interfira no resultado final e que podem contaminar a amostra, visto que ao trabalhar com palavras necessita-se que elas carreguem consigo algum significado que leve a uma interpretação, e essas palavras acabam sendo palavras "vazias"e sem relevância.

As listas de stop words devem ser construídas de acordo com a necessidade, porém existe uma biblioteca padrão, amplamente utilizada neste trabalho chamada NLTK. Esta biblioteca foi usada e em complemento, foi construída uma tabela de stop words auxiliar, visto que haviam idiomas como inglês, italiano, francês, entre outras, na amostra, que compreendem outra lista de stop words. Na Tabela 2.1 encontram-se algumas das *Stop Words* utilizadas neste trabalho. Nesse trabalho elas foram removidas por não possuírem significado para as análises.

Vale ressaltar que, pelo fato de haverem muitos idiomas incluídos nos dados de amostra, foi necessário utilizar várias bibliotecas para limpar adequadamente os dados, o que dificultou bastante a limpeza e tornou o trabalho mais laboroso.

i	me	my	myself	we	our	ours	ourselves
you	your	yours	yourself	yourselves	he	him	his
himself	she	her	hers	herself	it	its	itself
they	them	their	theirs	themselves	what	which	who
whom	this	that	these	those	am	is	are
was	were	be	been	being	have	has	had
having	do	does	did	doing	a	an	the
and	but	if	or	because	as	until	while
of	at	by	for	with	about	against	between
into	through	during	before	after	above	below	to
from	up	down	in	out	on	off	over
under	again	further	then	once	here	there	when
where	why	how	all	any	both	each	few
more	most	other	some	such	no	nor	not
only	own	same	so	than	too	very	s
t	can	will	just	don	should	now	de
de	os	tua	tem	estão	da	lhes	essas
e	é	foi	nossas	muito	o	se	tuas
tu	por	as	sua	aquele	entre	não	ele
delas	minhas	às	nos	pela	havia	me	como
ser	aqueles	nossa	vocês	eu	ter	tenho	suas
está	isso	pelos	estes	tinha	depois	foram	este
para	só	quem	deles	isto	um	eles	do
vos	mais	mesmo	num	dele	será	minha	a
no	teus	à	você	em	meus	esses	pelas
com	ao	dela	há	que	na	nosso	te
aos	dos	ou	aquela	era	uma	das	esta
teu	nem	já	até	seja	esse	mas	quando
aquelas	nossos	têm	também	seus	lhe	meu	seu
ela	elas	estas	nós	sem	essa	fosse	qual
	pela	pelo	nas	numa	aquilo	àquilo	àquela
a	abbastanza	abbia	abbiamo	abbiano	abbiate	accidenti	ad
adesso	affinche	agl	agli	ahime	ahimã	ahimè	ai
al	alcuna	alcuni	alcuno	all	alla	alle	allo
allora	altre	altri	altrimenti	altro	altrove	altrui	anche
ancora	anni	anno	ansa	anticipo	assai	attesa	attraverso
avanti	avemmo	avendo	avente	aver	avere	averlo	avesse
avessero	avessi	avessimo	aveste	avesti	avete	aveva	avevamo

Tabela 2.1: Lista parcial de *Stop Words* utilizadas neste trabalho

2.1.6 Bag of Words

Um dos grandes problemas em se trabalhar com texto é que não existe um formato padrão para eles, durante o processo de extração de dados, o que aumenta o grau de dificuldade em tratar e manipular estes dados. Por isso, quando se busca trabalhar com nuvens de palavras, como é o caso, é necessário aplicar uma técnica que organize esses dados e seja possível extrair informações relevantes, como a quantidade de ocorrências, por exemplo [31].

Neste trabalho, optou-se por utilizar um modelo mais simples de *bag of words* para gerar uma estrutura para iniciar as contabilizações dos fragmentos de texto, no caso, títulos de publicações. Nesse caso, foi gerada uma matriz simples, que apenas indica se determinado termo aparece ou não naquele título. A partir de booleanos, categoriza-se o aparecimento das palavras e a proporção de suas aparições. A matriz representa um exemplo do emprego de *bag of words*. Considerando a frase "Maria gosta de ir à feira"e "João vai à missa", já retirando as stop words, temos:

$$\begin{bmatrix} \text{Maria} & \text{gosta} & \text{ir} & \text{feira} & \text{João} & \text{vai} & \text{missa} \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Analisando isoladamente a primeira frase "Maria gosta de ir à feira", temos a segunda linha da matriz, já analisando a segunda frase de forma isolada, temos a terceira linha da matriz, onde o valor "1"indica que uma dada palavra aparece na frase, e o valor "0"indica que a palavra não aparece.

2.1.7 Term Frequency

Existe uma outra abordagem, comparado ao *bag of words* como forma de representação de dados, com uma performance superior, chamada de *Term Frequency*. Ela considera a quantidade de ocorrências de um determinado termo em relação ao total apresentado. A apresentação matemática desse modelo pode ser vista na Equação 2.1.

$$TF_{termos} = \frac{n_{termos}}{N}, \tag{2.1}$$

Sendo n_{termos} é a quantidade de vezes que o termo apareceu e N é a quantidade total de termos analisados.

2.2 Business Intelligence

O conceito de Business Intelligence, também chamado de Inteligência de Negócio ou BI, como é mais conhecido, tem tomado força nos últimos tempos atrelado ao conceito de Big Data. Esse conceito está presente nas mais diversas organizações, dentro de ambientes universitários e também nas ferramentas aliadas ao mercado de trabalho. O conhecimento de seus indicadores é um fator crítico para a saúde da organização em que se insere, principalmente quando se trata da necessidade de obter rápidas informações para tomar decisões de modo ágil. Devido a isso, muitas organizações não têm economizado esforços para garantir que seus dados estejam disponíveis em qualquer hora e lugar, de maneira atualizada [32].

A aplicação de BI é bastante difundida em várias áreas do conhecimento. Muitas vezes não paramos pra pensar, mas existe uma imensão de dados disponíveis nos processos do dia a dia, que não se consegue gerenciar, devido à quantidade de dados sem qualquer tratamento ou rápida visualização frente ao usuário final. Com o intuito de minimizar essa perda de informações, pela falta de capacidade técnica ou de tempo para processar os dados, é possível desenvolver dashboards ou painéis interativos, para que o usuário se mantenha sempre conectado e bem informado.

Business Intelligence trata-se de um conceito que surgiu em 1996 pela Gartner Research Group. Foi um conceito que já era rotina nas organizações, mas concretizado de maneira diferente do que temos hoje em dia [33]. Com a evolução das tecnologias e sistemas de informação, a perspectiva veio com a ideia de melhorar a qualidade da gestão estratégica nas organizações. Hoje em dia, é raro se ver trabalhar com BI sem estar atrelado a técnicas adjacentes, como processo de extração, transformação, processamento e apresentação de dados. Todas essas etapas dão a oportunidade de oferecer um trabalho final fluido e consistente para expor os dados necessários. Esses sistemas nascem com um propósito de ajudar gestores na tomada de decisão estratégica e fornecer informação em tempo real, e se completam fornecendo um ambiente agradável e prático para todos os interessados acessarem a informação de forma prática e lúdica.

Existem disponíveis no mercado diversas plataformas de BI, sendo elas gratuitas ou não. Atualmente, existe a IBM Watson Analytics, Adobe Analytics, Microsoft Power BI, Tableau, que fornecem ferramentas práticas para que o desenvolvedor seja capaz de gerar análises de dados e criar visualizações, a partir do cruzamento dos dados. Para escolher qual a melhor ferramenta, é necessário conhecer as necessidades, valores e aplicabilidade ao seu negócio.

Para este trabalho, foi escolhido o Microsoft Power BI, devido ao fato de haver uma parceria entre a Universidade de Brasília e a Microsoft, para que fosse disponibilizado gratuitamente o Office 365 para os alunos, o que facilitou bastante a escolha, visto que

com a licença, é possível publicar os trabalhos na Web e tornar disponível suas análises para quem desejar utilizar a plataforma.

2.2.1 Power BI

O Power BI é uma plataforma completa que oferece uma coleção de serviços, aplicativos e conexões com SGBDs, nuvem como a Azure, para executar análises com ou sem uso de programação robusta. O Power BI possui compatibilidade com diversas fontes de dados, e a escolhida para este caso, foram os dados diretamente do banco de dados em nuvem Azure. Os dados em nuvem já estão limpos e tratados, após processo completo de limpeza e tratamento executado previamente, com auxílio das linguagens C#, Python e R, além de acessos direto ao banco de dados, utilizando como SGBD o Microsoft SQL Server.

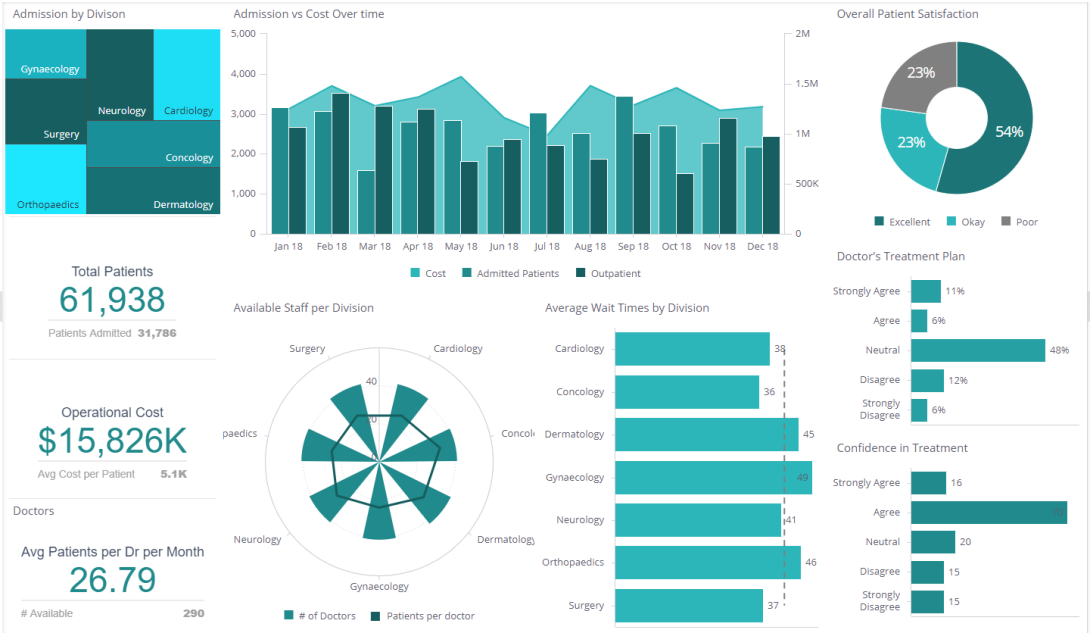

Figura 2.6: Exemplo de painel do Power BI. Fonte: Documentação Power BI

O Power BI fornece uma série de visualizações como padrão, e utilizando o serviço como desenvolvedor, é possível ainda importar novas *views* e gerar outras informações, indicadores, gráficos e análises, por meio do suporte oferecido ás linguagens R e Python, que exige uma habilidade de programação e também conhecimentos estatísticos. Para este trabalho, foram utilizadas outros tipos de gráficos e análises, utilizando o suporte para Python e também para linguagem R.

Foram desenvolvidos elementos visuais personalizados a algumas análises propostas, como por exemplo, a identificação das palavras com maior ocorrência nos títulos das publicações, para compreender em quais delas geram um maior Qualis ou menor; para determinar quais áreas possuem maior incidência de artigos de maior impacto. Além disso,

é possível ver as palavras mais evidentes nos artigos mais citados e quais assuntos geram mais citações. Como se tratam de análises que exigem uso de técnicas mais robustas de mineração de dados, optou-se por utilizar Python e R para programação, pela maior afinidade com as bibliotecas do ramo.

Capítulo 3

Plataforma de Visualização de Dados Proposta

Neste capítulo, é detalhado o projeto de uma plataforma de suporte computacional para os programas de pós-graduação da Universidade de Brasília, capaz de classificar professores e fornecer uma visão detalhada dos programas de pós-graduação que fazem parte. De modo geral, o projeto foi dividido em 6 blocos básicos, conforme pode ser visto na Figura 3.1. Os blocos serão discutidos detalhadamente das seções 3.1 a 3.6.

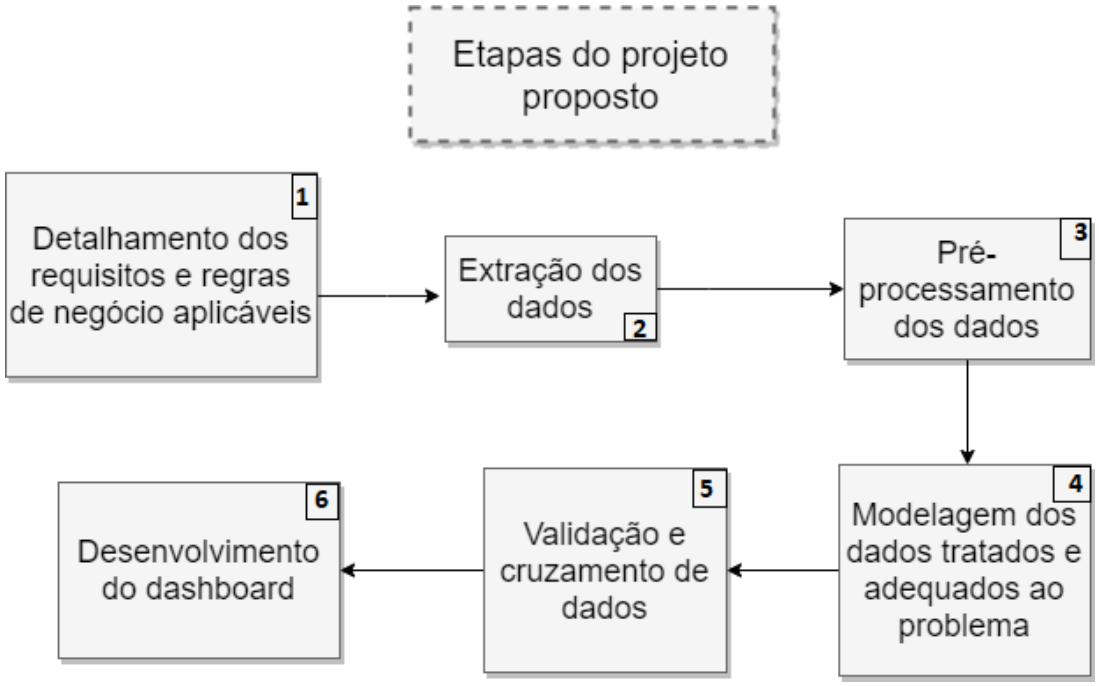

Figura 3.1: Diagrama de blocos do dashboard proposto para classificação de professores e programas de pós-graduação da Universidade de Brasília

3.1 Detalhamento dos requisitos e regras de negócio aplicáveis

O primeiro bloco da Figura 3.1 trata do detalhamento dos requisitos e regras de negócio aplicáveis ao projeto. Neste bloco, são delineados os problemas a serem resolvidos e as possíveis soluções, definindo quais informações seriam relevantes para estar presentes na plataforma final e atender as necessidades do usuário.

Para o caso, foi feita uma reunião com coordenadores de programas de pós-graduação da Universidade de Brasília, para entender quais informações são importantes para ter em fácil acesso a qualquer momento. Estas informações ficariam disponíveis em um painel interativo, com atualizações periódicas.

Algumas das informações solicitadas foram:

1. Quais professores recebem mais citações?
2. Quais artigos possuem mais visibilidade?
3. Em quais cursos o professor faz mais orientações?
4. Há interdisciplinaridade nas publicações?
5. Há uma tendência de quantidade de publicações ano a ano? Ou o pesquisador faz publicações relevantes, porém em baixa frequência?
6. O professor faz orientações focada em uma titulação ou orienta em vários títulos?
7. O pesquisador faz muitas publicações sozinho ou possui co-autores?
8. Suas publicações são em que idioma? Tem potencial para ser lida internacionalmente?
9. Quais assuntos são de domínio do pesquisador? Poderiam ser criadas redes de colaboração?
10. Como o programa, em geral, publica em periódicos? Há concentração de publicações em um grupo específico de professores?
11. Quais temas geram publicações de maior impacto? E menor?
12. Quais assuntos recebem mais citações?
13. Em quais publicações o pesquisador deve receber mais citações para aumentar seu H-index?

Tais informações são relevantes, tendo em vista que a avaliação feita pela Capes, leva em conta muitos desses quesitos e, além disso, uma ferramenta adequada que responde a estas perguntas, pode trazer inovação ao programa, visto que se identifica pontos fortes e fracos, dando a oportunidade de melhorias das diretrizes já traçadas pelas coordenações. Por exemplo, sabendo que determinado assunto está em alta e tem recebido bastantes citações, é interessante gerar aplicações inovadoras tomando como base este assunto, para que outras pessoas percebam este artigo e gere uma rede de citações. Outro ponto é que muitas vezes um autor publica um trabalho bastante relevante para área científica, porém em um idioma pouco falado.

Devido a estes pontos, essa etapa se torna tão importante e indispensável, pois é neste momento que é possível identificar as carências e melhorar o que pode ser aperfeiçoado, dentro do contexto dos programas de pós-graduação. Finalizada a etapa, será aplicada a etapa de extração dos dados, que é parte do segundo tópico do ciclo CRISP-DM, onde é feita a compreensão dos dados e do potencial que estes dados têm para responder as perguntas propostas na etapa 1.

O diagrama indicado pela Figura 3.2 detalha os passos a serem seguidos ao final dessa etapa.

Figura 3.2: Etapas do processo inicial de aquisição de dados.

Tabela 3.1: Tabela base para iniciar a Mineração de Dados. Descrição da tabela professor_id

Dado	Tipo de Dado
id_lattes	numeric
nome_lattes	string
nome_scholar	string
programa_pos	string

Este ciclo pode voltar ao Passo 1, caso haja a necessidade de inserir um novo pesquisador.

A Tabela 3.1 descreve a classificação das colunas da primeira tabela a ser utilizada, denominada professor_id.

3.2 Extração dos dados

Nessa etapa, os dados para nomear os pesquisadores já foram levantados, agora há a necessidade da extração dos dados do Currículo Lattes e do Google Scholar. De posse dos dados da tabela professor_id, descrita pela Tabela 3.1, o próximo passo é baixar os currículo um a um, na Plataforma Lattes. Às vezes, a Plataforma Lattes retira o reCaptcha da página e é possível utilizar um código de extração, consolidado na literatura, como *scriptLattes*, que lê os códigos de ID, da tabela inicial professor_id e extrai os currículos em XML para a pasta escolhida. Como isso é a minoria dos casos, a extração realizada em Setembro/2019 foi em parte realizada pelo *script* e a outra parte de acordo com os passos a seguir, demonstrado na Figura 3.3.

