UNIVERSIDADE FEDERAL DO PARANÁ

RAPHAEL MARZALEK BLASI

GESTÃO DE CONHECIMENTO BASEADA EM BUSCA POR SIMILARIDADE EM PROJETOS BÁSICOS PARA EMPREENDIMENTOS DE TRANSMISSÃO

CURITIBA

2024

RAPHAEL MARZALEK BLASI

GESTÃO DE CONHECIMENTO BASEADA EM BUSCA POR SIMILARIDADE EM PROJETOS PARA EMPREENDIMENTOS DE TRANSMISSÃO

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica, Área de Sistemas de energia, Departamento de Engenharia Elétrica, Setor de Tecnologia, Universidade Federal do Paraná, como parte das exigências para a obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Alexandre Rasi Aoki
Coorientador: Prof. Dr. Mateus Duarte Teixeira

CURITIBA
2024

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA CIÊNCIA E TECNOLOGIA

Blasi, Raphael Marzalek
Gestão de conhecimento baseada em busca por similaridade em projetos para empreendimentos de transmissão. / Raphael Marzalek Blasi. – Curitiba, 2024.
1 recurso on-line : PDF.

Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica.

Orientador: Prof. Dr. Alexandre Rasi Aoki
Coorientador: Prof. Dr. Mateus Duarte Teixeira

1. Gestão do conhecimento. 2. Linhas de transmissão. I. Universidade Federal do Paraná. Programa de Pós-Graduação em Engenharia Elétrica. II. Aoki, Alexandre Rasi. V. Teixeira, Mateus Duarte. III. Título.

Bibliotecária: Roseny Rivelini Morciani CRB-9/1585

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação ENGENHARIA ELÉTRICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **RAPHAEL MARZALEK BLASI** intitulada: **GESTÃO DE CONHECIMENTO BASEADA EM BUSCA POR SIMILARIDADE EM PROJETOS BÁSICOS PARA EMPREENDIMENTOS DE TRANSMISSÃO**, sob orientação do Prof. Dr. ALEXANDRE RASI AOKI, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 01 de Fevereiro de 2024.

Assinatura Eletrônica
05/02/2024 18:34:15.0
ALEXANDRE RASI AOKI
Presidente da Banca Examinadora

Assinatura Eletrônica
05/02/2024 17:36:38.0
CLODOMIRO UNSIHUAY-VILA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
06/02/2024 08:49:53.0
EDUARDO KAZUMI YAMAKAWA
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

"A tecnologia move o homem para frente, mas apenas a sabedoria o guia na direção certa." - Bertrand Russell

AGRADECIMENTOS

Primeiramente agradeço a Deus por me conceder o dom da vida, saúde e persistência em todos esses anos de estudo e trabalho.

Agradeço à minha família, meus pais Marília e Paulo Blasi e à minha irmã Thaís Blasi que sempre me ajudaram, apoiaram e me mostraram o caminho a seguir.

Agradeço ao Prof. Dr. Alexandre Rasi Aoki pela amizade que construímos durante essa orientação, pelos ensinamentos e conselhos sempre sinceros.

Agradeço ao suporte financeiro e às oportunidades e conhecimentos obtidos no âmbito do projeto de Pesquisa e Desenvolvimento PD-06491-0563/2019 – “INTELIGÊNCIA ARTIFICIAL BASEADA EM AUTOMAÇÃO COGNITIVA E SIMILARITY MATCHING APLICADA EM ESTUDOS ELÉTRICOS DE TRANSMISSÃO E GERAÇÃO PARA EFICIÊNCIA DA GESTÃO DE OBRAS”, coordenado pelo Dr. Milton Pires Ramos e realizado em conjunto com a Companhia Paranaense de Energia (COPEL Geração e Transmissão).

Por fim, mas não menos importante, aos professores do departamento de Engenharia Elétrica da UFPR, pelos ensinamentos, dedicação e paciência.

RESUMO

A gestão de conhecimento é fundamental para garantir a eficiência e a eficácia no compartilhamento de informações relevantes entre os projetos de uma equipe e redução no tempo de desenvolvimento de novos relatórios e estudos, e por sua vez, a busca por similaridade é uma ferramenta valiosa para empresas e organizações que precisam gerenciar grandes quantidades de informações de projetos. Mais especificamente, esta estratégia é utilizada para comparar e identificar documentos semelhantes de estudos elétricos para projetos de linhas de transmissão, podendo ser realizada por meio de algoritmos de processamento de linguagem natural e técnicas analíticas de dados. O principal benefício dessa busca é ajudar a encontrar informações relevantes mais rapidamente, economizando tempo e esforço da equipe técnica. Além disso, ela também pode ser útil para detectar possíveis erros ou duplicidade de informações em relatórios futuros, o que pode garantir a precisão e integridade desses documentos. Os relatórios de projetos de linhas de transmissão são armazenados em um banco de dados, o que permite a realização da busca por similaridade sem a necessidade de categorização, a qual pode impedir a identificação de documentos similares por estarem classificados em categorias distintas. A busca é realizada através de um script desenvolvido na linguagem de programação Python, permitindo o uso de bibliotecas próprias para o processamento e facilitando o acesso e a recuperação de informações relevantes de forma rápida e compartilhável. A avaliação da abordagem foi realizada através de testes em três casos de base de dados, sendo o primeiro uma comprovação da aplicabilidade dos conceitos através de documentos com frases criadas especialmente para a validação, o segundo e o terceiro caso contam respectivamente com 12 e 72 documentos de relatórios reais disponibilizados pela Companhia Paranaense de Energia (COPEL), os quais compõem uma base de dados variada, com estudos de diferentes equipamentos de projetos de linhas de transmissão, visando a validação dos resultados obtidos, incluindo a categorização do banco de dados com maior número de arquivos e a criação de uma interface de criação de um documento de referência, o qual possui as informações que se deseja buscar nos demais relatórios. Os resultados mostraram que a abordagem é eficaz na recuperação de informações relevantes presentes nos relatórios em todos os casos de estudo, porém, sem a categorização do banco de dados, verificou-se um aumento na taxa de similaridade entre estudos de componentes distintos, o que, apesar de mapear o conjunto total de documentos, apresentou resultados de similaridade não tão relevantes quanto aos resultados obtidos no banco de dados categorizado, uma vez que a busca é direcionada previamente para os documentos de interesse, além de representar um custo computacional muito menor.