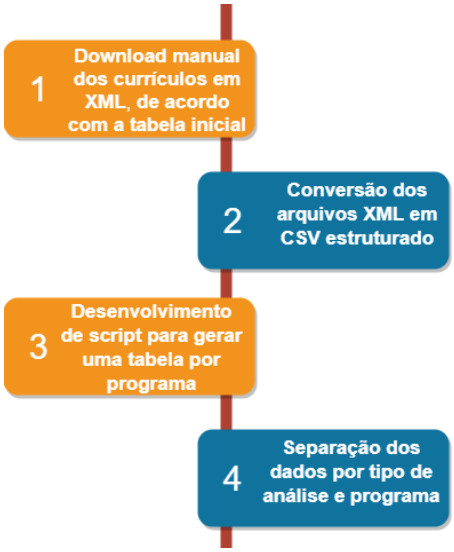

Figura 3.3: Processo inicial de extração dos dados.

Inicialmente, os currículos baixados vêm no formato .zip. Para isso, desenvolveu-se um script capaz de extrair o XML da compressão zip e a partir da leitura desse XML, extrair as tags com as informações desejadas. Por meio desta extração, foi gerada uma tabela para cada programa, de acordo com a tabela inicial professor_id.

Feito isso, é realizado um processo de ETL para separar as informações, que estão agrupadas em um único arquivo .csv, resultado dessa extração. Logo, esse arquivo .csv se divide em um arquivo .csv para cada análise (orientações, conferências e periódicos) e para cada programa. No total, 12 tabelas são geradas.

3.2.1 Extração da Plataforma Lattes

Para fazer o crawler dos currículos Lattes, foi utilizada a biblioteca BeautifulSoup. Esta biblioteca foi escolhida pelo fato de ser capaz de extrair dados de arquivos HTML e XML. Como os arquivos da base Lattes são em formato XML, é possível obter êxito em todo o processo.

Foram extraídos dados de pesquisa, orientação, financiamento, conferências, periódicos, livros e capítulos publicados. Pelo fato de essa ferramenta de suporte computacional abranger apenas dados que afetam a avaliação da qualidade de pesquisa, preferiu-se focar em 3 métricas: orientações, conferências e periódicos.

Ao fim do processo, foi percebido que havia a necessidade de acesso aos dados de citações dos pesquisadores. Como o currículo Lattes, não oferece esse recurso, foi necessário fazer um novo crawler com os perfis disponíveis no Google Scholar.

3.2.2 Extração da Google Scholar

Levantados os requisitos novamente, visto que foi necessário voltar para etapa 1 da Figura 3.1 após o processo de extração de dados, seguindo o proposto pelo CRISP-DM, foram detalhados os dados que seriam necessários para enriquecer a nova base de dados, com informações do Google Scholar, seguindo os passos disponíveis na Figura 3.4.

Para esse processo de extração, foram utilizadas as bibliotecas do Selenium WebDriver, pandas, random e time, todas do Python. As bibliotecas do Selenium WebDriver são, originalmente, para fazer testes automatizados, mas são capazes de fazer acesso direto ao navegador, tornando possível produzir ações automatizadas. Para este trabalho, foram feitos acessos diretos aos perfis do Google Scholar, utilizando os IDs captados para localização da página de cada pesquisador.

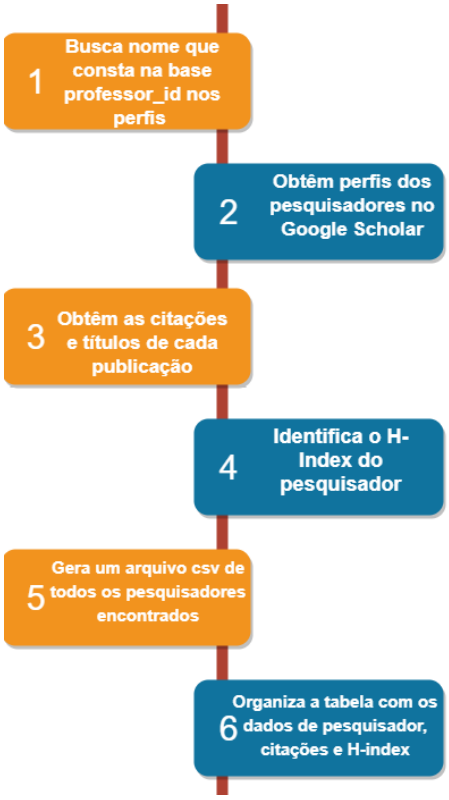

Figura 3.4: Esboço do processo de extração dos dados de pesquisador e citações, do Google Scholar.

Foram coletados dados de 31 pesquisadores, de um total de 64 professores listados como pertencentes aos programas de pós-graduação, da amostra avaliada. Logo, é possível pautar que apenas 50% possuem perfis cadastrados no Google Scholar ou com nome facilmente identificável. Após busca manual para validação e corretude dos dados, apenas 1 professor foi encontrado, totalizando uma amostra de citações para 32 pesquisadores.

Como resultado, foi gerado um arquivo .csv com os dados conforme Tabela 3.2. Os dados extraídos possuem algumas inconsistências e devem passar por um processo de validação, que será feito posteriormente.

Tabela 3.2: Tabela gerada após processo de extração de dados do Google Scholar.

Dado	Tipo de Dado
nome	string
ano	integer
hindex	integer
hindex2014	integer
citacoes	integer
titulo	string

Em complemento, já no processo de tratamento dos dados, os dados deverão ser remodelados, haja vista que o formato importado não é útil para as consultas. Como as bibliotecas utilizadas fazem acesso direto ao sistema da página, existem muitos dados comprometidos e que precisam ser tratados.

O processo de extração do Google Scholar foi realizado em Setembro/2019. Os dados estão na versão mais atualizada até este mês.

3.3 Pré-processamento de dados

O terceiro bloco da Figura 3.1 corresponde ao pré-processamento de dados. Pelo fato de tanto o Google Scholar quanto a Plataforma Lattes permitirem que os usuários apresentem seus dados de múltiplas maneiras, em diversos idiomas, com campos para adicionar dados textuais sem temática definida, é necessário passar pela fase de entendimento desses dados, e proceder para que eles tenham condições de serem utilizados na etapa subsequente.

Não há padrão de escrita definido para ser usado no preenchimento dos campos, logo a limpeza deve ser minuciosa. Esta etapa deve ser realizada diversas vezes, até que não sejam encontradas inconsistências nos dados. Além disso, deve ser ressaltado que o que se analisa são dados inseridos pelos próprios pesquisadores, que no caso do currículo Lattes, representam toda a jornada acadêmica do pesquisador. O fluxo de limpeza de dados pode ser visto na Figura 3.5.

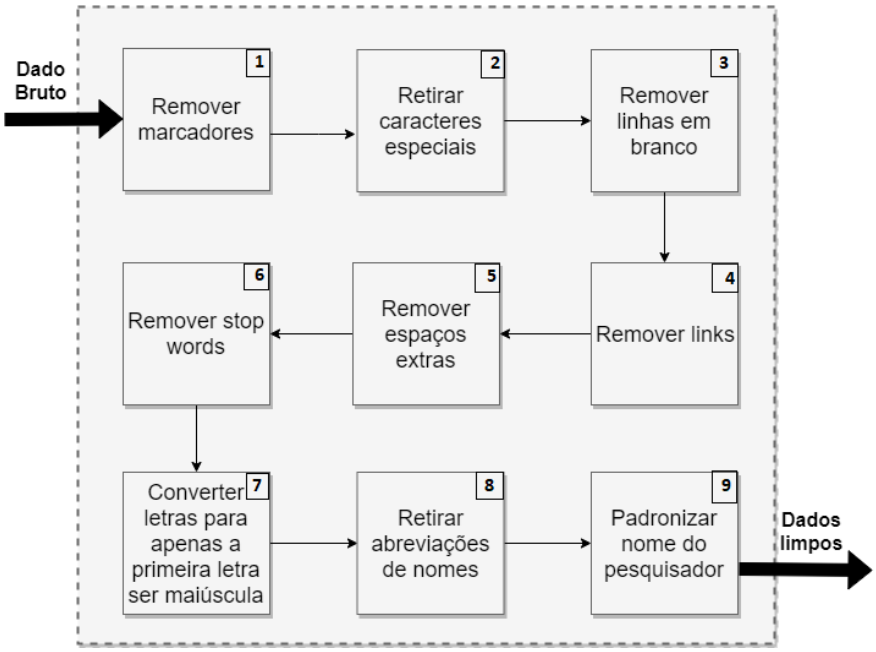

Figura 3.5: Fluxo de pré-processamento dos dados.

Foram definidas nove etapas para a limpeza dos dados, conforme pode ser visualizado na Figura 3.5. A primeira etapa é responsável por remover as tags que vierem juntos aos dados, durante a etapa de extração por acesso direto ao HTML (caso do Google Scholar) e as que porventura vierem na leitura do XML (caso do currículo Lattes) e qualquer marcador que estiver entre as palavras.

A segunda etapa tratam da remoção de caracteres especiais, como pontuações, acentos, letras que não fazem parte da codificação UTF-8, entre outros. Em seguida, são removidas todas as linhas em branco que constam no arquivo .csv. Após, faz a remoção de alguns links de sites, que podem estar presentes nos registros cujos campos não são definidos, além de remover os espaços desnecessários entre as palavras.

O sexto passo é remover todas as stopwords presentes nos dados textuais que serão utilizados para gerar informação. Por haver uma variedade de idiomas entre as publicações, foi necessário utilizar stopwords para vários idiomas, o que tornou essa etapa repetitiva. O próximo passo foi converter todas as letras do texto para que apenas a primeira ficasse maiúscula e as demais minúsculas. O oitavo passo foi retirar as abreviações dos nomes dos professores nas bases de Currículo Lattes, para que estivesse adequado caso fosse necessário cruzamento com outras bases.

Por fim, o último passo, no caso o nono, surgiu diante de uma necessidade de padronizar o nome no pesquisador para adequar o cruzamento dos dados com as bases auxiliares. Como não existe um ID padronizado ou valor numérico fixo para identificar o pesquisador, a chave utilizada foi o nome, por isso a necessidade de prezar pela padronização.

Na Tabela 3.3, mostra um exemplo da limpeza de um registro selecionado aleatoriamente do *dataframe* utilizado para análise.

No apêndice **??** o código utilizado para a limpeza e estruturação dos dados do currículo Lattes é apresentado.

3.4 Modelagem e Tratamento de Dados

O quarto bloco da Figura 3.1 corresponde à modelagem dos dados, que devem estar tratados e adequados para o problema. Neste trabalho foram utilizadas, principalmente, as bibliotecas pandas e numpy, para o processamento dos dados. Os dados foram modelados de acordo com a necessidade de utilização, separados em tabelas fato e tabelas dimensão.

Tabela 3.3: Exemplo de limpeza de dados extraídos do Google Scholar

Dado Bruto	RAFAEL TIMOTEO DE SOUSA JR, Acquisition# of </DIGITAL evidence in Android Smartphones/> Disponível em https://www.researchgate.net/
1) Remoção marcadores	RAFAEL TIMOTEO DE SOUSA JR, Acquisition# of DIGITAL evidence in Android Smartphones Disponível em https://www.researchgate.net/
2) Retirar caracteres especiais	RAFAEL TIMOTEO DE SOUSA JR, Acquisition of DIGITAL evidence in Android Smartphones Disponível em https://www.researchgate.net/
3) Remover linhas em branco	RAFAEL TIMOTEO DE SOUSA JR, Acquisition of DIGITAL evidence in Android Smartphones Disponível em https://www.researchgate.net/
4) Remover links	RAFAEL TIMOTEO DE SOUSA JR, Acquisition of DIGITAL evidence in Android Smartphones
5) Remover espaços extras	RAFAEL TIMOTEO DE SOUSA JR, Acquisition of DIGITAL evidence in Android Smartphones
6) Remover stop words	Acquisition DIGITAL evidence Android Smartphones
7) Apenas primeira letra maiúscula	Rafael Timoteo De Sousa Jr, Acquisition Digital Evidence Android Smartphones
8) Retirar abreviações de nomes	Rafael Timoteo De Sousa Junior, Acquisition Digital Evidence Android Smartphones
9) Padronizar nome do pesquisador	Rafael Timóteo De Sousa Júnior, Acquisition Digital Evidence Android Smartphones

3.4.1 Modelagem dos dados

Data Warehouses são repositórios de dados, que abrigam um conjunto de dados tratado e organizado. Ele visa acolher dados de alta qualidade, ou seja, depois de todo tratamento e modelagem necessários para resolução do problema [34]. Assim, pelo fato do conjunto de dados utilizados neste trabalho terem assuntos diversos, como atividades em orientações, publicações em periódicos, artigos em conferências e dados, é necessário ter mais que uma tabela fato. Para isso, optou-se por trabalhar com um modelo multidimensional, gerando cubos de dados para responder algumas perguntas. A Figura 3.6 esboça o conceito da matriz de dados multidimensional relatada como cubo.

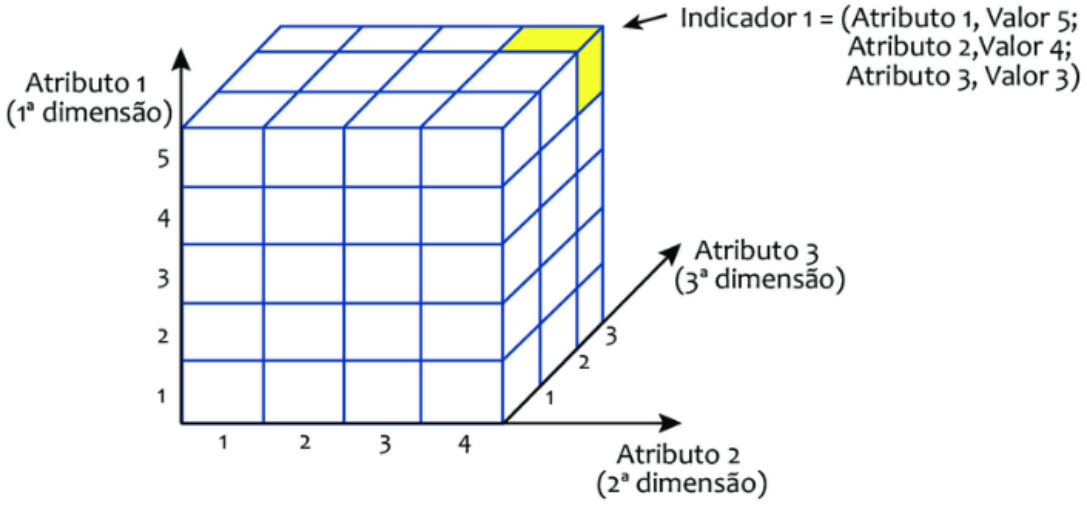

Figura 3.6: Exemplo de cubo de dados. Fonte: [4]

Os dados que fazem parte dessa matriz multidimensional, exemplificada pelos cubos, são construídos pelas relações entre as tabelas fato e dimensão. As tabelas fato representam um assunto do data warehouse e possui chaves para conectar-se aos às perspectivas ou atributos das tabelas dimensão, como exemplo: Um pesquisador se relaciona com N publicações e com N orientações. Logo, este projeto utiliza o esquema Estrela, bastante simples, que consiste em uma tabela fato se relacionando a algumas tabelas dimensão [35].

Dessa forma, a partir da tabela central professor_id exemplificada pela Tabela 3.1, foi possível gerar as conexões posteriores e cruzamentos posteriores.

Os dados do currículo Lattes foram modelados de modo distinto aos dados obtidos via Google Scholar. Os dados da Plataforma Lattes, foram dispostos em 12 tabelas dimensão, que segmentavam os programas e as áreas de estudo. Os dados do Google Scholar, foram dispostos em 2 tabelas dimensão, para os dados de H-index e citações.

A Figura 3.7 representa todo o processo desde o início até o processo de modelagem dos dados, acrescido das etapas subsequentes de tratamento, validação e desenvolvimento do dashboard.

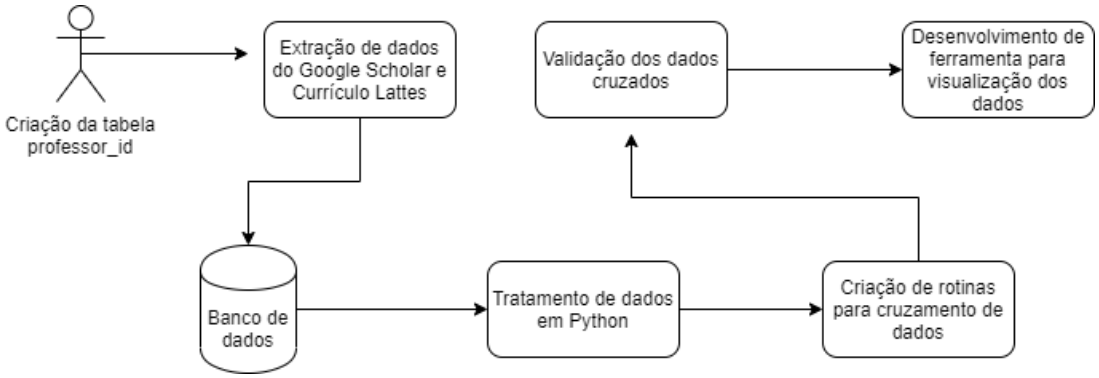

Figura 3.7: Diagrama elucidativo dos processos dos dados para o projeto.

3.4.2 Tratamento de dados

Nessa etapa foi feito o cruzamento dos dados e criação das rotinas automatizadas para validação dos dados. Como apoio para fase de tratamento foi usado o SGBD SQL Server, onde foram feitas consultas em SQL para gerar as informações para serem base de análise para a ferramenta de visualização. Ao conjunto de dados foi inserida uma tabela consolidada com as informações de Qualis, para o quadriênio 2013-2016, fornecida pela Plataforma Sucupira, da CAPES, na área Engenharias IV para os programas na área de Engenharia Elétrica e de Computação, base para os programas PPGEE, PGEA e PPEE; e Engenharias III para as áreas de Engenharia Mecânica, que é base do programa PPMEC, também analisado.