Palavras-chave: Gestão de conhecimento. Linhas de transmissão. Busca por similaridade. Eficiência.

ABSTRACT

Knowledge management is key to ensuring efficiency and effectiveness in sharing relevant information between a team's projects and reducing the development time of new reports and studies, and similarity matching is a valuable tool for companies and organizations that need to manage large amounts of project information. More specifically, this strategy is used to compare and identify similar electrical study documents for transmission line projects, and can be carried out using natural language processing algorithms and data analytics techniques. The main benefit of this search is that it helps to find relevant information more quickly, saving the technical team time and effort. It can also be useful for detecting possible errors or duplicate information in future reports, which can guarantee the accuracy and integrity of these documents. Transmission line project reports are stored in a database, which allows a similarity search to be carried out without the need for categorization, which can prevent similar documents from being identified because they are classified in different categories. The search is carried out using a script developed in the Python programming language, allowing the use of its own libraries for processing and facilitating access to and retrieval of relevant information in a quick and shareable way. The approach was evaluated by means of tests on three database cases, the first being a proof of the applicability of the concepts through documents with phrases created especially for validation, the second and third cases having respectively 12 and 72 documents from real reports made available by Companhia Paranaense de Energia (COPEL), These make up a varied database, with studies of different transmission line project equipment, with a view to validating the results obtained, including categorizing the database with the largest number of files and creating an interface for creating a reference document, which contains the information to be found in the other reports. The results showed that the approach is effective in retrieving relevant information from the reports in all case studies, but without categorizing the database, there was an increase in the similarity rate between studies of different components, which, despite mapping the total set of documents, showed similarity results that were not as relevant as the results obtained from the categorized database, since the search is previously directed to the documents of interest, as well as representing a much lower computational cost.

Keywords: Knowledge management. Transmission lines. Similarity search. Efficiency.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

ANEEL – Agência Nacional de Energia Elétrica

ATP – Alternative Transient Program

BD – Banco de Dados

CER – Compensação estática shunt

EPE – Empresa de Pesquisa Energética

ONS – Operador Nacional do Sistema Elétrico

PAR – Plano de Ampliações e Reforços

SIN – Sistema Interligado Nacional

TF-IDF - Term Frequency-Inverse Document Frequency

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTO

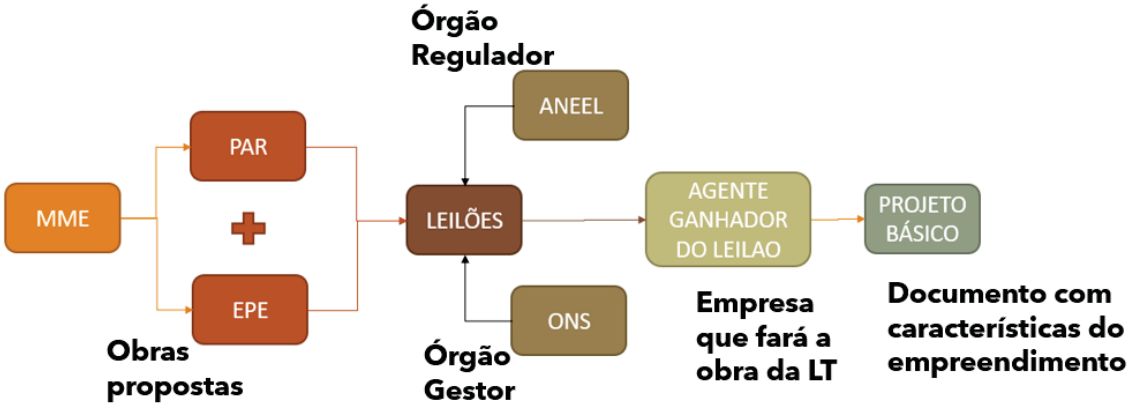
O sistema de transmissão de energia brasileiro é composto por aproximadamente 140 mil quilômetros de linhas, as quais são responsáveis pela transmissão da energia elétrica, gerada nas grandes centrais geradoras de energia, através de longas distâncias até os centros de distribuição.

Para um empreendimento de linhas de transmissão no sistema interligado nacional (SIN) o Operador Nacional do Sistema Elétrico (ONS) exige a realização de processos de definição e incorporação de novas instalações à rede elétrica básica.

Para compor as exigências, estudos realizados no Plano de Ampliação de Reforços (PAR) e pela Empresa de Pesquisa Energética (EPE) incorporam a visão da operação do sistema nas soluções propostas pelo planejamento para a proposta de novas obras. A partir disso, cabe à Agência Nacional de Energia Elétrica (ANEEL) a coordenação do processo do leilão para a realização das obras. No entanto, para a realização das obras, a ONS define requisitos técnicos para garantir a flexibilidade e confiabilidade operativa do SIN.

Após o término do leilão, o Agente que ganhou a concessão deve elaborar um projeto básico da nova instalação de transmissão, apresentando características técnicas, premissas de engenharia e especificações básicas de equipamentos, conforme apresentado na Figura 1.

Figura 1 - Fluxo para novo empreendimento de transmissão



Dessa forma, a necessidade da realização dos estudos elétricos exigidos demanda uma organização operacional das concessionárias, as quais possuem um tempo reduzido para a elaboração e entrega dos referidos estudos. Nesse ponto, a inserção de conceitos de gestão de conhecimento faz sentido na busca de eficiência e assertividade no desenvolvimento dos trabalhos relacionados.

Gestão do conhecimento refere-se ao processo de identificação, captura, armazenamento, compartilhamento e uso eficiente do conhecimento dentro de uma organização, envolvendo a criação de estratégias, políticas e práticas que visam organizar e potencializar o conhecimento existente, tanto experimental quanto documentado, para facilitar a tomada de decisões, inovação e aprendizado contínuo.

A importância da gestão do conhecimento reside na capacidade de transformar informações dispersas em ativos valiosos para a organização, facilitando o acesso rápido e preciso ao conhecimento relevante. As empresas podem tomar decisões mais embasadas, reduzir erros, promover a inovação e aprimorar seus processos fortalecendo a cultura organizacional, estimulando a aprendizagem contínua e a colaboração entre os colaboradores, o que se torna um diferencial competitivo em um ambiente empresarial cada vez mais dinâmico e complexo.

A aplicação da tecnologia de Processamento de Linguagem Natural (em inglês, *Natural Language Processing* - NLP), a qual se trata de uma subárea da inteligência artificial que tem como objetivo desenvolver técnicas e algoritmos capazes de compreender, interpretar e gerar linguagem natural humana, para a busca por similaridade em documentos representa uma opção na gestão de conhecimento corporativo. Ao utilizar algoritmos de NLP é possível analisar e compreender não apenas o conteúdo explícito, mas também o contexto e a semântica dos documentos, possibilitando a identificação de conexões e padrões anteriormente inacessíveis, permitindo não apenas a recuperação de informações relevantes, mas facilitando a disseminação do conhecimento, o compartilhamento de melhores práticas e o fortalecimento da tomada de decisões fundamentadas dentro das organizações.

1.2 OBJETIVOS

O presente estudo tem como objetivo principal o desenvolvimento de uma estratégia que utilize o método de busca por similaridade textual através da implementação de métodos de NLP, a fim de realizar comparações entre estudos

elétricos de empreendimentos de linhas de transmissão já realizados, e o novo estudo que deseja ser elaborado.

Dessa forma, os objetivos específicos são definidos por:

- Estudos das metodologias matemáticas possíveis de serem utilizadas na busca por similaridade;
- Escolha de uma metodologia para aplicação em software, bem como realização de testes e validação;
- Execução de buscas textuais em banco de relatórios referentes aos estudos elétricos de novos empreendimentos de linhas de transmissão já realizados pela Copel; e
- Apresentação dos resultados das buscas de similaridade com análise e validação em diferentes casos de estudo.

1.3 JUSTIFICATIVA

Projetos de geração e transmissão de energia elétrica usualmente são de grande porte, o que significa que falhas nos projetos básicos podem resultar em grandes prejuízos materiais e financeiros. Esses estudos são complexos e exigem o trabalho de engenheiros experientes que dispõem de um período de tempo reduzido pelo prazo definido pela ANEEL para a realização deles.

A elaboração de projetos básicos dos estudos elétricos de novos empreendimentos trata de um processo trabalhoso que segue padrões de desenvolvimento definidos por documentos disponibilizados pelo ONS, os quais são diferenciados pelas peculiaridades de cada empreendimento, porém com compartilhamento de informações entre os mesmos.

Dessa forma, com foco no curto prazo exigido pela ANEEL o presente trabalho busca facilitar e reduzir o tempo do desenvolvimento dos projetos básicos que são desenvolvidos pela concessionária.

O presente trabalho faz parte do projeto de Pesquisa e Desenvolvimento PD-06491-0563/2019 – “INTELIGÊNCIA ARTIFICIAL BASEADA EM AUTOMAÇÃO COGNITIVA E SIMILARITY MATCHING APLICADA EM ESTUDOS ELÉTRICOS DE TRANSMISSÃO E GERAÇÃO PARA EFICIÊNCIA DA GESTÃO DE OBRAS”,

coordenado pelo Dr. Milton Pires Ramos e realizado em conjunto com a Companhia Paranaense de Energia.

1.4 ESTRUTURA DA DISSERTAÇÃO

A estrutura do presente trabalho é composta por seis capítulos que visam abordar e analisar o tema proposto. O documento se inicia com a Introdução, onde são apresentados o contexto, os objetivos e a justificativa do presente estudo.

O capítulo 2, intitulado "Fundamentação Teórica", possui três subseções nas quais são apresentados conceitos e informações relevantes para o desenvolvimento do estudo, abordando as temáticas e contextualizações que compõem a evolução do presente estudo.

Já o capítulo 3, "Revisão da Literatura", utiliza um método de busca de trabalhos relevantes para a leitura, na qual são realizadas análises e filtragens dos resultados obtidos através de um software de busca em repositórios, e em seguida é abordada uma análise crítica dos estudos e pesquisas existentes sobre o tema em questão. Essa revisão tem como objetivo embasar teoricamente a pesquisa e identificar possíveis lacunas de conhecimento. Seguindo para o capítulo 4, intitulado "Material e Métodos”, são apresentados os materiais utilizados na pesquisa, como por exemplo a utilização da linguagem de programação Python e relatórios de estudos elétricos, além da descrição dos métodos adotados para o desenvolvimento da pesquisa, incluindo o fluxograma das ideias e objetivo, e o passo a passo da metodologia de desenvolvimento da ferramenta.

O capítulo 5, "Análise dos Resultados", apresenta e analisa casos de estudo, destacando os resultados obtidos e sua relevância para o estudo em questão. Por fim, o documento é finalizado com o capítulo 6, o qual contempla a conclusão do trabalho e possibilidades de estudos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados, de forma resumida, parte das diretrizes definidas pelo ONS, de forma que seja possível realizar uma contextualização técnica dos assuntos que compõem os relatórios dos projetos básicos de linhas de transmissão e consequentemente adquirir um maior entendimento da complexidade e detalhamentos necessários neste documento.

O objetivo dessa abordagem é demonstrar de forma prática a dificuldade referente a elaboração dos projetos básicos e dos estudos que fazem parte, e assim ficar mais clara a importância da gestão do conhecimento existente em relatórios já realizados anteriormente.