Durante a etapa de tratamento foram levadas em consideração as particularidades de cada indicador, bem como a necessidade da informação e o formato que ela deveria se apresentar. De modo geral, foram utilizadas as linguagens Python e R, para tratar dados e gerar análises para validação, bem como consultas via script SQL em banco de dados. A base de dados obtidas pelo Google Scholar representou a maior dificuldade da fase de tratamento de dados, visto que, por haver diversos homônimos e a mesma forma de citar diferentes pesssoas, existem artigos associados ao pesquisador, mas que de fato não pertencem a eles. Para tratar isso, foi feito um cruzamento com os arquivos presentes na base do Currículo Lattes e, os que não foram encontradas, foram novamente filtrados para verificar se tratava-se de uma publicação da área do pesquisador que não foi atualizada na base Lattes. Após essa etapa, os dados foram liberados para etapa de validação.

Vale ressaltar que, os cruzamentos feitos nessa etapa, tiveram o objetivo de verificar se as informações necessitariam ainda de alguma limpeza ou geração de informações adicionais. Nessa etapa, foi gerado um dado adicional para as tabelas extraídas do Google Scholar. Utilizando algoritmos de classificação, foi possível definir quais eram os artigos que teriam maior potencial para aumentar o H-index do pesquisador. Como parâmetros, foram utilizadas a quantidade de citações e o ano da publicação, levando em consideração que publicações mais novas tendem a ter um grau de inovação maior que publicações antigas, e por isso podem ser citadas com maior facilidade. Além disso, foi levado em consideração o fato de que é mais interessante receber citações em publicações cujo valor é igual ou próximo ao H-index, do que tentar receber várias citações em um artigo pouco citado e elevar o indicador a partir dessa publicação [36].

Os resultados encontrados estão disponíveis no Capítulo 4.

3.5 Validação dos dados

O quinto bloco da Figura 3.1 corresponde à validação e cruzamento dos dados, com o intuito de gerar uma informação consistente e fornecê-la ao usuário via plataforma de visualização.

Após todo processo de modelagem e tratamento dos dados, é necessário fazer a validação da qualidade do que foi produzido. Essa validação pode ser feita por meio da observação visual do dado original e daquele que foi extraído, para verificar se nenhuma informação foi perdida, e também pode ser realizada fazendo cruzamento dos dados com o auxílio de um SGBD e observando quantitativa e qualitativamente se os dados apresentam coerência e margem de erro controlada, que é de suma importância ao trabalhar com determinadas métricas [37].

A etapa subsequente de desenvolvimento da plataforma para visualização dos dados só se inicia após todos os dados apresentarem corretude. Após feito todos os cruzamentos, para todas as análises, é necessário validar se eles estão corretos e não apresentam qualquer duplicidade ou imprecisão. Devem ser verificadas também o tipo de dados que possui e como eles podem ser mostrados, como por exemplo, em forma de gráficos, tabelas, entre outros.

Foi utilizado o príncipio de Pareto para realização da validação, onde 20% dos dados foram separados como amostra para validar a fase de modelagem e tratamento e os outros 80% foram utilizados posteriormente para compor o conjunto e gerar os cruzamentos necessários [38].

3.5.1 Cruzamento de dados sobre orientações

Ao analisar os dados de orientações, foram levados em consideração os parâmetros de natureza da orientação, se houve financiamento ou não durante o trabalho, quantas orientações são feitas por ano para cada programa e por cada pesquisador, qual o tipo dela, se é orientador principal ou co-orientador, entre outras. Tais análises são oriundas das informações obtidas exclusivamente na Plataforma Lattes.

Natureza da orientação

Neste caso, foi realizada uma contagem das orientações. Vale ressaltar que houve a preocupação de contar apenas orientações distintas, após validada a inexistência de inconsistência nos dados.

As orientações por natureza foram divididas em:

- Trabalho de Conclusão de Curso para Graduação
- Iniciação Científica
- Dissertação de mestrado
- Tese de doutorado
- Monografia em Especialização
- Outra Natureza

A classificação "Outra natureza"é oriunda da classificação "Outros"feita pelo pesquisador no preenchimento de seu currículo Lattes. As contagens foram feitas considerando um único pesquisador, um grupo selecionado ou mesmo de todos os pesquisadores membros do programa. Após obtenção de êxito em relação à contagem dos dados originais, os dados foram validados e estruturados para serem apresentados em forma de gráfico de barras.

Orientação recebeu financiamento?

Este tópico visa tratar os dados para validar o estudo acerca do financiamento da orientação realizada. Em geral, as orientações que tendem a receber mais financiamento são as de iniciação científica, mestrado e doutorado, pelo fato de agências de fomento incentivarem essas pesquisas com o pagamento de bolsas para os orientandos. Esta análise visa verificar quais as naturezas de orientação mais recebem financiamentos e as que menos recebem.

Após cruzamento dos dados, para verificar a consistência, foi constatado que os dados condizem com os dados originais fornecidos pelo XML da Plataforma Lattes. Logo, estão apropriados para o uso e construção dos cubos de informação.

Orientações por Ano

Essa fase de validação dos dados, para este caso, visa compreender a disposição das orientações ano e ano, bem como qual é o comportamento dos pesquisadores do programa ao realizar orientações. Se em determinado ano houve pico de orientações de determinada natureza, enquanto em outros anos houve brusca diminuição, é preciso avaliar o que ocorreu no programa nessas datas.

Os cruzamentos analisados nessa fase levaram em consideração orientações realizadas a partir do ano de 1970 e estão em conformidade com o XML original. Utilizando o SGBD SQL Server, foram feitas consultas para validação dos dados. Os dados foram tratados de modo se adequar ao cubo de informações proposto.

Tipo de orientação

Nessa fase é analisado como o pesquisador se comportou no decorrer do projeto de orientação. Existem duas opções:

- Orientador principal
- Co-orientador

Para cada programa, foi feito o cruzamento por scripts SQL e também validado o dataframe por meio de códigos em linguagem Python e R. Os dados de tipo de orientação também foram validados com o XML original e modelados para compor o cubo de dados.

Orientações por Instituição

Essa validação foi feita por meio da comparação dos dados, cruzados com o XML original que fornece dados brutos sobre as instituições em que foram feitas as orientações. Essa análise considera a quantidade de orientações feitas nas instituições, permitindo compreender se o orientador possui uma participação ativa em outras universidades ou foca seus trabalhos apenas nas dependências da Universidade de Brasília.

Foram necessários alguns reparos nos dados das instituições, que possuíam algumas inconsistências nos nomes. Após resolução, os cruzamentos foram feitos e os valores validados e prontos para serem utilizados nas análises.

Orientações por Curso

Nessa fase, foi levantada a quantidade de orientações por curso. Um orientador pode atuar em vários cursos, aumentando o grau de interdisciplinaridade do programa. Isso serve para medir quão expressivas são orientações do pesquisador no curso/área do programa em que se analisa e será avaliado.

Como forma de validação, foi criado um programa em Python, que verifica a quantidade de cursos no dataframe disponibilizado, no caso, o XML original. A partir desse cruzamento, foram extraídos os dados do XML do currículo Lattes e feita uma quantificação dos termos iguais, para determinar quantas orientações em cada área.

3.5.2 Cruzamento de dados sobre conferências

Para validar os dados sobre conferências, foram considerados os idiomas utilizados para publicação, os anos em que foram feitas, se publica como primeiro autor ou aceita coautorias, o impacto que um pesquisador tem dentro do programa em número de publicações e as palavras mais utilizadas nos títulos das publicações aceitas em conferências. Os dados coletados foram em sua totalidade obtidos por meio da Plataforma Lattes.

Publicações por Idioma

A validação desses dados serviu de subsídio para verificar quais idiomas são mais utilizados nas publicações em conferências. Essa análise facilita na identificação dos padrões dos pesquisadores dentro dos programas e se existe uma preferência de idioma em determinado ano ou posição de autoria.

Os dados foram validados por meio de scritps SQL, que cruzaram os dados obtidos com os valores disponíveis no XML original. Os valores quantitativos foram conferido duas vezes, um por meio do SQL e por meio de código em Python, para conferência do dataframe.

Publicações por Ano

Foram validadas as publicações a partir do ano de 1970 e gerado um dado quantitativo ano a ano, para verificar se há consistência com os dados do XML original. A quantidade de publicações foi conferida e os dados preparados para fase final de visualização dos dados. Essa análise permite traçar o perfil do programa ou dos pesquisadores selecionados e compreender se há regularidade nas publicações durante todo ano, que é uma métrica importante para a CAPES, no âmbito de avaliação.

Posição de Autoria

Os autores podem publicar sozinhos ou em coletivo. É ideal que um pesquisador compartilhe publicações com seus alunos ou mesmo entre colegas de seu progrma, pois, como forma de avaliação, isso é bem visto, o que pode contribuir para que o conceito do programa aumente. O entendimento é que um programa onde todos os docentes são qualificados, aumenta a chance de um aluno ser orientado por um professor capacitado e que haja uma homogeneidade no acesso ao potencial de inovação.

Se houver concentração de publicações em um só pesquisador, esta se torna uma métrica prejudicial, pois mostra que a fonte de "pontuação"do programa concentra-se em um só ponto. Desse modo, essa fase de validação visa verificar se os dados, para classificar a posição do pesquisador de acordo com a quantidade de autores em uma publicação, tem capacidade de gerar a informação que se deseja e apresentam consistência para tal. Após cruzamento de dados por meio de consultas SQL e programa na linguagem R, foi possível verificar que os dados apresentam corretude. Na etapa anterior, de tratamento e modelagem dos dados, foi criado um algoritmo para conferir a posição do autor dentro de uma lista de autores, e informar sua posição dentro dessa lista. Desse modo, foi possível qualificá-lo como primeiro, segundo, terceiro, até o n-ésimo autor de um artigo.

Total de publicações

Para validar os dados de publicações em conferências, foi feita uma quantificação dos títulos já publicados. Um algoritmo em Python, desenvolvido para buscar o título, conferir com o XML original e retornar a quantidade de registros válidos, foi executado e os resultados indicaram uma precisão de 99% entre os dados tratados e o dado original. Dessa forma, retornou-se para a etapa de tratamento de dados para indicar os dados que foram perdidos e fornecer a informação com completude. Com os dados completos, a etapa de validação mostrou que os dados apresentam consistência e qualidade para formar os cubos de informação, que serão dispostos em gráficos e tabelas na fase de desenvolvimento da plataforma de visualização dos dados.

Palavras mais utilizadas em títulos

Para medir quais palavras apareceram em mais publicações, o que enfatiza a ideia dos temas que mais são aceitos em conferências, é necessário gerar uma lista que faça a gestão da frequência de ocorrência de cada palavra. Para isso, foi feita uma validação dos títulos, por meio de algoritmo na linguagem Python, que gerava uma lista de dados paralela com a quantidade de ocorrência de cada palavra, removidas as stop words e as classificavam como alta, média e baixa ocorrência. Dessa forma, foram feitas validações acerca dessas

palavras e se os resultados condiziam com o esperado. Assim, como forma de facilitar a confirmação desses dados, foi gerada uma wordcloud, que é muito importante para este tipo de análise [39], para compreender se os dados têm potencial de gerar informação.

Assim, os dados obtidos foram comparados com os esperados e foram validados com sucesso, estando prontos para gerarem uma visualização clara e precisa ao usuário.

3.5.3 Cruzamento de dados sobre citações

É muito importante conhecer o impacto das publicações e a proporção das citações dos professores do programa. Diante disso, uma forma de aumentar o conceito do programa, é aumentando um índice muito utilizado para quantificar a produtividade dos pesquisadores, o H-index. Para aumentar este índice, é necessário que os artigos sejam citados. Um tática é criar uma rede de colaboração para que os membros do programa possam fazer citações entre si, e aumentar seus índices. Para determinar quais algoritmos devem receber citações, durante esta fase, foi desenvolvido um algoritmo em Python, com auxílio da biblioteca Scikit-learn, para determinar quais as chances de um artigo receber citações e com isso indicar em qual artigo deve buscar receber citações. Por exemplo, um pesquisador possui o H-index geral igual a 10 e o H-index a partir de 2014 em 8, logo é possível que um artigo de 2014 para frente, com a quantidade de citações próximas ao valor do H-index, seja o responsável por aumentar este indicador. Os dados obtidos para essa análise se encontram disponível na plataforma Google Scholar.

Publicações e citações por ano

A validação desses dados é feita através do cruzamento entre as informações de citações, publicações e do ano de cada uma delas, propondo uma visão detalhada com comportamento do pesquisador durante uma série histórica analisada. Os dados referente a esta análise foram adquiridos no mês de Setembro/2019, e estão em conformidade com os dados originais, disponíveis nos perfis de cada pesquisador.

H-index total e H-index 2014+

Os valores foram obtidos por meio de crawler diretamente na página de perfil de cada pesquisador. Assim, foi realizado um teste automatizado para verificar a consistência desses dados, que não passaram por um processo de transformação, visto que durante o processo de extração, foram obtidos em sua totalidade e completos. Por ser uma medida quantitativa, é possível mostrá-la por meio de gráficos na ferramenta de visualização que será desenvolvida.

Citações por ano

Essa validação busca cruzar os dados entre a quantidade de citações e o ano que o artigo foi publicado. Isso informa se existe um pico de citações em artigos mais recentes ou se há crescimento harmônico ano após ano, ou mesmo estagnação no número de citações dos artigos mais antigos. Os dados foram validados por meio de cruzamentos realizados por consultas SQL e algoritmos em Python e R, bem como análise manual em acesso direto aos perfis dos pesquisadores, onde se encontra disponível a informação.

Palavras mais utilizadas em títulos

Para medir quais palavras apareceram em mais publicações, o que enfatiza a ideia dos temas que mais recebem citações, assim é necessário gerar uma lista que faça a gestão da frequência de ocorrência de cada palavra. Para isso, foi feita uma validação dos títulos, por meio de algoritmo na linguagem Python, com a quantidade de cada palavra, sem stop words e contabilizada sua ocorrência. Dessa forma, foram feitas validações acerca dessas palavras e também se os resultados condiziam com o esperado. Assim, como forma de facilitar a confirmação desses dados, foi gerada uma wordcloud, para compreender a informação gerada pelos dados.

Artigos que devem receber citações

Conforme definido anteriormente, para que o H-index aumente, é necessário aumentar citações em determinados artigos. Para isso, foi desenvolvido um algoritmo em Python, utilizando dados de citações coletados nos meses de junho, julho, agosto e setembro de 2019, a fim de gerar uma série histórica e utilizar o algoritmo de classificação Naive Bayes, para determinar a probabilidade de receber citações nesse artigo, a partir de citações já ocorridas. Este algoritmo é implementado de forma simples, já que leva em conta a probabilidade condicional para os valores de saída a partir da entrada definida. A probabilidade condicional pode ser calculada como:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{3.1}$$

Também é levado em conta que é mais interessante receber citações em artigos cuja quantidade de citações está próxima do valor do H-index, bem como considerada a hipótese de que artigos dos últimos 5 anos têm chances maiores de serem citados, caso o H-index 2014 desse pesquisador esteja próximo do H-index geral. Assim, foram obtidos resultados dentro do esperado por meio do cruzamento desses dados e aplicação do algoritmo de classificação Naive Bayes. Para fins de comparação entre o artigo que foi predicto

e o que de fato recebeu alguma citação, foi calculada a precisão do algoritmo para aquele conjunto de dados, dado por meio da equação [40]:

$$Precisão = \frac{PC}{PC + PE} \tag{3.2}$$

onde PC indica o total de previsões corretas e PE o total de previsões errôneas.

3.5.4 Cruzamento de dados sobre periódicos

Para validar os dados sobre periódicos, foram considerados os idiomas utilizados para publicação, os anos em que foram feitas, se publica como primeiro autor ou aceita coautorias, o impacto que um pesquisador tem dentro do programa em número de publicações, palavras mais utilizadas nos títulos das publicações aceitas em periódicos e a proporção de publicações por estrato. Os dados coletados foram obtidos por meio da Plataforma Lattes e do cruzamento dos dados do Qualis, fornecido pela Plataforma Sucupira, que contém informação de ISSN, Evento e Estrato.

Publicações por Idioma

A validação desses dados serviu para verificar os idiomas mais utilizados nas publicações em periódicos. Essa análise facilita na identificação dos padrões dos pesquisadores dentro dos programas e se existe uma preferência de idioma em determinado ano ou posição de autoria. Os dados foram validados por meio de scritps SQL, que cruzaram os dados obtidos com os valores disponíveis no XML original.

Publicações por Ano

Foram validadas as publicações a partir do ano de 1970 e gerado um dado quantitativo ano a ano, para verificar a consistência com os dados originais. A quantidade de publicações foi conferida e validada. Essa análise permite traçar o perfil do programa ou dos pesquisadores selecionados e compreender se há regularidade nas publicações durante todo ano, que é uma métrica importante para a CAPES, no âmbito de avaliação.

Posição de Autoria

Conforme explicado na subseção 3.5.2, os autores podem publicar sozinhos ou em coletivo. O pesquisador deve compartilhar suas publicações com seus alunos ou mesmo entre colegas de seu progrma, para contribuir para o aumento do conceito do programa. Entende-se que um programa onde todos os docentes são qualificados, aumenta a chance de um aluno ser orientado por um professor capacitado e que haja uma homogeneidade

no acesso ao potencial de inovação, por isso o aumento do conceito. Os dados foram validados e preparados para fase final de visualização de dados.

Total de publicações

Para validar os dados de publicações em periódicos, foi feita uma quantificação dos títulos já publicados. Um algoritmo em Python, desenvolvido para buscar o título, conferir com o XML original e retornar a quantidade de registros válidos, foi executado e os resultados indicaram que os dados estavam precisos e consistentes. A etapa de validação mostrou que os dados apresentam qualidade o suficiente para formar os cubos de informação, de acordo com [41] que serão dispostos em gráficos e tabelas na fase de desenvolvimento da plataforma de visualização dos dados.

Palavras mais utilizadas em títulos

Para medir a aparição de palavras, o que enfatiza a ideia dos temas que geraram mais citações, é necessário gerar uma lista com a frequência de ocorrência de cada palavra. Para isso, foi feita uma validação dos títulos, por meio de algoritmo na linguagem Python. Dessa forma, foram feitas validações acerca dessas palavras e os resultados estiveram em conformidade com o esperado. Assim, como forma de facilitar a confirmação desses dados, foi gerada uma wordcloud, para compreender a informação gerada pelos dados.