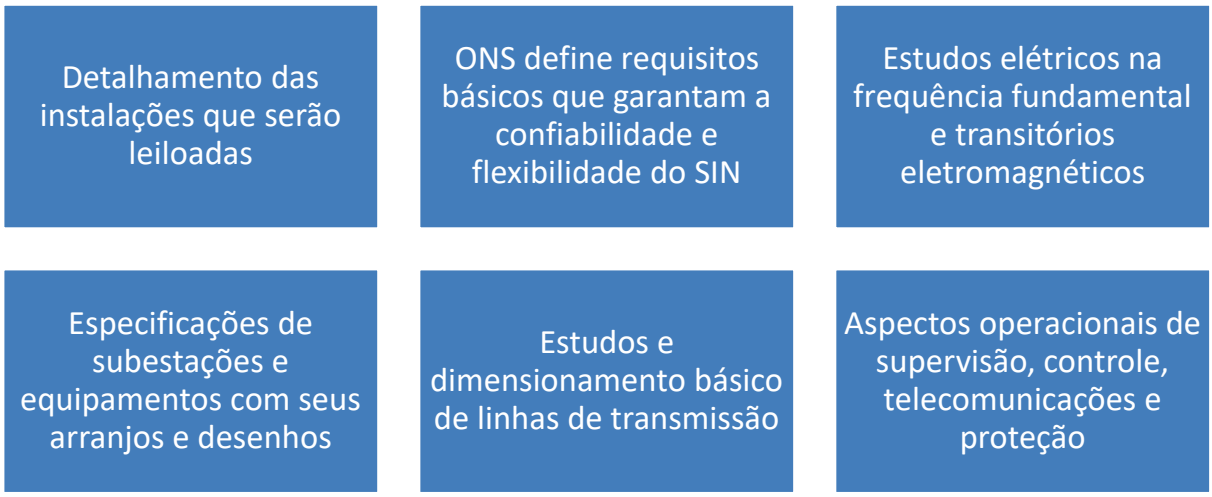
2.1 PROJETO BÁSICO DE LINHAS DE TRANSMISSÃO DO ONS

Para a incorporação de novas instalações à rede básica do sistema elétrico o ONS inclui a visão de operação do sistema nas soluções propostas pelo planejamento. Dessa forma, através dos Procedimentos de Rede para detalhamento das instalações que serão leiloadas, o ONS define requisitos básicos que garantam a confiabilidade e flexibilidade do SIN.

Dessa forma, o projeto básico que deve ser desenvolvido e entregue ao ONS trata-se do conjunto de soluções básicas propostas pela transmissora de forma que atendam aos requisitos e vise sua aprovação pela ANEEL.

Os projetos básicos podem ser divididos em quatros grupos de documentos principais, os quais abrangem os estudos elétricos na frequência fundamental e transitórios eletromagnéticos, especificações de subestações e equipamentos com seus arranjos e desenhos, estudos e dimensionamento básico de linhas de transmissão e aspectos operacionais de supervisão, controle, telecomunicações e proteção, como mostra a Figura 2.

Figura 2 - Temas abordados nos projetos básicos de transmissão



2.1.1 Diretrizes do ONS – Um breve resumo

O ONS disponibiliza um documento que apresenta as diretrizes para os projetos básicos de empreendimentos de linhas de transmissão, seguindo a divisão dos tópicos normalmente aplicáveis. A seguir serão relatados resumidamente as diretrizes listadas pelo ONS, bem como suas particularidades.

Primeiramente abordando os estudos elétricos na frequência fundamental, com o objetivo de demonstrar a conformidade do projeto básico da instalação com os requisitos estabelecidos no lote do leilão de transmissão:

- Aspectos gerais
  - Configuração da Rede e Base de Dados

    Para a realização de simulações em conformidade com os horizontes a serem avaliados no estudo, devem ser utilizadas as seguintes bases de dados:

    - Base de dados do ONS – PAR (Plano de ampliação e Reforços), para análises referentes ao ano de entrada em operação do empreendimento.

- Base de dados da EPE – PD (Plano Decenal), para as análises relacionadas com a evolução da rede em anos mais à frente, adotando como referência o horizonte do plano decenal.

- Tipos de estudos

  Os estudos elétricos a serem apresentados, independentemente da configuração proposta, devem englobar fluxo de carga, rejeição de carga, energização de linha de transmissão, estudos dinâmicos e fluxo de potência nos barramentos das subestações.

2.1.2 Estudo de fluxo de carga

Deve fazer a análise das condições da rede em regime normal de operação e sob condição de emergência de um dos componentes da transmissão (regra n-1), observando os limites de tensão nos barramentos e de carregamento das linhas de transmissão e de equipamentos, a fim de comprovar a adequação da instalação aos requisitos.

2.1.3 Estudos de energização de linha de transmissão

O objetivo desse estudo está em identificar se a compensação reativa da linha é adequada e se os recursos de controle de tensão a montante da manobra são suficientes para garantir as condições de pré-manobra. Esse estudo deve ainda pesquisar as máximas tensão em regime permanente e dinâmico na extremidade da linha e nas barras das subestações.

2.1.3 Estudo de rejeição de carga

Pesquisa das tensões máximas de regime permanente e dinâmico na extremidade da linha de transmissão e nas barras das subestações, a fim de verificar a adequação aos limites e identificar condições de abertura desfavoráveis para o disjuntor a montante da rejeição de carga. Juntamente com o estudo de energização

de linha de transmissão identifica a necessidade de compensação reativa fixa na linha de transmissão.

2.1.4 Estudos dinâmicos

Os estudos dinâmicos visam subsidiar a especificação dos equipamentos e apoio aos estudos de transitórios eletromagnéticos, se mostrando necessário para as seguintes situações:

- Geradores e compensadores síncronos
  Estudos de estabilidade eletromecânica englobando análises de manobras de energização, rejeição de carga e emergências.

- Compensação estática shunt (CER)
  Estudo da resposta dinâmica do CER frente a manobras de energização, rejeição de carga e emergências.

- Compensação série
  Estudo dinâmico para obtenção da corrente máxima de swing (corrente transitória temporária que ocorre durante distúrbios ou eventos de perturbação em sistemas elétricos de potência) sujeito ao banco de capacitor em série durante oscilações de potência no sistema elétrico.

- Viabilidade do religamento monopolar
  Demonstração que a adoção de tempo morto para religamento monopolar superior a 500 ms não compromete o desempenho dinâmico do sistema.

- Abertura em oposição de fase
  Identificação da pior condição de defasagem angular a ser imposta ao disjuntor, simulando a separação de blocos de geração com perda de sincronismo entre eles.