Publicações por Estrato Qualis

O Qualis é uma forma de estratificar a produção científica. Logo, é possível detectar se o artigo faz parte dos estratos: A1 (o mais alto), A2, B1, B2, B3, B4, B5 e C. Geralmente, os artigos A1 tratam de assuntos da mais alta relevância, seguido dos estratos posteriores, o que caracteriza o potencial de inovação presente no programa. Logo, um programa que possuem bastante publicações de estrato A, é um programa que produz pesquisas de alto impacto e por isso, deve ter seu conceito aumentado. Por isso, uma estratégia para aumentar o conceito do programa de pós-graduação é possuir várias publicações de estratos superiores em sua composição. Nesta etapa, foram cruzados dados do currículo Lattes e do Qualis CAPES para verificar a quantidade de publicações por estrato, atrelados às métricas mencionadas anteriormente. Os dados foram validados e verificados por meio de consultas em SQL ao banco de dados de origem. A quantificação das publicações por estrato apresentou precisão e consistência.

3.6 Desenvolvimento de plataforma para visualização de dados

Após realizar o tratamento dos dados, o dataset está pronto para ser visualizado. Para isso, deve ser desenvolvida uma plataforma para visualização de dados. O foco dessa ferramenta é servir de suporte computacional para gestão dos programas de pós-graduação, que permitirá que o usuário tenha acesso a todos os indicadores, em qualquer local, devido ao suporte online da ferramenta.

A plataforma foi desenvolvida utilizando os conceitos de Business Intelligence, Big Data, unido ao conceito de Machine Learning, no caso da utilização do algoritmo de classificação Naive Bayes para sugerir qual artigo deve receber mais citações para que o H-index aumente. A ferramenta escolhida para desenvolvimento foi o Power BI, visto que ela possui suporte à conexão com o banco de dados em nuvem da Azure. Assim, os dados que serão consumidos pelo Power BI serão apenas resultados das consultas já realizadas previamente na nuvem, o que tende a gerar visualizações mais rápidas [42]. O esquemático desse projeto pode ser visto na Figura 3.8.

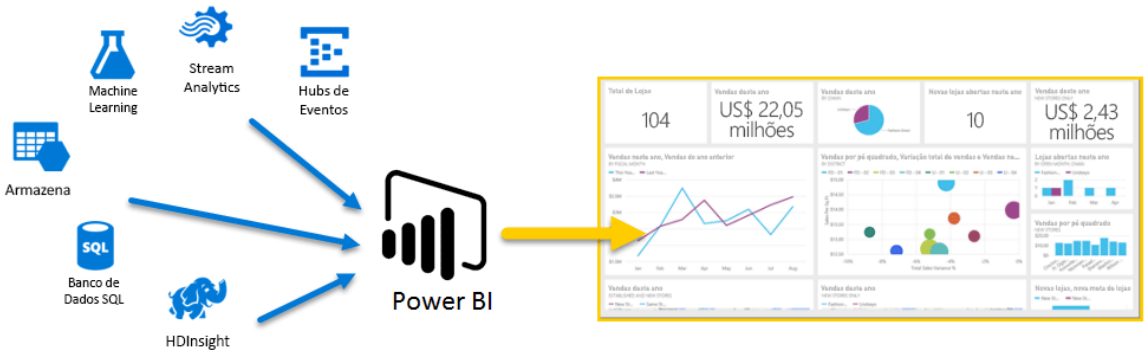

Figura 3.8: Esquema representativo acerca das etapas envolvidas na visualização com Power BI. Fonte: [5]

A ferramenta de criação dos relatórios é chamada Power BI Desktop, e tem possui pacotes de visualizações com vários tipos de indicadores, para que sejam escolhidos os tipos mais adequados para a análise desejada. A representação gráfica de dados visa representar os resultados obtidos de forma clara, simples e com veracidade, permitindo que se chegue às conclusões sobre como se relaciona o conjunto de dados ou mesmo verificar a evolução do evento observado [43]. Dessa forma, foi feito um estudo para verificar a forma mais clara e simples, que evite criar viés, para que os dados sejam visualizados.

Feito isso, os gráficos foram posicionados e associados aos parâmetros obtidos durante as consultas ao banco de dados, sendo escolhidos aqueles que melhor expressaram a informação desejada. Os resultados obtidos estão disponíveis no Capítulo 4.

Primeiramente, os dados são colocados em nuvem e passam por todos os processos de limpeza, tratamento e validação citados anteriormente. Dessa forma, todos os cruzamentos de dados são executados em nuvem e a plataforma de visualização apenas consome os resultados dessas consultas, otimizando o processamento, que é uma vantagem de utilizar um banco de dados em nuvem.

3.6.1 Escopo do projeto

Foram definidos, na primeira etapa da Figura 3.1, os requisitos e detalhamento para o desenvolvimento de uma ferramenta de suporte computacional aos programas de pós-graduação, e entre eles estavam o fato de que esta ferramenta deveria ser online para que pudesse ser acessada de qualquer lugar e a qualquer momento. Assim, foi desenvolvida na parte de visualização de dados, uma ferramenta que tivesse como características básicas a facilidade e a rapidez em consultar indicadores importantes para definir métricas e tomar decisões com agilidade, durante uma reunião, por exemplo.

Desse modo, foi levantadas as quatro análises principais, sobre: orientações efetuadas, atuação em conferências, publicações em periódicos e citações em artigos publicados. A partir disso, foi criada uma página por programa, para cada análise, gerando 16 páginas de relatórios, acrescida de mais uma página inicial contendo atalhos para cada página. Devido ao baixo número de pesquisadores com perfil no Google Scholar, fonte de dados para análise de citações, foi decidido manter apenas uma página geral, com informações sobre os quatro programas analisados. Por fim, o projeto final manteve 14 páginas no total, sendo uma páginal inicial e mais 13 páginas de análise para os quatro programas.

O esboço do projeto pode ser visto na Figura 3.9. No próximo capítulo serão analisados os resultados obtidos após todo o processo apresentado.

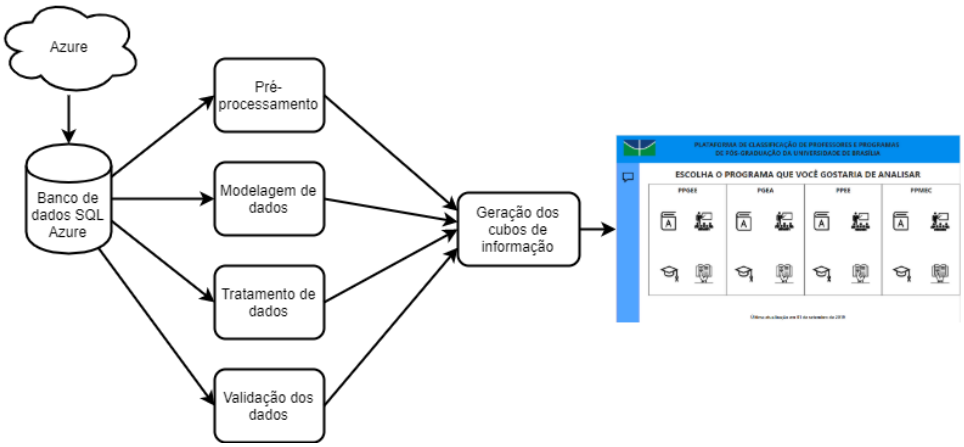

Figura 3.9: Esboço do projeto final para desenvolvimento de ferramenta para visualização de dados.

Capítulo 4

Resultados

Esta seção apresenta os resultados obtidos a partir das análises realizadas, bem como o resultado final da plataforma de visualização dos dados. Foram comparados os resultados fornecidos pela plataforma desenvolvida, com os resultados das consultas ao banco já validados na etapa de Validação dos Dados.

4.1 Página inicial do dashboard

Foi criada uma página principal para exposição dos atalhos, que redirecionam o usuário para a análise de sua preferência. Os atalhos filtram a análise por programa, que por sua vez, permite fazer estudos com todos os professores do programa, ou mesmo com apenas alguns selecionados, para avaliar o impacto daquele pesquisador no programa. A página inicial pode ser vista na Figura 4.1.

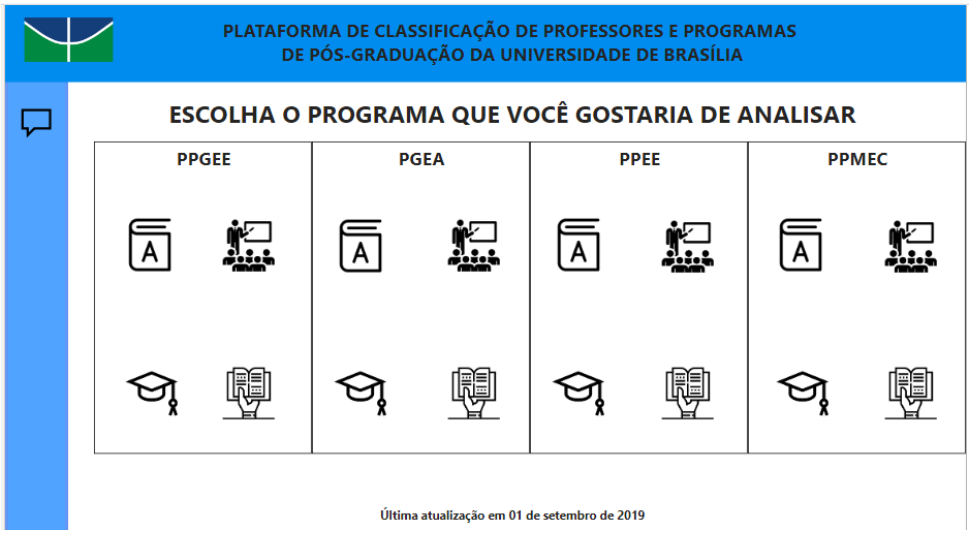

Figura 4.1: Página inicial da plataforma de classificação de professores e programas de pós-graduação da Universidade de Brasília.

Essa tela conta com ícones que sinalizam os estudos de orientações, conferências, citações e periódicos, respectivamente, para cada um dos programas selecionados: PPGEE, PGEA, PPEE e PPMEC. Basta clicar sobre o ícone, que será redirecionado para a opção escolhida.

4.2 Análise de orientações

Conforme dito anteriormente, a avaliação do programa feita pela CAPES visa a permitir que o aluno que ingresse em um curso de pós-graduação seja capaz de ser orientado por professores de alta qualidade e competência e, a partir disso, melhorar junto ao seu orientador a sua produção intelectual. Desse modo, é muito importante compreender se o professor está atuando adequadamente como orientador, para manter a excelência e melhoria contínua do programa que participa.

Levando isso em consideração, foi desenvolvido um relatório pensado nas necessidades de compreensão da produtividade dos professores no quesito de orientações. Logo, foram gerados indicadores diversos para responder as perguntas levantadas em etapa anterior.

Foi criado um relatório para o programa PPGEE, que pode ser visto na Figura 4.2 considerando as seguintes perguntas:

- Os professores fazem orientações de quais naturezas?
- Existe algum tipo de financiamento durante o período de orientação?
- Quantas orientações são feitas por ano e a natureza delas?
- O professor tem atuado mais como orientador principal ou co-orientador?
- As orientações estão concentradas na Universidade de Brasília ou orienta em outras instituições?
- Em quais cursos houve participação com orientador principal ou co-orientador?

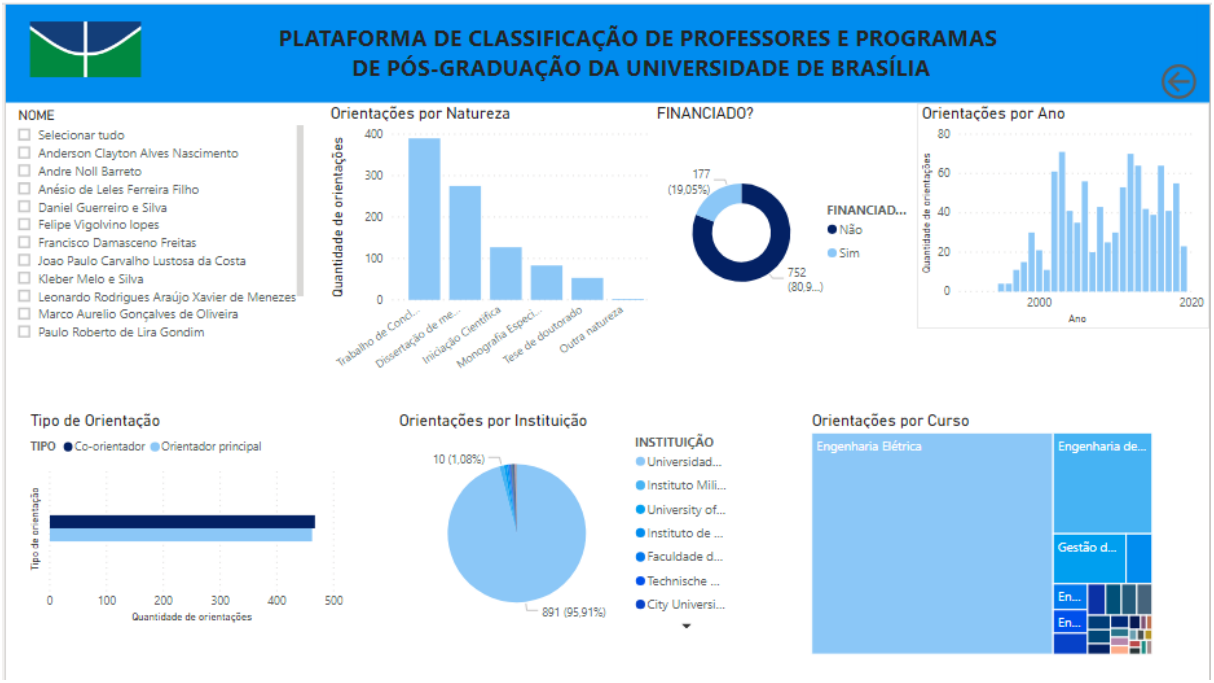

Figura 4.2: Relatório sobre orientações para o programa PPGEE.

Primeiramente, foi colocado um filtro de segmentação para que fosse possível selecionar todos os participantes do programa e também analisar somente aqueles que forem de escolha. Isso ajuda a perceber qual o impacto do professor em relação aos indicadores gerais, visto que foram utilizados modelos gráficos que possuem um efeito de transparência que permite ver a proporção do resultado atual sobre o resultado total. Os gráficos também podem ser aumentados para que sejam exportados e utilizados para outros fins.

4.2.1 Orientações por natureza

O programa PPGEE, numa perspectiva geral, é composto por um corpo docente com maior enfoque em: orientações em Trabalho de Conclusão de Curso, para a graduação, seguido de orientações em dissertações de mestrado, iniciação científica, monografia de especialização, tese de dourado e outros, respectivamente. O gráfico de colunas, da Figura 4.3 demonstra esse comportamento para o PPGEE, porém todos os programas analisados demonstram que a principal natureza de orientação é nos trabalhos de conclusão de curso, até mesmo pela quantidade de alunos que ingressam ser maior do que as outras.

No PGEA, em ordem decrescente de quantidade de orientações, temos: Trabalho de Conclusão de Curso, Iniciação Científica, Dissertação de Mestrado, Tese de Doutorado, Outra natureza e Monografia em Especialização.

No PPEE, a ordem decrescente é: Trabalho de Conclusão de Curso, Monografia em Especialização, Dissertação de Mestrado, Iniciação Científica, Tese de Doutorado e Outra

natureza.

Já no PPMEC, a ordem é Trabalho de Conclusão de Curso, Dissertação de Mestrado, Iniciação Científica, Outra natureza, Tese de Doutorado e Monografia em Especialização.

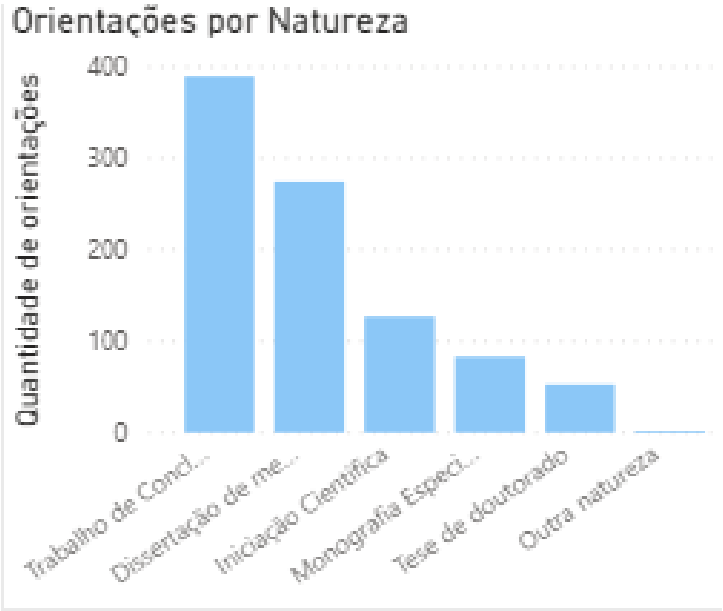

Figura 4.3: Gráfico de orientações por natureza geral do programa PPGEE.

As orientações para graduação, em geral, iniciação científica e trabalho de conclusão de curso, podem ser convidativas para ingresso nos programas de pós-graduação. Então, uma estratégia para aumentar a captação de alunos na pós-graduação, seria um investimento maior em orientações em nível de graduação. Além disso, é importante avaliar se os professores focam seus esforços em naturezas de orientação variadas ou se atuam incisivamente em alguma natureza. Isso pode ser feito facilmente com a ajuda do filtro localizado à esquerda da tela.

Um fato interessante sobre esta análise é que é possível verificar uma relação entre a orientação de iniciação científica com a orientação em dissertação de mestrado. Os professores que possuem iniciação científica em quantidade razoável, também possuem interessantes índices de orientações em mestrado. O fato pode demonstrar uma tendência dos alunos que ingressam na iniciação científica, continuarem seus estudos no mestrado.

4.2.2 Financiamento durante a orientação

Existem agências de fomento que incentivam a pesquisa, financiando o período em que se realiza o estudo. Assim, foi feita uma análise para verificar em quais naturezas, há maior ocorrência de financiamentos. Os dados encontrados podem ser demonstrados na tabela 4.1. No total, 19,05% das orientações feitas pelos professores do programa

receberam financiamento, enquanto as outras 80,95% não receberam. Percebeu-se que há mais financiamentos em iniciação científica e mestrado, o que pode explicar também o fenômeno dos alunos da iniciação científica continuarem os estudos até o mestrado, devido às bolsas e financiamento da pesquisa.

Tabela 4.1: Tabela com a natureza da orientação e a proporção de financiamentos recebidos para o programa PPGEE.

Natureza da orientação	Qtd de financiamentos	Porcentagem
TCC Graduação	7	0,75%
Iniciação Científica	73	7,86%
Monografia Especialização	0	0
Dissertação de mestrado	76	8,18%
Tese de doutorado	21	2,26%
Outra natureza	0	0

Para o PGEA, há uma maior quantidade de orientações financiadas, se comparado ao PPGEE. No PGEA, 36,87% das orientações são financiadas, enquanto no PPEE, apenas 16,29% recebem subsídio. Já o PPMEC, possui 41,43% das orientações financiadas.

4.2.3 Orientações por Ano

Pelo fato da plataforma online possuir interatividade entre seus gráficos, é possível a partir de uma análise, obter outros parâmetros. Logo, é possível verificar ao mesmo tempo, quantas orientações tiveram por ano, por natureza, curso, tipo de orientação e instituição. Logo, uma análise importante é a do comportamento do pesquisador em relação a sua atuação. Por meio de realce no gráfico, é possível ver a proporção de uma análise em relação a outra, conforme Figura 4.4.

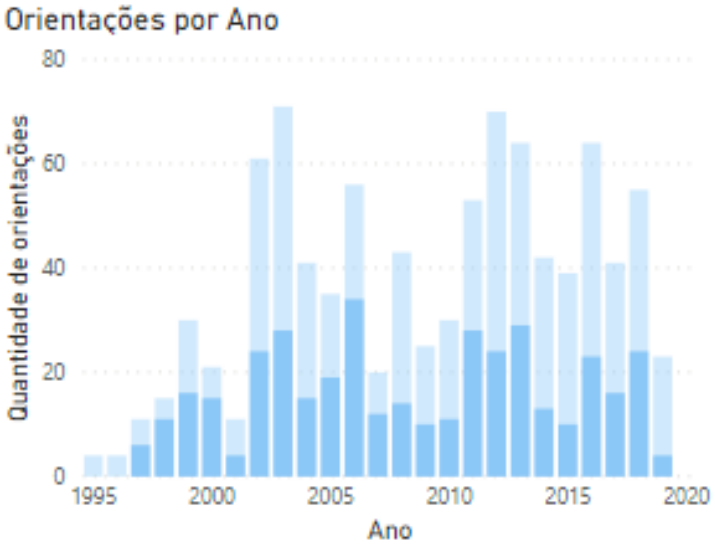

Figura 4.4: Gráfico de orientações em TCC em relação à quantidade total ano a ano do PPGEE.

Além dessa análise, é possível fazer diversas outras, devido ao fato de trabalhar com a ideia de tabelas fato e dimensão, onde todas se comunicam e propõem interatividade a partir de qualquer clique nos gráficos definidos. A partir disso, é possível ver o impacto que os trabalhos de conclusão de curso causam na quantidade de orientações.

4.2.4 Tipo de orientação

O pesquisador pode atuar como orientador principal ou co-orientador em trabalhos. Assim, com a ideia de cubo de informação, pode-se comparar que o maior número de co-orientações se devem a orientações em curso distintos ao de origem. Por exemplo, pesquisadores do programa de pós-graduação em Engenharia Elétrica possuem orientações, como co-orientadores, na Engenharia Mecatrônica, Ciência da Computação, Ciências da Saúde, entre outras. Isso traz interdisciplinaridade para o programa, e com isso, aumenta sua rede de colaboração, incidindo diretamente na quantidade de citações de suas produções.

Além disso, a proporção entre orientações principais e co-orientações é bem dividida, porém com maioria de co-orientações, conforme pode ser visto na Figura 4.5.

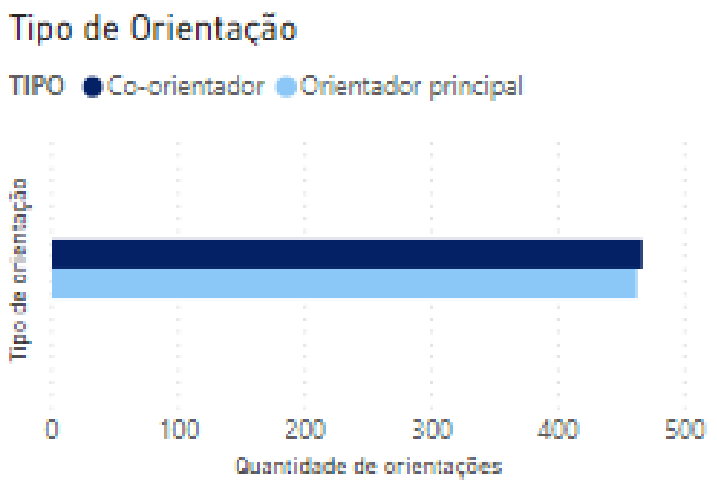

Figura 4.5: Gráfico de orientações por tipo para o PPGEE.

No programa PGEA, há uma grande diferença entre orientação principal e co-orientação. 71% das orientações são como co-orientador, enquanto apenas 29% são como orientador principal. No PPEE, 63,18% são co-orientações, enquanto 36,82% são orientações principais. Por fim, no PPMEC, 60,6% são co-orientações e 39,4% são orientações principais.

A partir disso, observa-se que o programa em que há maior homogeneidade entre orientações principais e co-orientações é o programa PPGEE, nos demais, a maioria ainda é de co-orientações.

4.2.5 Orientações por Instituição e por Curso

Na análise de orientações por instituição, foi visto que o corpo docente do programa PPGEE, faz mais orientações na Universidade de Brasília, o que é já era esperado, totalizando 95,91% das orientações. O gráfico de pizza da Figura 4.6 demonstra esse comportamento.

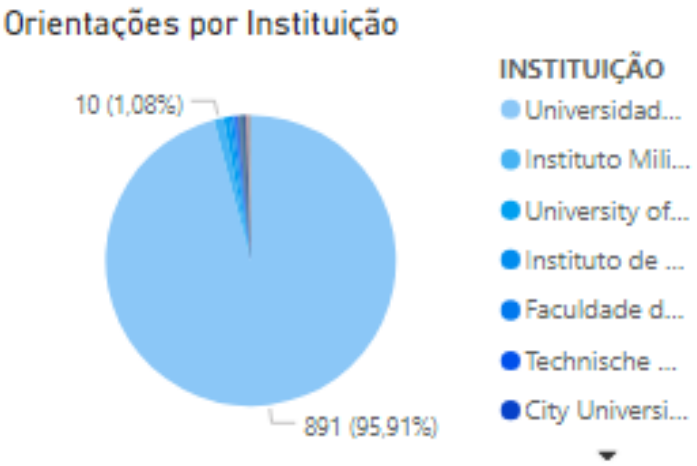

Figura 4.6: Gráfico de quantidade de orientações por instituição do PPGEE.

Além disso, é possível ver o comportamento dos cursos, prevalencendo o curso de Engenharia Elétrica, para o caso do PPGEE, seguido de Engenharia de Redes de Comunicação, como as de maior quantidade de orientação. A Figura 4.7 esboça isto.

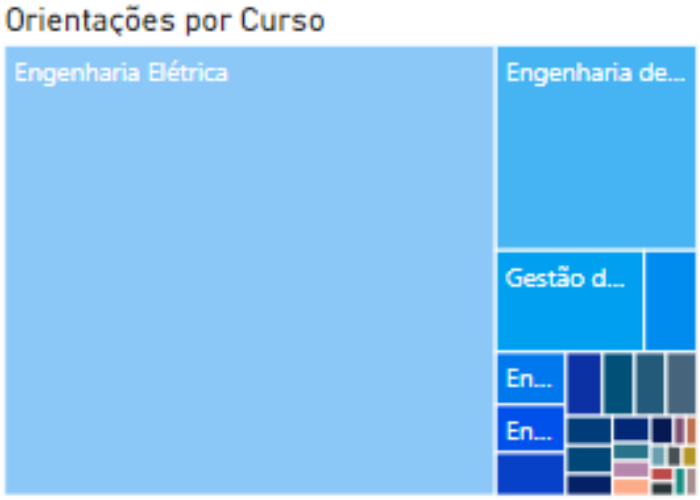

Figura 4.7: Gráfico de quantidade de orientações por curso, do programa PPGEE.

O PGEA e o PPEE apresentou comportamento semelhante ao PPGEE, no quesito de cursos, com prevalência da Universidade de Brasília como instituição, onde mais ocorrem orientações e de principal curso Engenharia Elétrica, como era esperado. Já no PPMEC, o cenário é diferente. Por se tratar de um programa em Sistemas Mecatrônicos, os cursos de maior foco em orientação são Engenharia Mecânica e Engenharia Mecatrônica.

4.3 Análise de conferências

A análise de conferências levou em consideração os idiomas em que eram publicados, a quantidade por ano de publicações, a posição de autoria e as palavras mais utilizadas.

No programa PPGEE, foram 1167 publicações em conferências, enquanto no PGEA foram 1892, no PPEE 991 e no PPMEC 1745 publicações. Esses valores surgem do total acumulado desde o ano de 1970. A Tabela 4.2 esboça a porcentagem de publicações por idioma dos programas analisados.

Tabela 4.2: Tabela com os dados da porcentagem de publicações por idioma em conferências para os programas analisados.

Programa	Idioma	Porcentagem
PPGEE	Português	56.04%
	Inglês	42.5%
	Francês	0.69%
	Bretão	0.43%
	Alemão	0.09%
	Espanhol	0.09%
	Iidiche	0.09%
	Japonês	0.09%
PGEA	Português	51.43%
	Inglês	47.94%
	Francês	0.21%
	Italiano	0.11%
	Alemão	0.05%
	Espanhol	0.16%
	Não informado	0.05%
	Outros	0.05%
PPEE	Português	50.35%
	Inglês	47.23%
	Francês	1.21%
	Bretão	0.20%
	Alemão	0.10%
	Espanhol	0.61%
	Iidiche	0.10%
	Africâner	0.10%
PPMEC	Português	50.20%
	Inglês	47.45%
	Francês	0.34%
	Bretão	0.17%
	Alemão	0.06%
	Espanhol	1.72%
	Não informado	0.06%

Além disso, foi verificada a posição em que o pesquisador ocupa na listagem de autoria. Por exemplo, se numa publicação, ele atua como primeiro, segundo, terceiro autor, e assim sucessivamente. Isso é uma análise importante, pois um indicador positivo para o programa é a rede de colaboração gerada entre pesquisadores que publicam juntos e com seus alunos. Isso demonstra que a pesquisa está sendo difundida, e os alunos estão tendo acesso a um ensino de ponta, o que eleva o conceito do programa de pós-graduação na avaliação da CAPES. Também, há informações da quantidade de publicações por ano, o que auxilia ver a frequência que cada pesquisador faz publicações.

Outro ponto de análise são as palavras mais utilizadas em artigos aceitos em conferência. Isso visa identificar os temas mais relevantes para a métrica Conferências. Assim, no programa PPGEE, os cinco nomes mais frequentes foram *systems*, *analysis*, *networks*, tensão e linhas. No programa PGEA, os mais frequentes foram *control*, *systems*, sinais, *video* e *networks*.

No PPEE, as palavras com mais relevância foram: Redes, *networks*, *estimation*, *software* e *design*. Por fim, no programa PPMEC foram: Desenvolvimento, *Robot*, *System*, Estudo e *Control*. Essa análise condiz com os perfis de atuação do curso, visto que no PPGEE trabalha-se com sistemas de potência e telecomunicações; no PGEA, com sistemas eletrônicos e automação; no PPEE, com a área de segurança da informação, computação e telecomunicações; e no PPMEC o foco são os sistemas mecatrônicos, que envolvem controle, automação, gestão da produção, entre outros.

4.4 Análise sobre publicações em Periódicos

Neste estudo foram levantados os dados de idioma, para comparar se há uma discrepância entre os trabalhos aceitos em conferências e periódicos, verificar também a proporção de publicação por ano, o perfil de posição de autoria dos pesquisadores, as publicações por estrato Qualis e as palavras com maiores ocorrências nas publicações em periódicos.

O programa PPGEE possui um total de 354 publicações em periódicos, o PGEA um total de 506 publicações, o PPEE 415 publicações e o PPMEC 584, registradas até o mês de setembro de 2019. Existe uma diferença grande entre as publicações em periódicos e conferências, no quesito de idioma. A Tabela 4.3 traz a informação da porcentagem de publicações por idioma e programa. A partir dessa tabela, é possível identificar os pontos de discrepância entre as publicações de periódicos e conferências. Em geral, existem poucos idiomas que os pesquisadores publicam nos periódicos, sendo que em conferências, há uma maior diversidade de idiomas listados. Uma hipótese para este fato é que os *journals* possuem como idioma, geralmente, o inglês. Assim, os pesquisadores são obrigados a se adequarem às condições impostas para publicarem seus trabalhos.

Tabela 4.3: Tabela com os dados da porcentagem de publicações por idioma em periódicos para os programas analisados.

Programa	Idioma	Porcentagem
PPGEE	Português	35.59%
	Inglês	64.12%
	Espanhol	0.28%
PGEA	Português	21.74%
	Inglês	77.87%
	Espanhol	0.20%
	Francês	0.20%
PPEE	Português	20.24%
	Inglês	79.28%
	Espanhol	0.48%
PPMEC	Português	24.49%
	Inglês	71.58%
	Espanhol	2.23%
	Chinês	1.37%
	Holandês	0.34%

Outro fato observado é de que as publicações de maior estrato são, em sua maioria, no idioma inglês, o que torna incentiva os pesquisadores a preferirem a língua. Além disso, foi constatado que no programa PPGEE, houve um pico de publicações em 2019, bem como nos programas PGEA e PPEE. Porém, no programa PPMEC, houve uma queda de publicações neste ano.

Para todos os progrmas, foram geradas *word clouds* para verificar quais os termos geram mais publicações em cada estrato. Por exemplo, no programa PPGEE os assuntos que mais geraram publicações em estrato A1 foram *power*, *systems* e *formula*. No PGEA os assuntos mais abordados em publicações A1 foram *systems*, *robust* e *compression*. No PPEE foram *formula*, *systems*, *math* e *cryptography*. Já no PPMEC, os assuntos foram *analysis*, *system* e *control*. Na Figura 4.8 pode ser visto um exemplo de *Word Cloud* utilizado nas análises.

Figura 4.8: Wordcloud gerada para periódicos do programa PPMEC.

4.5 Análise sobre citações

Essa análise tem o objetivo de verificar as publicações, nas quais os pesquisadores têm mais citações, bem como auxiliá-lo a conseguir citações em artigos específicos para aumentar seu H-index. Para verificar em qual artigo o pesquisador deveria receber citações, foi utilizado o algoritmo Naive Bayes, tomando como base a série histórica dos dados de citações por publicação dos meses de junho, julho, agosto e setembro. O Naive Bayes, por ser um algoritmo probabilístico, indica a probabilidade de receber uma citação.

Primeiramente, os dados dos quatro meses são carregados e gerada uma lista com as classes (recebeu ou não recebeu citações) e suas porcentagens de ocorrência. O algoritmo desenvolvido apresenta uma ideia simples e didática, que pode ser vista por meio do exemplo a seguir. De 100 publicações analisadas, considere 20% receberam citações neste intervalo de tempo. Dessas publicações que receberam citações, 90% possuíam quantidade de citações próximas em até 3 unidades ao valor do H-index e 15% das que não tinham quantidade de citações próximas em até 3 unidades do valor do H-index, também receberam citações. Logo, qual a probabilidade de que numa próxima carga de dados, eu receba citação num artigo de interesse (valor próximo ao H-index)?

O algoritmo gerou uma precisão de 62%, de acordo com o cálculo efetuado pelas ferramentas da biblioteca utilizada *scikit-learn*. A partir disso, foi gerado outro algoritmo para sugerir os artigos nos quais, ao ser citado, poderia exercer aumento no valor do H-index.

Ao utilizar um algoritmo de classificação, como no caso, foi possível obter uma perspectiva de que se um artigo está sendo citado, num curto intervalo de tempo, o mesmo

tem condições de ser citado em um curto período. Atualmente, o H-index considerando os dados a partir de 2014, auxilia a compreender se há uma tendência de recebimento de citação em artigos mais atuais, podendo ser incrementado pelo uso de *Machine Learning* como no caso, enriquecendo ainda mais a análise. Isso pode ajudar o coordenador do programa a traçar metas e incentivar o aumento da produtividade de seus pesquisadores.

De modo geral, o painel criado para as análises de citação possui uma visão parecida com a presente no Google Scholar, porém consolidada para todos os pesquisadores cadastrados, em um só lugar. Isso permite fazer comparações entre eles, sem qualquer esforço extra. Este relatório conta com os títulos das publicações e suas respectivas citações, valor do H-index geral e H-index 2014, informações sobre os artigos que receberam mais citações em cada ano, bem como as palavras presentes nos títulos que geraram mais citações.

O pesquisador que possui o maior H-index geral entre os analisados, no valor de 38, é o professor Ricardo Lopes de Queiroz, que compõe o programa PPMEC. As palavras presentes nos títulos que mais receberam citações são: *Image*, *Compression*, *Coding* e *Estimation*. Logo, percebe-se que seu nicho de atuação é a área de processamento de sinais e imagens. O ano de sua publicação que mais recebeu citações é 2003 e conta com 376 citações.

Apesar de ser o maior H-index geral, o pesquisador não possui o maior H-index 2014, o que pode ser percebido diante da diminuição da quantidade de citações recebidas nos anos subsequentes a 2003. Assim, o professor que possui o maior H-index 2014 é o Demétrio Antônio da Silva Filho, cujo valor é 24. Isso indica que possui um alto índice de citações em períodos atuais e continua publicando trabalhos relevantes. Este pesquisador também possui o artigo mais citado, entre os pesquisadores analisados, cujo título é *"Charge Transport in Organic Semiconductors"*.

Este relatório permite que o programa se posicione acerca da colaboração mútua, para que um pesquisador cite o outro em sua área. As *Word Clouds* servem para pesquisar termos afins e identificar um docente que compartilhe da mesma linha de pesquisa, para que possam gerar uma rede de citações.

Capítulo 5

Conclusão

Neste trabalho foi desenvolvida uma plataforma capaz de apoiar a tomada de decisão a partir da análise de dados bibliométricos extraídos do Google Scholar e Currículo Lattes. Foram utilizadas técnicas de Machine Learning e Mineração de Dados para realização dos estudos, que contou com análise dos perfis de orientação, publicações em conferências e periódicos dos pesquisadores relacionados, bem como aplicou-se o algoritmo Naive Bayes para determinar a probabilidade de um artigo determinado receber uma citação e aumentar o H-index do pesquisador.