2.1.5 Estudo de fluxo de potência em barramentos

Os estudos de fluxo de potência nos barramentos têm como objetivo o dimensionamento dos mesmos, e dessa forma viabilizar a seleção de condutores e equipamentos que constituem os vãos de interligação, de forma que não se tornem um elemento limitador para o futuro. Para o desenvolvimento desses estudos, devem ser consideradas as seguintes premissas:

- Conhecimento da disposição física: posições exatas a serem ocupadas por cada conexão ao longo do barramento;
- Carregamentos admissíveis: conhecimento prévio de carregamentos máximos admissíveis em cada conexão;
- Condições de carga e configurações de pátio: referem-se às diferentes situações operacionais ou cenários nos quais o sistema é analisado, o que envolve examinar como o fluxo de potência se distribui nos vários barramentos do sistema sob diferentes demandas de carga, configurações de geração, e possivelmente outras condições operacionais. A condição de carga corresponde à configuração da ampliação da subestação, enquanto a configuração de pátio utiliza a base de dados de fluxo de potência da EPE e respectiva configuração da subestação; e
- Contingências: o estudo deve ser desenvolvido a partir da condição de configuração íntegra da subestação, simulando a perda de elementos internos e externos (regra n-1). A escolha contingência deve buscar o maior impacto de circulação de corrente.

2.1.6 Estudos de transitórios eletromagnéticos

Em seguida são descritas as diretrizes para os estudos de transitórios eletromagnéticos que devem ser realizados, com cada uma das particularidades dos equipamentos que compõem os empreendimentos.

- Aspectos Gerais

  De forma geral é possível dividir as grandezas a serem analisadas durante a operação transitória em quatro grupos, sendo eles: sobretensões, sobrecorrentes, formas de onda (harmônicas) e transitórios eletromagnéticos/eletromecânicos. Os tipos de estudos transitórios eletromagnéticos necessários para o projeto básico podem variar de acordo

com os equipamentos que fazem parte do empreendimento e características das linhas de transmissão.

- Modelagem da Rede

  Considerado um dos aspectos mais importantes juntamente com o ajuste da base de dados para as simulações com o software ATP (Alternative Transient Program), sendo seccionado nos seguintes subtópicos:

  - Barras de fronteira – Equivalentes de curto-circuito

    Ao contrário das simulações de fluxo de potência, onde a rede é equilibrada e pode ser representada de forma monofásica, nas simulações de transitórios eletromagnéticos deve-se considerar a representação trifásica, levando em consideração possíveis acoplamentos. Tendo em vista que um dos efeitos transitórios mais significativos é causado pela aplicação ou remoção de defeito, os valores de magnitude de curtos-circuitos devem ser próximos aos obtidos em regime permanente, caso a rede de sequência positiva e de sequência zero sejam bem representadas. Quando a rede é reduzida aos itens mais elementares, composta de resistências, indutâncias e capacitâncias, deve-se dar atenção ao amortecimento da rede elétrica, o qual se manifesta nas representações das cargas e nos fatores de qualidade dos equipamentos. Dessa forma, os equivalentes de rede a serem utilizados devem ser indicados por suas impedâncias de curto-circuito de sequência zero e positiva, vistas a partir das barras de fronteira. Para a definição das barras de fronteira, são selecionadas as que possuem menos influência sobre o comportamento transitório do sistema, e para a inserção dos equivalentes na localização dessas barras deve-se adotar a seguinte regra: “Entre a(s) barra(s) focalizada(s) no estudo e as barras de fronteira devem existir, pelo menos, 2 (duas) outras barras”. Todo projeto básico deve apresentar um diagrama unifilar de rede representada, identificando claramente o posicionamento dos equivalentes de curto-circuito, da geração representada, das cargas e das linhas incluídas no caso base.

  - Dados dos componentes da Rede

Devem ser apresentados todos os dados dos componentes utilizados na modelagem de rede no programa ATP, sendo destacados:

- Linhas de Transmissão

  Apresentar em forma de tabela os dados de sequência zero e positiva, por classe de tensão por quilometro de todas as linhas representadas. Parâmetros das linhas de transmissão pertencentes ao Lote do Leilão de Transmissão. Geometria das torres consideradas no estudo com estudos de transitórios eletromagnéticos.

- Transformadores

  Devem apresentar os dados de todos os transformadores apresentados bem como quais foram representados com as respectivas curvas de saturação. Para modelagem da histerese magnética devem ser apresentados o valor de tensão e reatância saturada.

- Reatores

  Devem ser apresentados todos os reatores informando a potência (Mvar), tensão correspondente (kV) e se é manobrável ou fixo, bem como o valor de aterramento de neutro (ohms).

- Bancos de capacitores shunt

  Apresentar tabela com todos os bancos de capacitores informando potência (Mvar) e tensão (kV) correspondente. Caso existam reatores limitados em série com o banco essas características também devem ser apresentadas. No caso de adoção de disjuntores com resistores, suas características devem ser informadas no item disjuntores.

- Compensação série

  Apresentar o valor da reatância Capacitiva em ohms e o valor em Mvar correspondente para a tensão respectiva.

- Compensador estático

  Detalhar o ponto de operação do ajuste de regime permanente, onde a desconsideração do efeito desse equipamento, para avaliações comuns do projeto básico, é considerada uma hipótese conservadora. Os modelos de compensador são de

responsabilidade da transmissora e devem ser aferidos pelo fornecedor do equipamento com informações de curva de saturação, malha de aquisição de dados da rede e suas filtragens, malha de controle, capacidade de sobrecarga indutiva, entre outros.

- Para-raios

  Além dos dados de tensão nominal, capacidade de dissipação de energia e pontos utilizados na modelagem, nos estudos de manobra deve ser apresentada a característica da curva de descarga para 30 x 60 μs. Já para estudos de coordenação de isolamento deve ser apresentada a curva de 1,2 x 50 μs.

- Equivalente de rede

  Apresentação dos dados dos equivalentes próprios e as impedâncias de transferência, informando também os parâmetros de sequência zero e positiva.

- Unidades geradoras

  No caso de modelagem simplificada deve apresentar as reatâncias subtransitórias e as potências das unidades geradoras (MVA). Já para casos específicos de modelagem de máquina completa deve apresentar as reatâncias e as constantes de tempo das máquinas, levando em conta o efeito de reguladores de tensão.

- Disjuntores

  Para as simulações estáticas, informar o número de chaveamentos adotados, o desvio padrão com o truncamento das chaves correspondentes ao contato principal e o auxiliar do disjuntor. Quando empregados resistores de pré-inserção, deve ser informado o valor do resistor (ohms) e o tempo da inserção. Já para dispositivo sincronizador, devem ser fornecidos os parâmetros do conjunto mecânico-eletrônico.