O objetivo do trabalho foi explorar os dados educacionais e validar os dados por meio de uma plataforma online, que permite consolidar todos os membros dos programas analisados, em uma só página, cruzando dados e proporcionando uma informação completa e acessível para todos os usuários. A página utiliza conceitos de Business Intelligence e conta com relatórios criados para atender a demanda por respostas acerca do perfil dos pesquisadores que fazem parte do programa. Dessa forma, a plataforma cumpriu seu papel de classificar tanto o pesquisador quanto o programa no qual este faz parte.

Foi verificado que, quanto melhor a fase de pré-processamento dos dados, melhor será o resultado final. Durante o trabalho foram feitos vários ajustes nessa etapa, que demandou bastante tempo, para que o painel de visualização consumisse os dados da melhor maneira possível.

A plataforma, como solução proposta, conseguiu responder, de forma eficaz, as perguntas levantadas sobre os programas, bem como associar os indicadores aos meios de avaliação da CAPES, para que a ferramenta apresentasse eficiência ao realizar as consultas necessárias. Ao comparar os resultados obtidos pela análise de dados com os dados extraídos das próprias plataformas, foi possível constatar que a plataforma apresentou bons resultados. Além disso, o trabalho realizado pode ser aplicado em outras áreas da educação, pois tem o intuito de analisar dados bibliométricos em geral, o que torna a ferramenta versátil.

As maiores dificuldades neste trabalho surgiram na parte de extração de dados, pré-processamento e tratamento, visto que, por se tratar de dados obtidos por meio de web crawler, várias inconsistências estiveram presentes, principalmente na despadronização dos dados e alta quantidade de variáveis a serem analisadas. A fase de pré-processamento bem realizada serviu para melhorar a precisão do algoritmo Naive Bayes, que na primeira utilização apresentou precisão de 49% e após ajustes nesta fase, proporcionou um aumento de 13% em seu valor, atingindo uma precisão de 62%.

A precisão do algoritmo Naive Bayes foi considerada razoável, visto que seu valor foi de 62%. Possivelmente, esse valor relativamente baixo, se deve ao fato de que as coletas de dados para geração da série histórica para aplicação do algoritmo e estudo de citações foi realizada em apenas quatro meses, onde poucos artigos tiveram a quantidade de citações modificadas. Logo, acredita-se que se o projeto contasse com uma coleta de dados num período de um ano, por exemplo, esta análise poderia apresentar resultados superiores.

Outro ponto que traria melhora à análise de citações e, consequentemente, aos resultados obtidos, seria o incentivo de que os professores dos programas aderissem ao uso do Google Scholar, criando seus perfis e fazendo o controle das citações associadas a eles, visto que existem várias publicações vinculadas a um pesquisador que não pertencem a ele e também faltam publicações cuja autoria realmente é daquele membro.

De contribuições ao meio científico, entende-se que o desenvolvimento dessa plataforma proporcione uma melhoria na qualidade dos programas de pós-graduação, pois apresenta análises que facilita o intercâmbio de informações relevantes, entre áreas de pesquisa e pesquisadores diferentes, possibilitando a criação de uma rede de colaboração para que as pesquisas de um programa sejam mais difundidas e assim contribuam positivamente para trazer inovações à comunidade, bem como consequente aumento do conceito do programa perante a CAPES.

5.1 Trabalhos Futuros

Para trabalhos futuros, uma proposta seria a utilização de outros algoritmos de classificação para verificar se o comportamento seria melhor ou pior em relação ao Naive Bayes, e também, expandir a análise dos dados para todos os programas de pós-graduação da Universidade de Brasília.

Além disso, poderia otimizar a criação das matrizes utilizadas nas *Bag-of-words* para melhoria das Word Clouds, que geram diminuição do tempo de processamento e melhoria da qualidade de análise dos assuntos presentes nas publicações.

Outra sugestão para trabalhos futuros seria a implementação de um processo mais robustos, utilizando IA, para determinar quais as publicações deveriam receber citações

para aumentar o H-index do pesquisador, bem como indicar professores com áreas de pesquisa parecidas, para que estes possam publicar em parceria ou mesmo compartilhar citações.

Referências

[1] Imasato, Takeyoshi, Marcelo Scherer Perlin e Denis Borenstein: *Análise do perfil dos acadêmicos e de suas publicações científicas em administração*. RAC-Revista de Administração Contemporânea, 21(1):62–83, 2017. x, 8, 9

[2] Fayyad, Usama, Gregory Piatetsky-Shapiro e Padhraic Smyth: *From data mining to knowledge discovery in databases*. AI magazine, 17(3):37–37, 1996. x, 12

[3] Baia, Carlos: *Introdução ao machine learning*. `http://carlosbaia.com/2016/07/17/introducao-ao-machine-learning/`. x, 17

[4] Ceci, Flávio: *Business Intelligence*. janeiro 2012, ISBN 9788578174651. x, 34

[5] Microsoft: *Documentação do Power BI Embedded do Azure*. Disponível em: https://docs.microsoft.com/pt-br/power-bi/service-azure-and-power-bi. x, 45

[6] Sidone, Otávio José Guerci, Eduardo Amaral Haddad e Jesús Pascual Mena-Chalco: *A ciência nas regiões brasileiras: evolução da produção e das redes de colaboração científica*. Transinformação, 28(1):15–32, 2016. 1, 2

[7] Da Silva, Marcio Bezerra *et al.*: *Analysis of scientific production in information technology: A panoramic study of articles published by librarianship professors from the unb*. Biblios, (59):18, 2015. 1

[8] Reis, Luciano Gomes dos e Jaqueline Horvath: *Uma análise sobre a produção acadêmica dos docentes das universidades estaduais paranaenses de 2008 a 2012*. Revista Gestão Universitária na América Latina-GUAL, 7(3):22–42, 2014. 2

[9] Balancieri, Renato, Alessandro Botelho Bovo, Vinícius Medina Kern, RCS dos Pacheco e Ricardo Miranda Barcia: *A análise de redes de colaboração científica sob as novas tecnologias de informação e comunicação: um estudo na plataforma lattes*. Ciência da informação, 34(1):64–77, 2005. 2, 5

[10] Dias, Thiago Magela Rodrigues, Gray Farias Moita e Patrícia Mascarenhas Dias: *Adoção da plataforma lattes como fonte de dados para caracterização de redes científicas*. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, 21(47):16–26, 2016. 3

[11] Silva, Bruna Guedes Martins da: *Estudo panorâmico da publicação científica em tecnologia da informação pelos professores de biblioteconomia da unb*. 2013. 3

[12] Santos Estácio, Letícia Silvana dos: *A importância do currículo lattes como ferramenta que representa a ciência, tecnologia e inovação no país.* Revista ACB: Biblioteconomia em Santa Catarina, 22(2):300–311, 2017. 3

[13] Digiampietri, Luciano Antonio, Rogério Mugnaini, Jesús Pascual Mena Chalco, Karina Valdivia Delgado e José de Jesús Pérez Alcázar: *Análise macro das últimas atualizações dos currículos lattes.* Em Questão, 20(3):88–113, 2014. 3

[14] Ferraz, Renato Ribeiro Nogueira, Luc Quoniam, Denise Nacif Pimenta, Jesús Pascual Mena-Chalco e Carolina Alencar Nigro: *Extração e disponibilização on line de indicadores de desempenho e prospecção dos resultados das pesquisas em dengue com a utilização da ferramenta computacional scriptlattes.* Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, 20(43):93–114, 2015. 3

[15] BRITO, Aline Grasiele Cardoso de, Luc Quoniam e Jesús Pascual Mena-Chalco: *Exploração da plataforma lattes por assunto: proposta de metodologia.* Transinformação, 28(1):77–86, 2016. 3

[16] Amaral, Roniberto Morato, Aline Grasiele Cardoso Brito, Luc Marie Quoniam, Leandro Innocentini Lopes de Faria *et al.*: *Panorama da inteligência competitiva no brasil: os pesquisadores e a produção científica na plataforma lattes.* Perspectivas em Ciência da Informação, 21(4):97–120, 2016. 3

[17] Mascarenhas, Fábio, Natanael Vitor Sobral, Guilherme Alves Santana, Tatyane Lucia Cruz *et al.*: *Mapeamento da produção científica brasileira sobre acesso aberto: 2001 a 2011.* Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, 17(2):19–35, 2012. 4

[18] Perucchi, Valmira e Suzana Pinheiro Machado Mueller: *Produção dos professores dos institutos federais de educação, ciência e tecnologia no currículo da plataforma lattes.* Informação & Informação, 22(1):111–128, 2017. 4

[19] Aguillo, Isidro F: *Is google scholar useful for bibliometrics? a webometric analysis.* Scientometrics, 91(2):343–351, 2012. 5

[20] Haddaway, Neal Robert, Alexandra Mary Collins, Deborah Coughlin e Stuart Kirk: *The role of google scholar in evidence reviews and its applicability to grey literature searching.* PloS one, 10(9):e0138237, 2015. 6

[21] Bergman, Elaine M Lasda: *Finding citations to social work literature: The relative benefits of using web of science, scopus, or google scholar.* The journal of academic librarianship, 38(6):370–379, 2012. 6

[22] Capes, Portal Periódicos: *Disponível em:< http://www. periodicos. capes. gov. br/>.* Acesso em 10 de setembro, 2019. 7

[23] Beuren, Ilse Maria e José Carlos De Souza: *Em busca de um delineamento de proposta para classificação dos periódicos internacionais de contabilidade para o qualis capes.* Revista Contabilidade & Finanças, 19(46):44–58, 2008. 7

[24] Herculano, Rondinelli Donizetti e Ana Maria Q Norberto: *Análise da produtividade científica dos docentes da universidade estadual paulista, campus de marília/sp*. Perspectivas em Ciência da Informação, páginas 57–70, 2012. 9

[25] Ferraz, Renato Ribeiro Nogueira, Carolina Alencar Nigro e Luc Quoniam: *Apoio da ferramenta computacional scriptsucupira para prestação de contas à capes em relação ao quadriênio 2013-2016 por um programa de pós-graduação stricto sensu em direito*. Prisma. com, (35):51–72, 2018. 9

[26] Ferraz, Renato Ribeiro Nogueira, Luc Marie Quoniam, Emerson Antônio Maccari e Vladmir Oliveira da Silveira: *Análise e gestão de análise de redes de colaboração entre pesquisadores de programas de pós-graduação stricto sensu com a utilização da ferramenta computacional scriptlattes*. Perspectivas em Gestão & Conhecimento, 4:133–147, 2014. 9

[27] Cardoso, Olinda Nogueira Paes e Rosa Teresa Moreira Machado: *Gestão do conhecimento usando data mining: estudo de caso na universidade federal de lavras*. Revista de Administração Pública-RAP, 42(3):495–528, 2008. 10, 11

[28] Wu, Xindong, Xingquan Zhu, Gong Qing Wu e Wei Ding: *Data mining with big data*. IEEE transactions on knowledge and data engineering, 26(1):97–107, 2013. 11, 12

[29] Larose, Daniel T e Chantal D Larose: *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014. 13

[30] Aranha, Christian e Emmanuel Passos: *A tecnologia de mineração de textos*. Revista Eletrônica de Sistemas de Informação, 5(2), 2006. 18, 19

[31] Wallach, Hanna M: *Topic modeling: beyond bag-of-words*. Em *Proceedings of the 23rd international conference on Machine learning*, páginas 977–984. ACM, 2006. 21

[32] Both, Eder Luis e Sérgio Luis Dill: *Business intelligence aplicado em saúde pública*. Anais SULCOMP, 1, 2012. 22

[33] Costa, Sérgio e Maribel Santos: *Sistema de business intelligence no suporte à gestão estratégica*. Em *Atas da Conferência da Associação Portuguesa de Sistemas de Informação*, volume 12, páginas 162–174, 2014. 22

[34] Ribeiro, Marcela Xavier *et al.*: *Mineração de dados em múltiplas tabelas fato de um data warehouse.* 2004. 34

[35] Han, Jiawei, Jian Pei e Yiwen Yin: *Mining frequent patterns without candidate generation*. Em *ACM sigmod record*, volume 29, páginas 1–12. ACM, 2000. 34

[36] Bornmann, Lutz e Hans Dieter Daniel: *What do we know about the h index?* Journal of the American Society for Information Science and technology, 58(9):1381–1385, 2007. 36

[37] Agranonik, Marilyn e Vânia Naomi Hirakata: *Cálculo de tamanho de amostra: proporções*. Revista HCPA. Porto Alegre. Vol. 31, n. 3,(2011), p. 382-388, 2011. 36

[38] Jin, Yaochu e Bernhard Sendhoff: *Pareto-based multiobjective machine learning: An overview and case studies*. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(3):397–415, 2008. 36

[39] Nausheen, Farha e Sayyada Hajera Begum: *Sentiment analysis to predict election results using python*. Em *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, páginas 1259–1262. IEEE, 2018. 41

[40] Praciano, Bruno Justino Garcia, João Paulo Carvalho Lustosa da Costa, João Paulo Abreu Maranhão, Fábio Lúcio Lopes de Mendonça, Rafael Timoteo de Sousa Júnior e Juliano Barbosa Prettz: *Spatio-temporal trend analysis of the brazilian elections based on twitter data*. Em *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, páginas 1355–1360. IEEE, 2018. 43

[41] Moreira Tanuro, Carla: *Uma arquitetura de software para descoberta de regras de associação multidimensional, multinível e de outliers em cubos olap: um estudo de caso com os algoritmos apriori e fpgrowth*. Tese de Mestrado, Universidade Federal de Pernambuco, 2010. 44

[42] Lachev, Teo e Edward Price: *Applied Microsoft Power BI: Bring your data to life!* Prologika Press, 2018. 45

[43] CORREIA, Maria Sonia Barros Barbosa: *Probabilidade e estatística*. 2003. 45

Apêndice A

A.1 Função para extração dos dados de orientações do XML

```
1
2  def getorient(zipname):
3      nomezip = './xml_zip' + '/' + str(zipname)
4      archive = zipfile.ZipFile(nomezip, 'r')
5      dadoslattes = archive.open('curriculo.xml')
6      soup = BeautifulSoup(dadoslattes, 'lxml',
7                           from_encoding='ISO-8859-1')
8      verify = soup.find_all('outra-producao')
9      if len(verify) == 0:
10         print('Outras producoes nao encontradas para', zipname)
11     else:
12         orientacoes_concluidas = verify[0].find_all('orientacoes-concluidas
       ')
13         if len(orientacoes_concluidas) == 0:
14             print('Orientacoes nao encontradas para', zipname)
15         else:
16             ano = []
17             natureza = []
18             instituicao = []
19             curso = []
20             nome_aluno = []
21             tipo_orientacao = []
22             financiamento = []
23             mestrado_orien = orientacoes_concluidas[0].find_all(
24                 'orientacoes-concluidas-para-mestrado')
25             if len(mestrado_orien) == 0:
26                 print('Orientacoes concluidas de mestrado nao encontradas
       para', zipname)
```

```
27            else:
28                for i in range(len(mestrado_orien)):
29                    base = mestrado_orien[i].find_all(
30                        'dados-basicos-de-orientacoes-concluidas-para-
    mestrado')
31                    base = str(base)
32                    resultado = re.search('ano=\"(.*)\" doi',
33                                          base)
34                    if resultado is None:
35                        dados1 = 'VAZIO'
36                    else:
37                        dados1 = resultado.group(1)
38                    ano.append(dados1)
39                    resultado = re.search('natureza=\"(.*)\" pais',
40                                          base)
41                    if resultado is None:
42                        dados1 = 'VAZIO'
43                    else:
44                        dados1 = resultado.group(1)
45                    natureza.append(dados1)
46                    detalhamento = mestrado_orien[i].find_all(
47                        'detalhamento-de-orientacoes-concluidas-para-
    mestrado')
48                    detalhamento = str(detalhamento)
49                    resultado = re.search('nome-da-instituicao=\"(.*)\"
    nome-do-curso=',
50                                          detalhamento)
51                    if resultado is None:
52                        dados1 = 'VAZIO'
53                    else:
54                        dados1 = resultado.group(1)
55                    instituicao.append(dados1)
56                    resultado = re.search('nome-do-curso=\"(.*)\" nome-do-
    curso-ingles',
57                                          detalhamento)
58                    if resultado is None:
59                        dados1 = 'VAZIO'
60                    else:
61                        dados1 = resultado.group(1)
62                    curso.append(dados1)
63                    resultado = re.search('nome-do-orientado=\"(.*)\" nome-
    orgao',
64                                          detalhamento)
65                    if resultado is None:
66                        dados1 = 'VAZIO'
```

```
67                      else:
68                          dados1 = resultado.group(1)
69                      nome_aluno.append(dados1)
70                      resultado = re.search('tipo-de-orientacao=\"(.*)\">',
71                                            detalhamento)
72                      if resultado is None:
73                          dados1 = 'VAZIO'
74                      else:
75                          dados1 = resultado.group(1)
76                      tipo_orientacao.append(dados1)
77                      resultado = re.search('flag-bolsa=\"(.*)\" nome-da-
    agencia',
78                                            detalhamento)
79                      if resultado is None:
80                          dados1 = 'VAZIO'
81                      else:
82                          dados1 = resultado.group(1)
83                      financiamento.append(dados1)
84              orienconc_dout = orientacoes_concluidas[0].find_all(
85                  'orientacoes-concluidas-para-doutorado')
86              if len(orienconc_dout) == 0:
87                  print(
88                      'Orientacoes concluidas de doutorado nao encontradas
    para', zipname)
89              else:
90                  for i in range(len(orienconc_dout)):
91                      base = orienconc_dout[i].find_all(
92                          'dados-basicos-de-orientacoes-concluidas-para-
    doutorado')
93                      base = str(base)
94                      resultado = re.search('ano=\"(.*)\" doi',
95                                            base)
96                      if resultado is None:
97                          dados1 = 'VAZIO'
98                      else:
99                          dados1 = resultado.group(1)
100                     ano.append(dados1)
101                     resultado = re.search('natureza=\"(.*)\" pais',
102                                           base)
103                     if resultado is None:
104                         dados1 = 'VAZIO'
105                     else:
106                         dados1 = resultado.group(1)
107                     natureza.append(dados1)
108                     detalhamento = orienconc_dout[i].find_all(
```

```
109                         'detalhamento-de-orientacoes-concluidas-para-
    doutorado')
110                     detalhamento = str(detalhamento)
111                     resultado = re.search('nome-da-instituicao=\"(.*)\"
    nome-do-curso=',
112                                           detalhamento)
113                     if resultado is None:
114                         dados1 = 'VAZIO'
115                     else:
116                         dados1 = resultado.group(1)
117                     instituicao.append(dados1)
118                     resultado = re.search('nome-do-curso=\"(.*)\" nome-do-
    curso-ingles',
119                                           detalhamento)
120                     if resultado is None:
121                         dados1 = 'VAZIO'
122                     else:
123                         dados1 = resultado.group(1)
124                     curso.append(dados1)
125                     resultado = re.search('nome-do-orientado=\"(.*)\" nome-
    orgao',
126                                           detalhamento)
127                     if resultado is None:
128                         dados1 = 'VAZIO'
129                     else:
130                         dados1 = resultado.group(1)
131                     nome_aluno.append(dados1)
132                     print(dados1)
133                     resultado = re.search('tipo-de-orientacao=\"(.*)\">',
134                                           detalhamento)
135                     if resultado is None:
136                         dados1 = 'VAZIO'
137                     else:
138                         dados1 = resultado.group(1)
139                     tipo_orientacao.append(dados1)
140                     resultado = re.search('flag-bolsa=\"(.*)\" nome-da-
    agencia',
141                                           detalhamento)
142                     if resultado is None:
143                         dados1 = 'VAZIO'
144                     else:
145                         dados1 = resultado.group(1)
146                     financiamento.append(dados1)
147
```

```
148         outrasorientacoes = verify[0].find_all('outras-orientacoes-
    concluidas')
149         if len(outrasorientacoes) == 0:
150             print('Outras orientacoes nao encontradas para', zipname)
151         else:
152             for j in range(len(outrasorientacoes)):
153                 base = outrasorientacoes[j].find_all(
154                     'dados-basicos-de-outras-orientacoes-concluidas')
155                 base = str(base)
156                 resultado = re.search('ano=\"(.*)\" doi',
157                                       base)
158                 if resultado is None:
159                     dados1 = 'VAZIO'
160                 else:
161                     dados1 = resultado.group(1)
162                 ano.append(dados1)
163                 resultado = re.search('natureza=\"(.*)\" pais',
164                                       base)
165                 if resultado is None:
166                     dados1 = 'VAZIO'
167                 else:
168                     dados1 = resultado.group(1)
169                 natureza.append(dados1)
170                 detalhamento = outrasorientacoes[j].find_all(
171                     'detalhamento-de-outras-orientacoes-concluidas')
172                 detalhamento = str(detalhamento)
173                 resultado = re.search('nome-da-instituicao=\"(.*)\" nome-do
    -curso=',
174                                       detalhamento)
175                 if resultado is None:
176                     dados1 = 'VAZIO'
177                 else:
178                     dados1 = resultado.group(1)
179                 instituicao.append(dados1)
180                 resultado = re.search('nome-do-curso=\"(.*)\" nome-do-curso
    -ingles',
181                                       detalhamento)
182                 if resultado is None:
183                     dados1 = 'VAZIO'
184                 else:
185                     dados1 = resultado.group(1)
186                 curso.append(dados1)
187                 resultado = re.search('nome-do-orientado=\"(.*)\" numero-de
    -paginas',
188                                       detalhamento)
```

```
                if resultado is None:
                    dados1 = 'VAZIO'
                else:
                    dados1 = resultado.group(1)
                nome_aluno.append(dados1)
                resultado = re.search('tipo-de-orientacao-concluida=\"(.*)
    \">',
                                      detalhamento)
                if resultado is None:
                    dados1 = 'VAZIO'
                else:
                    dados1 = resultado.group(1)
                tipo_orientacao.append(dados1)
                resultado = re.search('flag-bolsa=\"(.*)\" nome-da-agencia'
    ,
                                      detalhamento)
                if resultado is None:
                    dados1 = 'VAZIO'
                else:
                    dados1 = resultado.group(1)
                financiamento.append(dados1)
            df_orientacoes = pd.DataFrame({'ANO': ano,
                                           'NATUREZA': natureza,
                                           'INSTITUICAO': instituicao,
                                           'CURSO': curso,
                                           'ALUNO': nome_aluno,
                                           'TIPO': tipo_orientacao,
                                           'FINANCIADOR': financiamento})
            id_lattes = zipname.split('.')[0]
            pathfilename = str('./csv_producao/' +
                               id_lattes + '_orientacoes'  '.csv')
            df_orientacoes.to_csv(pathfilename, index=False)
            print(pathfilename, ' gravado com', len(
                df_orientacoes['YEAR']), ' çõorientaes')
```

A.2 Código para extração dos dados de conferências do XML

```
def get_eventos(zipname):
    nomezip = './xml_zip' + '/' + str(zipname)
    archive = zipfile.ZipFile(nomezip, 'r')
```

```
    dadoslattes = archive.open('curriculo.xml')
    soup = BeautifulSoup(dadoslattes, 'lxml',
                         from_encoding='ISO-8859-1')
    curriculum = soup.find_all('curriculo-vitae')
    if len(curriculum) == 0:
        print('curriculo vitae nao encontrado para', zipname)
    else:
        for i in range(len(curriculum)):
            dadosgerais = curriculum[i].find_all('dados-gerais')
            if len(dadosgerais) == 0:
                print('Dados gerais nao encontrados para', zipname)
            else:
                for j in range(len(dadosgerais)):
                    nomeabrev = str(dadosgerais[j])
                    resultado = re.search('nome-completo=\"(.*)\" nome-em-
citacoes',
                                          nomeabrev)
                    if resultado is None:
                        dados1 = 'VAZIO'
                    else:
                        dados1 = resultado.group(1)
                    nomecompleto = dados1
    producaobibliografica = soup.find_all('producao-bibliografica')
    if len(producaobibliografica) == 0:
        print('Producoes bibliograficas nao encontradas para', zipname)
    else:
        trabevent = producaobibliografica[0].find_all('trabalhos-em-eventos
')
        if len(trabevent) == 0:
            print('Artigos publicados nao encontrados para', zipname)
        else:
            tituloconf = []
            anoconf = []
            doiconf = []
            idiomaconf = []
            eventconf = []
            issnconf = []
            autoresconf = []
            autores_na_ordem = []
            qual_ordem = []
            publication = trabevent[0].find_all('trabalho-em-eventos')

            for i in range(len(publication)):
                dadosconf = publication[i].find_all('dados-basicos-do-
trabalho')
```

```
46                  dadospaperconf = str(dadosconf)
47                  resultado = re.search('titulo-do-trabalho=\"(.*)\" titulo-
    do-trabalho-i',
48                                        dadospaperconf)
49                  if resultado is None:
50                      dados1 = 'VAZIO'
51                  else:
52                      dados1 = resultado.group(1)
53                  tituloconf.append(dados1)
54
55                  resultado = re.search('ano-do-trabalho=\"(.*)\" doi',
56                                        dadospaperconf)
57                  if resultado is None:
58                      dados1 = 'VAZIO'
59                  else:
60                      dados1 = resultado.group(1)
61                  anoconf.append(dados1)
62
63                  resultado = re.search('doi=\"(.*)\" flag-divulgacao-c',
64                                        dadospaperconf)
65                  if resultado is None:
66                      dados1 = 'VAZIO'
67                  else:
68                      dados1 = resultado.group(1)
69                  doiconf.append(dados1)
70
71                  resultado = re.search('idioma=\"(.*)\" meio-de-divulgacao='
    ,
72                                        dadospaperconf)
73                  if resultado is None:
74                      dados1 = 'VAZIO'
75                  else:
76                      dados1 = resultado.group(1)
77                  idiomaconf.append(dados1)
78                  resultado = re.search('nome-do-evento=\"(.*)\" volume',
79                                        paperdt)
80                  if resultado is None:
81                      dados1 = 'VAZIO'
82                  else:
83                      dados1 = resultado.group(1)
84                  eventconf.append(dados1)
85                  resultado = re.search('issn=\"(.*)\" local-de-public',
86                                        paperdt)
87                  if resultado is None:
88                      dados1 = 'VAZIO'
```

```
                else:
                    dados1 = resultado.group(1)
                    dados1 = str(dados1[0:4]) + '-' + str(dados1[4:])
                issnconf.append(dados1)
                author = publication[i].find_all('autores')
                autores_total = []
                autores_total1 = []
                autores_ordena = []
                for j in range(len(author)):
                    auth = str(author[j])
                    resultado = re.search(
                        'nome-completo-do-autor=\"(.*)\" nome-para-citacao'
,
                        auth)
                    if resultado is None:
                        dados1 = 'VAZIO'
                    else:
                        dados1 = resultado.group(1)
                        complete = resultado.group(1)
                    autores_total.append(dados1)

                    resultado = re.search(
                        'ordem-de-autoria=\"(.*)\"',
                        auth)
                    if resultado is None:
                        dados1 = 'VAZIO'
                    else:
                        dados1 = resultado.group(1)
                        ncao = resultado.group(1)
                    autores_total1.append(dados1)
                    if nomecompleto == complete:
                        autores_ordena.append(ncao)

                autoresconf.append(autores_total)
                autores_na_ordem.append(autores_total1)
                qual_ordem.append(autores_ordena)

            df_papers = pd.DataFrame({'TITLE': tituloconf,
                                      'YEAR': anoconf,
                                      'DOI': doiconf,
                                      'LANG': idiomaconf,
                                      'EVENT': eventconf,
                                      'ISSN': issnconf,
                                      'AUTHOR': autoresconf,
                                      'ORDER': autores_na_ordem,
```

```
                                        'ORDER_OK': qual_ordem})
            id_lattes = zipname.split('.')[0]
            id_parse = str('./saida_csv/' + id_lattes + '_conferences'  '.
    csv')
            df_papers.to_csv(id_parse, index=False)
            print(id_parse, ' gravado com', len(
                df_papers['YEAR']), 'çõpublicaes em conferencias')
```