- Parâmetros da Simulação
  - Tempo total de simulação (ms)

Para estudos das manobras de linhas de Transmissão e transformadores deve-se considerar o tempo total em torno de 300 a 500 ms após a manobra em análise. Para o caso de religamento, acrescentar o tempo morto a fim de obter o tempo total de simulação. Para a manobra de energização de transformadores, deve estender a simulação por cerca de um segundo para observação da tendência de amortecimento das formas de onda.

- Passo de integração (ms)
  O passo de integração deve estar adequado em função da faixa de frequências envolvidas no fenômeno em análise e características dos componentes modelados na base de dados do ATP.

- Avaliação da adequação de compensação shunt de linhas de transmissão
  Recomenda-se realizar o estudo com o objetivo de investigar a possibilidade de ocorrência de ressonância e na própria linha, o que pode causar a indução de tensões elevadas, resultando em dificuldades na extinção do arco secundário e no atraso do decaimento da carga residual da linha de Transmissão, o que inviabiliza as manobras de religamento tripolar e monopolar, além de existir o risco de danificação dos equipamentos instalados. Uma vez identificada esta condição, devem ser propostas medidas de mitigação, como a instalação de reatores ou resistores de neutro e a adequação dos esquemas de transposição da linha quando esta for empregada, uma vez que essa interação influencia diretamente as manobras das chaves de aterramento da linha em função dos acoplamentos eletrostáticos e eletromagnéticos entre linhas na mesma estrutura ou mesma faixa de servidão.
  As principais avaliações que devem ser realizadas são:
  - Do grau de compensação da linha de transmissão e da possibilidade de ocorrer ressonância na frequência fundamental; e
  - Existência de ressonância por indução entre os circuitos ou entre linhas paralelas, sob condição de abertura tripolar ou monopolar.

Após a modelagem completa das linhas, deve-se efetuar a investigação da existência de ressonâncias entre os circuitos ou linhas em paralelo, considerando as compensações shunts, com a simulação da resposta em frequência vista dos terminais da linha associados as fases ou circuitos investigados. Esta avaliação deve verificar se todos os reatores terminais da linha estão conectados, representando adequadamente o aterramento utilizado nesses reatores.

- Estudo de energização de linhas de transmissão

  Este estudo tem por objetivo avaliar as máximas sobretensões transitórias a serem impostas aos barramentos das subestações e aos terminais das linhas de transmissão, as energias dissipadas nos para-raios de linha tendo em vista o dimensionamento desses equipamentos sob o ponto de vista da capacidade de absorção de energia e a adequação da coordenação de isolamento das estruturas de linhas de transmissão frente a surtos de manobras.

  As diretrizes para os estudos de energização consideram a possibilidade de energização por ambos os sentidos da linha, com o sistema íntegro e degradado com contingência (n – 1). As avaliações devem ser efetuadas com e sem a aplicação de defeitos na linha, e quando aplicados devem estar, pelo menos, localizados no terminal energizado, no meio da linha e na extremidade oposta.

  O estudo das manobras de energização deve ser realizado de maneira estática considerando a amostragem de 200 chaveamentos. Os procedimentos de rede do ONS destacam que o disjuntor deve ser representado pela chave estatística, com tempos de fechamento caracterizados por distribuição gaussiana.

- Estudo de religamento tripolar de linhas de transmissão

  Em conjunto aos estudos de energização de linhas de Transmissão e rejeição de carga esse estudo tem como objetivos: avaliar máxima sobretensões transitórias; avaliar as energias dissipadas nos para-raios; e adequar a coordenação de isolamento das estruturas

  De forma geral, as manobras de religamento tripolar resultam os piores valores de sobretensão em linhas de transmissão.

A sequência de eventos a ser considerada é: defeito monofásico em um dos terminais, abertura tripolar do terminal mais próximo, abertura da outra extremidade da linha com tempo correspondente ao disparo da proteção, extinção do defeito, contagem do tempo morto para o religamento e simulação do religamento estatístico.

Assim como no estudo de energização de linhas de transmissão o disjuntor deve ser representado através de chave estatística. Já as análises devem focar nos aspectos relevantes para o projeto básico da instalação, como por exemplo, valor das sobretensões nos equipamentos localizados nas subestações e nos terminais, a energia dissipada pelos para-raios de linha, e verificação da coordenação de isolamento frente a sobretensões fase-fase e fase-terra.

- Rejeição de carga

  O estudo de rejeição de carga tem por objetivo avaliar as máximas sobretensões transitórias que serão impostas ao barramento das subestações e aos equipamentos terminais das linhas de transmissão.

  Entre as diretrizes deste estudo devem ser consideradas as manobras sem aplicação de defeito prévio, bem como com aplicação de defeito monofásico prévio, aplicação de defeito posterior à rejeição, análise dos tempos de eliminação dos defeitos e indisponibilidade dos reatores manobráveis.

  Na simulação, o fluxo de potência deve ser ajustado de maneira que a linha de transmissão em estudo esteja na condição de carregamento com valor o mais próximo possível do limite de longa duração da linha.

- Estudo de religamento monopolar de linhas de transmissão

  Com o objetivo da implantação do religamento monopolar esse estudo prioriza as soluções técnicas para a viabilização, sem comprometer o desempenho do sistema, abrangendo o estudo das sobretensões transitórias de manobra e o estudo da extinção do arco secundário.

  O estudo de manobra do religamento monopolar considera os seguintes eventos: defeito monofásico em um dos terminais ou no meio da linha, abertura monopolar terminal mais próximo ao defeito, abertura monopolar da outra extremidade da linha com tempo correspondente de disparo da proteção,

extinção do defeito, contagem do tempo morto para religamento, e simulação do religamento estatístico.

- Estudo de energização de transformadores

  Busca identificar as tensões e correntes resultantes da manobra da energização de transformadores impostas aos próprios equipamentos, orientando quanto à necessidade ou não da instalação de resistores pré-inserção ou dispositivos de manobra controlada para os disjuntores de manobra dos transformadores.

  Para a realização do estudo, deve ser considerada a possibilidade da energização por ambos os terminais do transformador, sob indisponibilidade de um componente da rede (n - 1), tendo em vista as manobras de recomposição no sistema.