A.3 Relatórios completos por análise e programa

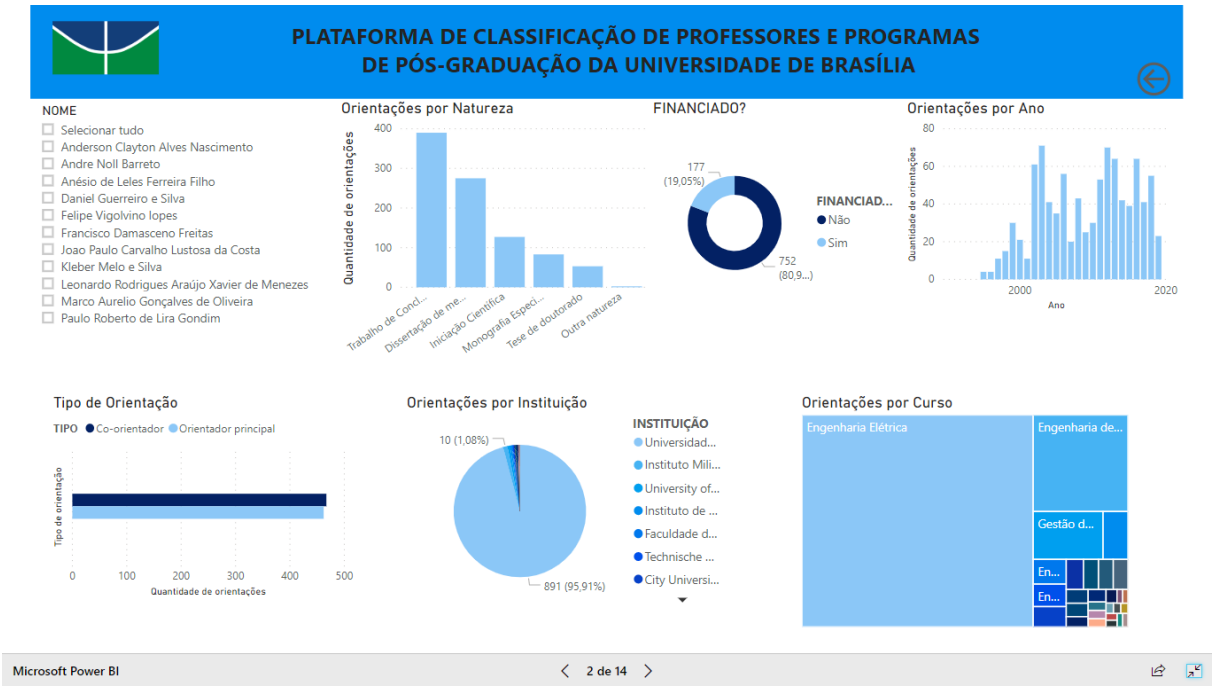

Figura A.1: Relatório total de orientações para o programa PPGEE

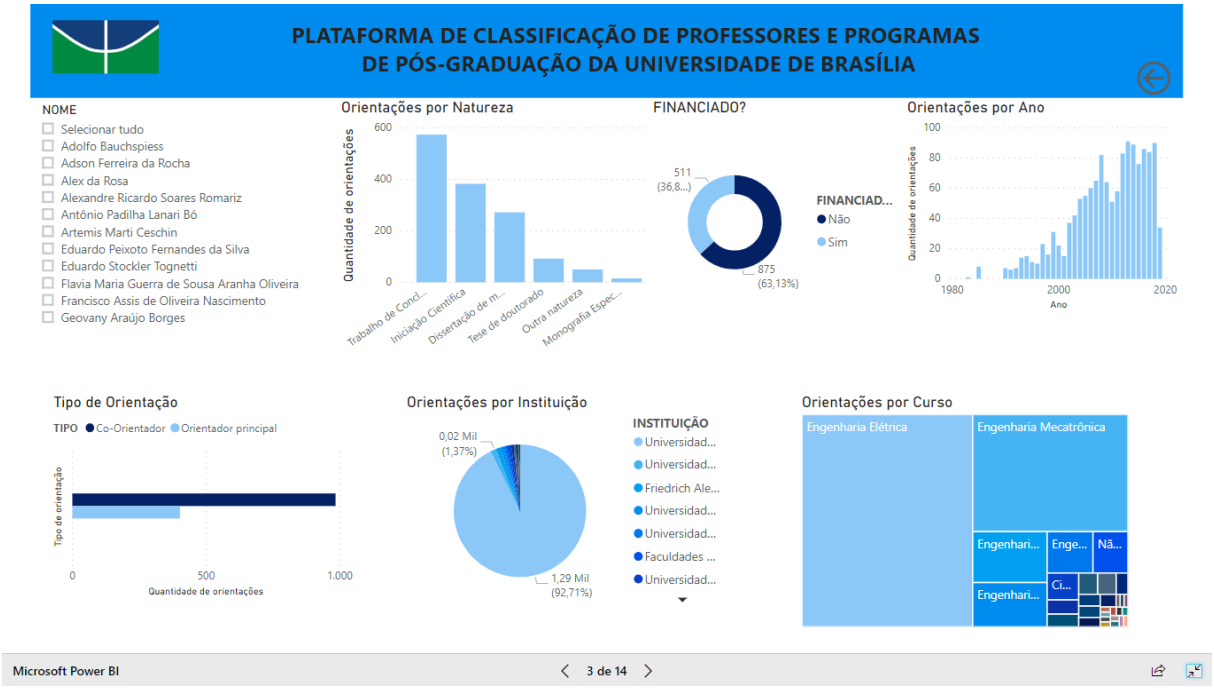

Figura A.2: Relatório total de orientações para o programa PGEA

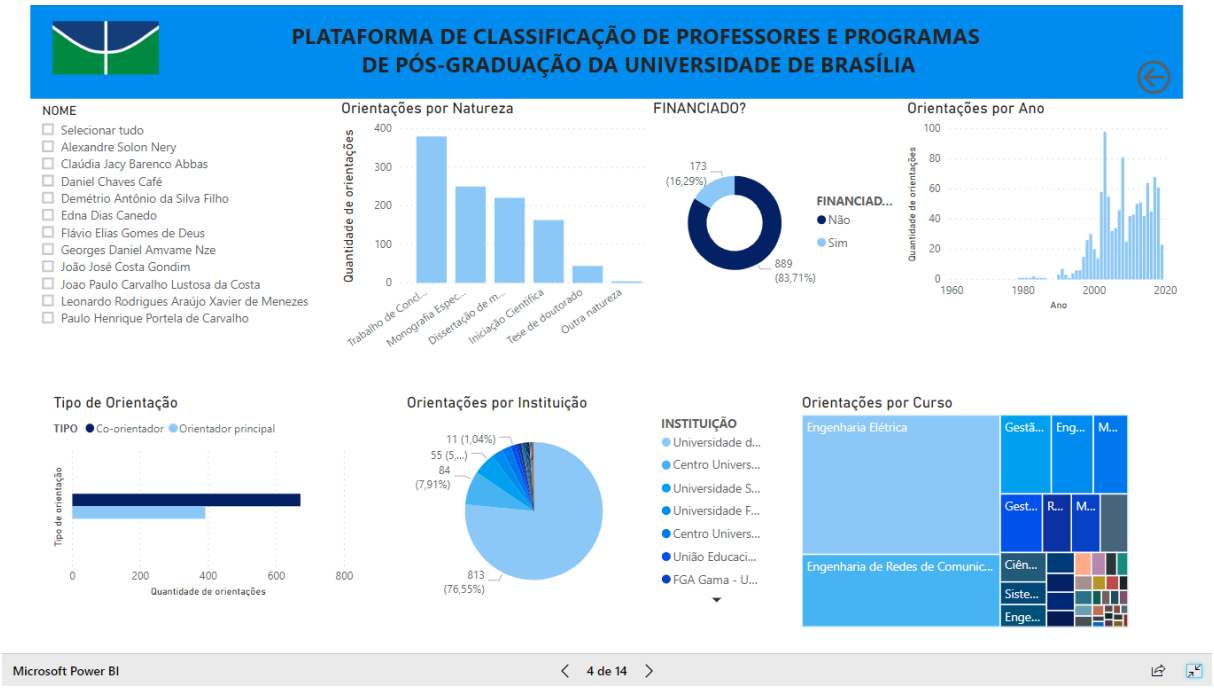

Figura A.3: Relatório total de orientações para o programa PPEE

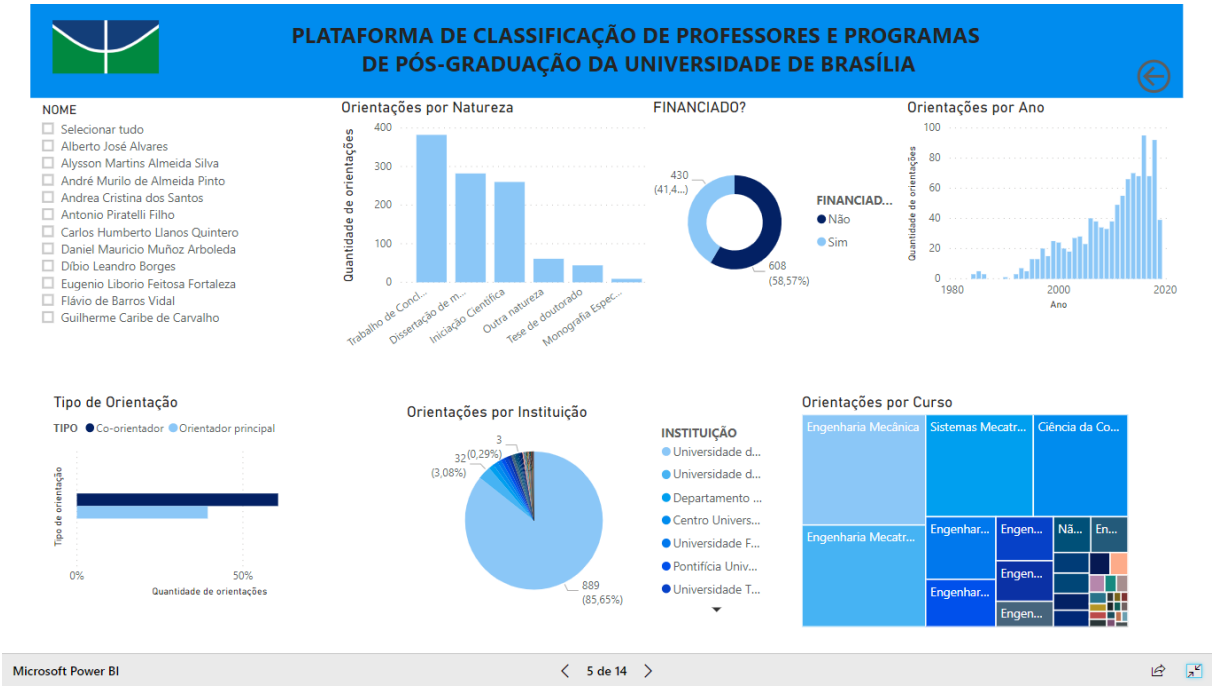

Figura A.4: Relatório total de orientações para o programa PPMEC

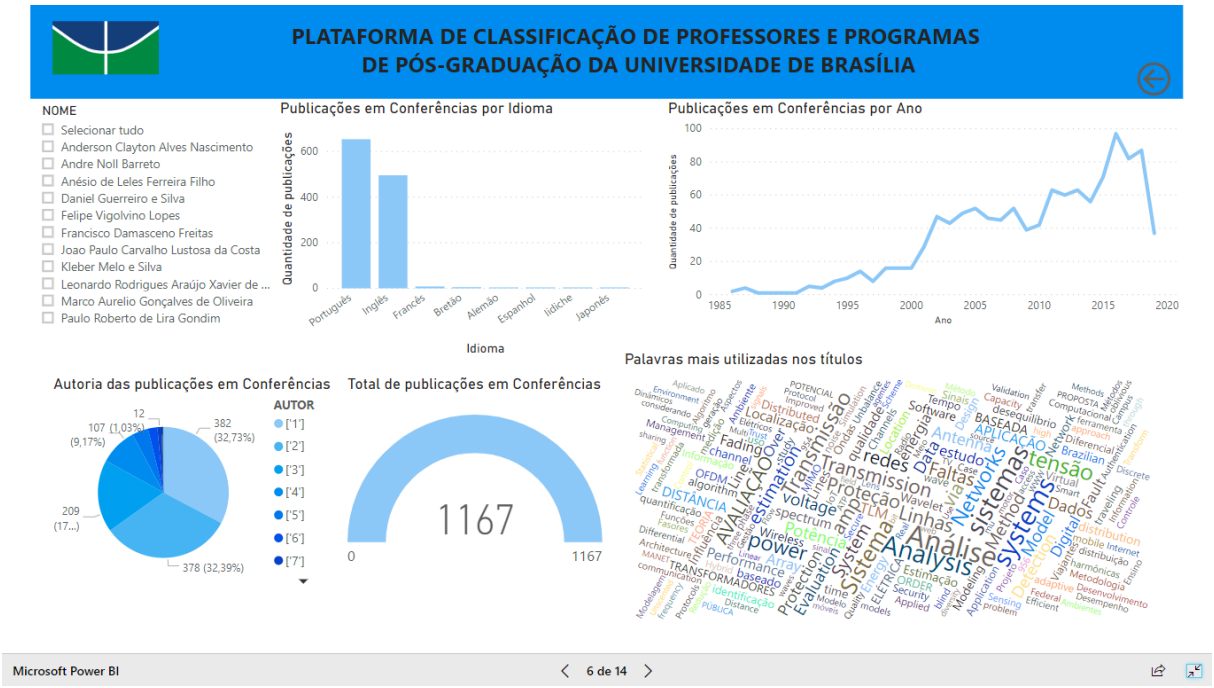

Figura A.5: Relatório total de conferências para o programa PPGEE

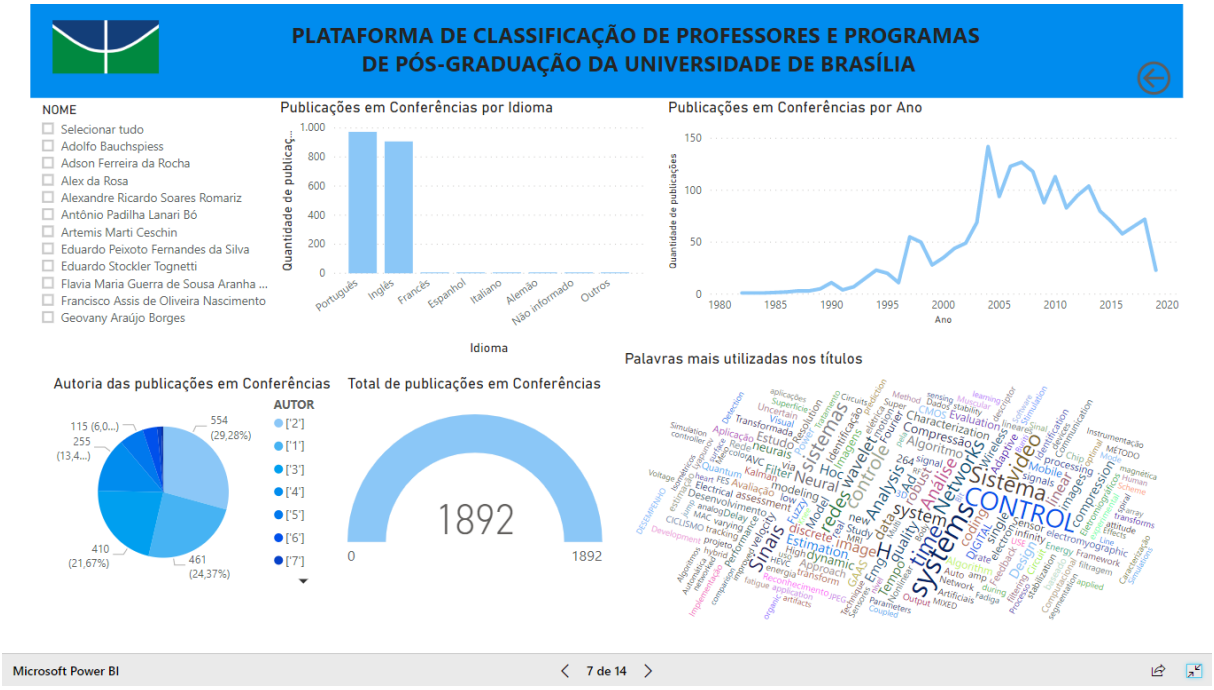

Figura A.6: Relatório total de conferências para o programa PGEA

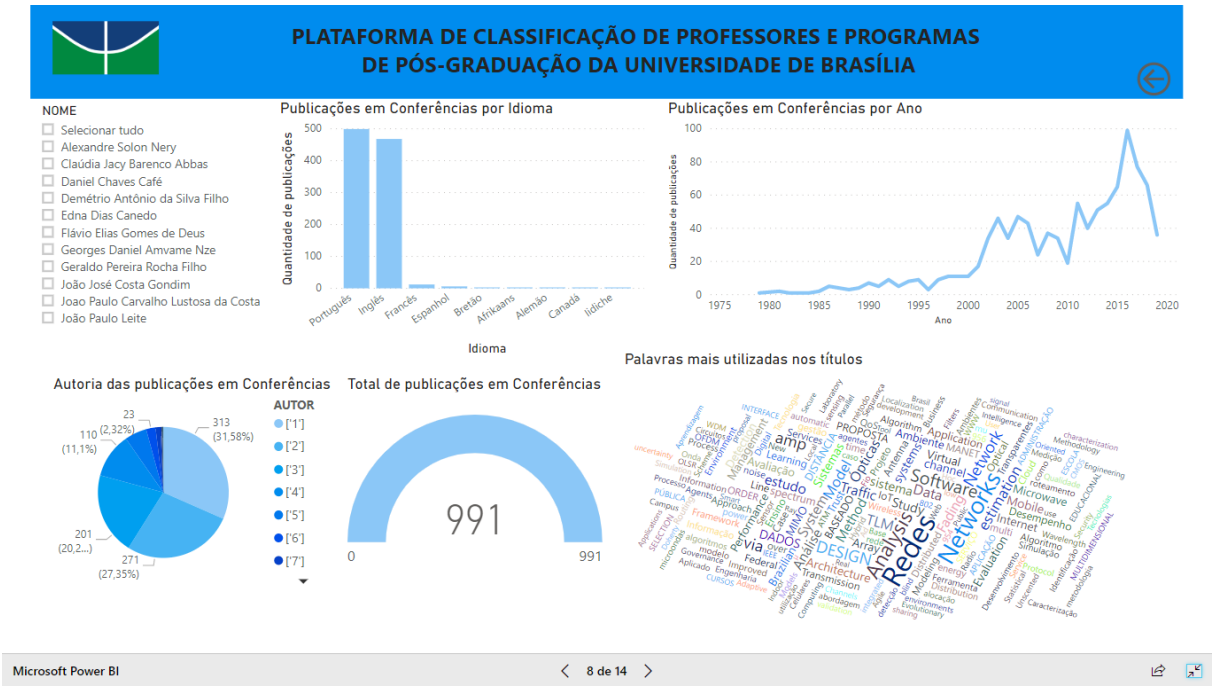

Figura A.7: Relatório total de conferências para o programa PPEE

Figura A.8: Relatório total de conferências para o programa PPMEC

Figura A.9: Relatório total de periódicos para o programa PPGEE

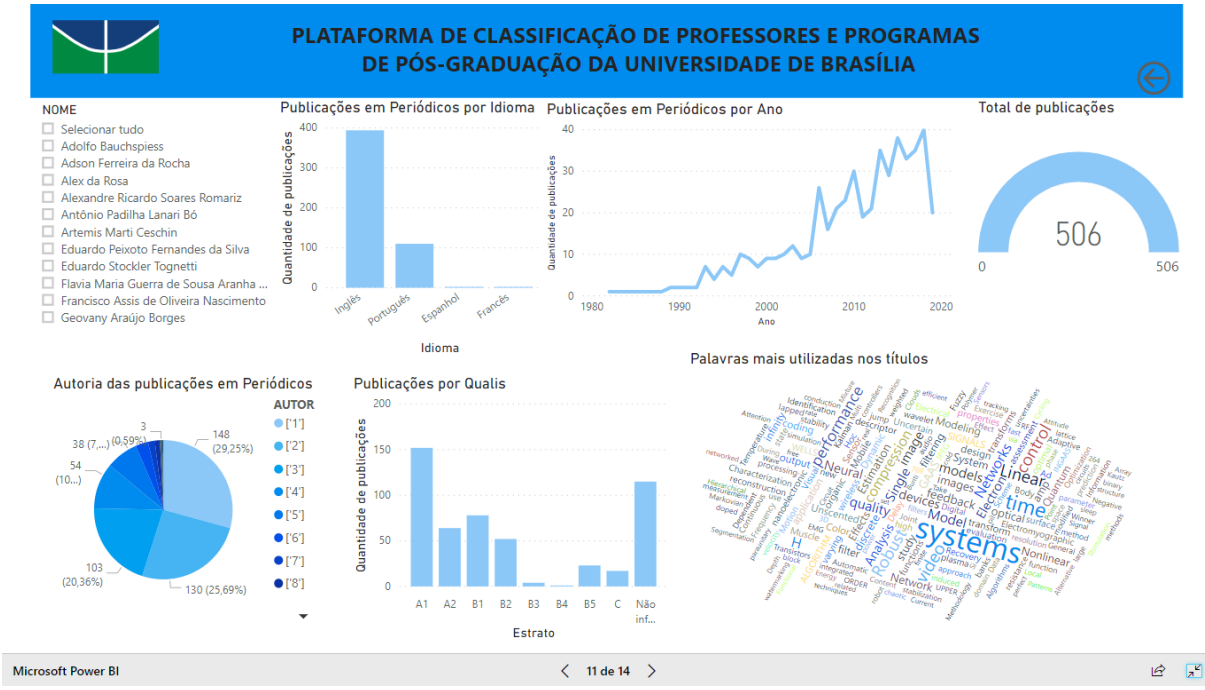

Figura A.10: Relatório total de periódicos para o programa PGEA

Figura A.11: Relatório total de periódicos para o programa PPEE

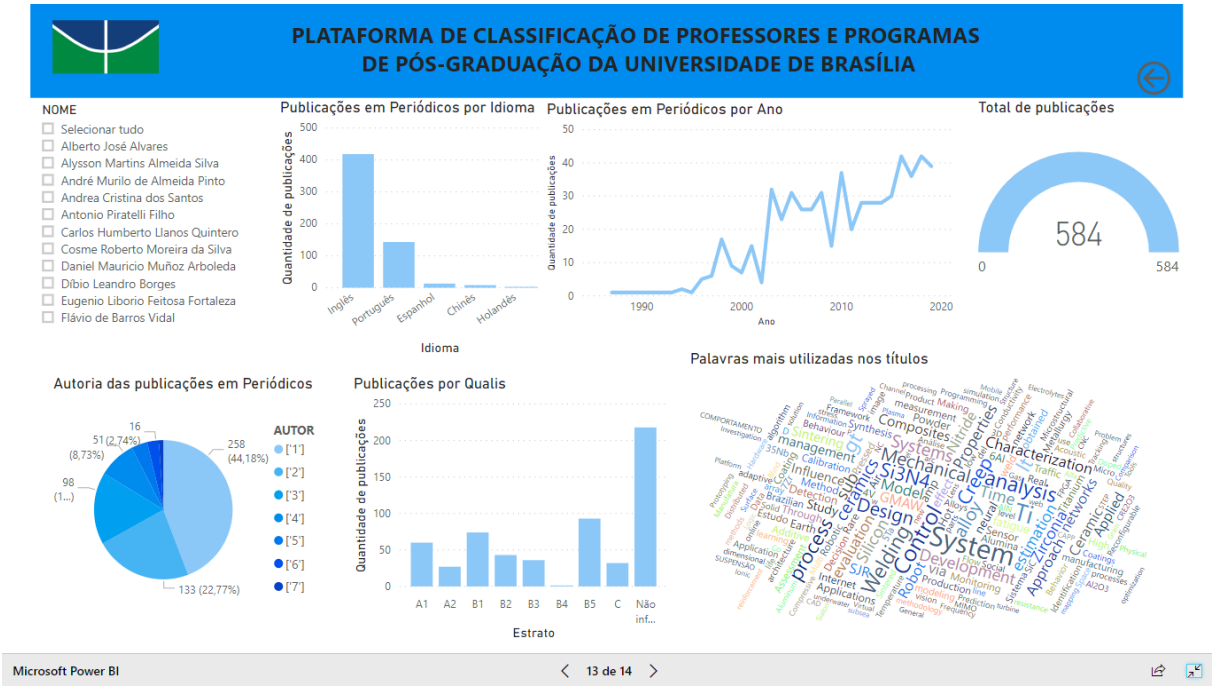

Figura A.12: Relatório total de periódicos para o programa PPMEC

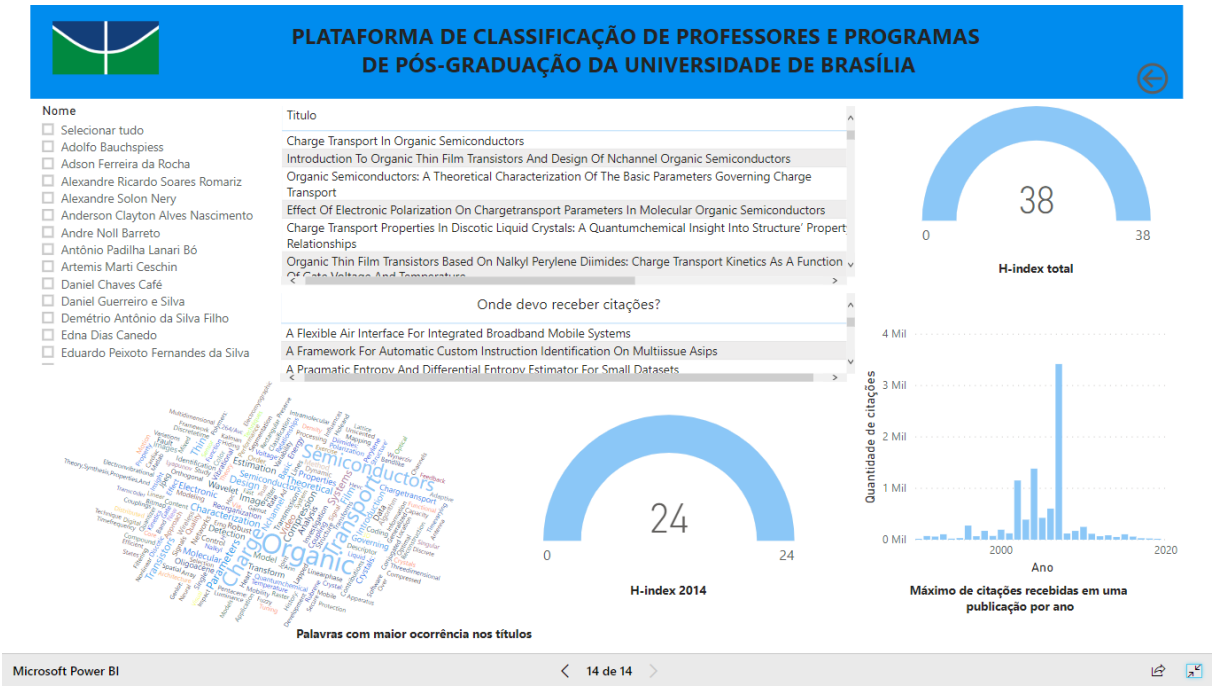

Figura A.13: Relatório total de citações para todos os programas