  A fim de impor a condição mais crítica, o fluxo residual deve ser considerado com seu valor máximo numa das fases, e abranger o fechamento do disjuntor no instante de polaridade de fluxo inverso em relação ao fluxo residual.

  Caso a curva de saturação do transformador ainda não esteja disponível na etapa de projeto básico, devem ser adotados dados típicos.

  Em função da aleatoriedade do fechamento dos polos do disjuntor, o estudo de manobras deve ser realizado de maneira estatística, com modelagem do disjuntor no programa ATP, aplicando as recomendações do item 9.2.1 dos Procedimentos de Rede do ONS.

- Energização de banco de capacitores

  Avalição dos transitórios de corrente e tensão resultantes das manobras de energização, aplicação e eliminação de defeito associado a banco de capacitores em derivação, cujos transitórios podem resultar em impactos sobre os demais equipamentos locais, considerando a necessidade de indutores limitadores de corrente a serem instalados em série com o banco de capacitores, e para controle das sobretensões além da recomendação de disjuntores dotados com resistores de pré-inserção ou de sincronizadores, além do dimensionamento dos para-raios.

- Tensão de restabelecimento transitória (TRT)

A interrupção da corrente de circuito é dividida em três etapas, sendo a primeira predominada pelo efeito Joule, operando com correte de carga nos contatos fechados. Na segunda, com os contatos em movimento, ocorre a criação do arco elétrico, iniciando a fase térmica da interrupção. A última etapa é a fase dielétrica da interrupção, com a corrente passando para zero, ocorre a resposta do sistema surgindo uma diferença de tensão nos contatos do disjuntor. A caracterização da TRT é dada através dos parâmetros de taxa de crescimento da tensão transitória e o valor de pico da onda da TRT. Esses valores são utilizados para comparativos com valores normalizados a fim de concluir se o equipamento está adequado para a manobra.

- Assimetria de corrente de curto-circuito

  Esse estudo busca subsidiar tanto a definição da capacidade de estabelecimento nominal em curto-circuito dos disjuntores, como a corrente suportável nominal em curto-circuito dos equipamentos. A definição da capacidade de interrupção nominal em curto-circuito de disjuntores é caracterizada por dois valores, sendo eles, o valor eficaz de sua componente CA (corrente alternada) e a porcentagem da componente CC (corrente contínua).

- Tensões e correntes induzidas em lâminas de terra de seccionadoras

  Para o cálculo dos valores das correntes e tensões induzidas em lâmina de terra das seccionadoras instaladas nos terminais das linhas de transmissão e em manobras, considera-se as condições mais severas dos acoplamentos existentes entre circuitos paralelos e de carregamento máximo. A indução dessas correntes surge do acoplamento eletromagnético e eletrostático entre circuitos.

  Ao efetuar a manobra de abertura das chaves de aterramento, é possível que a corrente a ser interrompida tenha valores elevados, juntamente com o surgimento de tensão de restabelecimento transitória de abertura.

- Estudo de coordenação de isolamento

  Com o objetivo de avaliar sobretensões que atingem os equipamentos de entrada de linha e do interior da subestação, decorrentes da ocorrência de

descargas elétricas nas linhas de transmissão, esse estudo estabelece os níveis básicos de isolamento para os equipamentos e as características nominais dos para-raios com suas localizações na instalação.

As descargas elétricas que atingem a rede elétrica podem ser classificadas como diretas ou indiretas, o que diferencia os cálculos a serem realizados para a coordenação de isoladores. Já o estudo para descargas atmosféricas em subestações é desenvolvido de maneira determinística.

- Estudos associados à compensação estática shunt (CER)

  Na primeira etapa desse estudo são estabelecidas as condições de contorno sistêmicas a fim de subsidiar o projeto do equipamento pelo fabricante, a qual deve ser desenvolvida em conjunto com o projeto básico da instalação.

  A segunda etapa do estudo é desenvolvida pelo fabricante com base nos resultados da primeira etapa desenvolvida pela Transmissora. Nela o fabricante deve fornecer os modelos computacionais da CER com base nas características do equipamento a ser utilizado em futuros estudos de sistema.
- Estudos associados à compensação série

  Esse estudo também é desenvolvido em duas partes, sendo que a primeira tem como objetivo estabelecer os requisitos de cunho sistêmico necessários à sua especificação e à efetiva tomada de preços, desenvolvido pela transmissora. Já na segunda etapa, desenvolvida pelo fabricante, tem como objetivo o seu dimensionamento e o atendimento dos requisitos estabelecidos na especificação de compra.

2.2 Processamento de Linguagem Natural (NLP)

A Processamento de Linguagem Natural (do inglês, Natural Language Processing - NLP) é uma subárea da inteligência artificial que tem como objetivo desenvolver técnicas e algoritmos capazes de compreender, interpretar e gerar linguagem natural humana, sendo um campo de estudo interdisciplinar que envolve a computação, a linguística e outras áreas afins, e tem como um dos principais desafios a complexidade e diversidade da linguagem natural, que pode ser ambígua, subjetiva e contextualmente dependente.

Entre as tarefas que a NLP busca realizar, uma das mais importantes é a busca por similaridade de documentos. Isso é possível graças a técnicas de processamento de texto que permitem identificar a similaridade entre diferentes documentos, permitindo que sejam agrupados ou comparados de acordo com seus conteúdos. Essa tarefa é particularmente útil em áreas como a classificação de documentos, a detecção de plágio, a análise de sentimentos e a recuperação de informações, entre outras. (JURAFSKY, 2023)

Um dos principais desafios enfrentados pela NLP na busca por similaridade de documentos é a variação na representação dos textos, que pode afetar a precisão e a acurácia dos algoritmos utilizados. Diferentes documentos podem conter informações semânticas similares, mas apresentadas de maneira distinta, o que pode dificultar a identificação de sua similaridade, e para lidar com essa variação, são utilizadas técnicas de pré-processamento de texto, como a normalização e a filtragem de termos, bem como técnicas de modelagem de linguagem, como os modelos de espaço vetorial e os modelos de tópicos.

A Tabela 1 apresenta uma análise dos principais métodos de aplicação do NLP com exemplos de aplicações e seus respectivos pontos positivos e negativos.

Tabela 1 - Análise dos principais métodos de NLP

Método de NLP	**Exemplos de aplicações**	**Pontos positivos**	**Pontos negativos**
Análise de sentimento	Classificação de sentimentos em textos, avaliações de produtos, redes sociais	Entender como as pessoas se sentem em relação a algo; melhorar a experiência do usuário	Difícil determinar o tom sarcástico ou irônico; pode não ser tão preciso em línguas diferentes do inglês
Reconhecimento de entidades nomeadas (NER)	Identificação de nomes de pessoas, lugares, organizações em textos	Criar bancos de dados de informações; ajuda a automatizar tarefas de processamento de texto	Difícil identificar nomes próprios que não são comuns em uma determinada língua; erros de reconhecimento

Modelagem de tópicos	Identificação de tópicos e temas em conjuntos de textos	Entender o conteúdo de grandes volumes de texto; agrupar documentos	Resultados podem não ser precisos ou relevantes; difícil escolher o número correto de tópicos
Reconhecimento de padrões	Identificação de padrões em textos, como endereços de e-mail, números de telefone, datas, etc.	Extrair informações importantes de grandes volumes de texto; usado para automatizar tarefas de preenchimento de formulários	Pode haver erros de reconhecimento; difícil lidar com diferentes formatos e estilos de texto

Fonte: O Autor (2023)

2.3 Cálculo de Similaridade

O termo similaridade refere-se a mostrar computacionalmente, através de um valor, o quanto dois objetos são semelhantes entre si, ou seja, características que são compartilhadas entre eles.

O objeto pode ser entendido com textos, imagens, documentos, estruturas de dados, sons, entre outros.

O conceito de similaridade pode ser subjetivo de acordo com o observador, sendo dessa forma, necessária a criação de métricas para o cálculo. O grau de similaridade pode ser gerado através de diversos mecanismos de cálculos já desenvolvidos, os quais utilizam conceitos matemáticos bem estruturados. (DORNELES, 2022)

Uma das técnicas mais comuns para identificar documentos mais similares envolve a aplicação de algoritmos que comparam o conteúdo dos documentos com o objetivo de determinar o grau de similaridade entre eles.

Para realizar o cálculo de similaridade entre documentos, geralmente é utilizada uma técnica de comparação de várias características dos documentos, como palavras-chave, tópicos e estrutura do documento, a fim de identificar os documentos mais similares. Uma técnica comum de análise de similaridade de texto é o cálculo de

distância de Levenshtein, que mede a distância entre duas cadeias de caracteres, ou a análise de tópicos, que utiliza algoritmos de aprendizado de máquina para identificar os tópicos abordados em cada documento e compará-los com os tópicos dos outros documentos.

A análise de similaridade de texto pode ser aplicada a vários tipos de documentos, incluindo textos de mídia social, artigos científicos, documentos empresariais e utilizada para identificar documentos relevantes em grandes conjuntos de dados, seguindo os conceitos da Figura 3.

Figura 3 - Sequência de conceitos para a busca por similaridade

2.3.1 Conceitos matemáticos utilizados para o cálculo de similaridade

Existem diversas técnicas matemáticas possíveis de serem utilizadas para encontrar a distância entre objetos. Essa distância representa o escore de similaridade, sendo quanto menor a distância, maior a similaridade entre os objetos analisados.

Para ser possível aplicar modelos matemáticos de cálculo de similaridade, é necessário preparar os dados, de forma que estejam de acordo com as fórmulas numéricas dos métodos, sendo chamado de pré-processamento dos dados.

Essa modificação dos dados varia com o tipo de objeto que será analisado, por exemplo, textos, imagens e páginas da web, cada um tem um processo diferente de transformação das informações dependendo do método matemático do cálculo da distância e formato original.

O presente trabalho tem foco especial na análise de textos, e dessa forma envolve a seleção dos dados que constituem a base de dados de texto de interesse e o trabalho inicial e dessa forma poder expressar o conteúdo desses arquivos. No pré-processamento devem ser mantidas informações como morfologia e significados

intrínsecos das palavras, bem como contexto em que foram utilizados, e, ao mesmo tempo, promover uma redução dimensional do texto analisado. (FREDIGO et al., 2013)

A seguir são apresentadas a sequência de algumas das técnicas de pré-processamento de texto mais utilizadas na literatura:

- Filtering

  Nessa técnica é realizada a filtragem de sinais de pontuação, ou seja, recursos prosódicos que, apesar da remoção, não alteram o significado do documento, apesar de correr o risco de alterar o significado específico de frases.

- Tokenization

  Uma vez removida a pontuação, essa técnica busca identificar as palavras que geram o significado das frases. Dessa forma o significado é estruturado através de relação entre palavras.

- Stemming

  Após serem identificadas as palavras que geram significado, essas são reduzidas para a sua forma mais básica, ou seja, seus radicais. O objetivo dessa etapa é remover estilos de escrita e expressar o significado direto, e assim tornar a comparação entre textos mais precisa.

- Stopword Removal

  Usualmente textos possuem uma grande quantidade de palavras, o que pode ser considerados dados, o que impacta diretamente no tempo de processamento, sendo assim necessária a relaxação de filtragem de dados, ou seja, remoção de palavras que não agregam grande valor ao texto. De forma geral, palavras que não sejam substantivo, adjetivo ou verbos tendem a ser removidas.

- Thesaurus

  Mesmo com a remoção de palavras pouco significantes, é possível fazer uma segunda filtragem de palavras com base em seus significados semelhantes e repetição.

- Cálculo de relevância de palavra

  Algumas palavras em um texto possuem maior relevância que outras, de forma que a ideia do cálculo de relevância objetiva vincular um peso referente ao uso do termo dentro do texto. Para isso, existem diferentes formas de definir o peso, a maioria baseada em cálculos simples de frequência:

  - Frequência absoluta

    Conhecida como frequência do termo, representa a quantidade de vezes que um termo aparece no documento. Trata-se de uma medida mais simples que as demais, e não leva em conta a quantidade total de palavras do texto completo, e dessa forma, reduzindo a precisão, pois uma palavra muito frequente em um texto curto pode ter o mesmo peso de uma palavra pouco comum em um texto longo.

  - Frequência relativa

    Esse tipo de análise corrige o ponto fraco da frequência absoluta, considerando a quantidade total de palavras no documento, e normaliza os valores dos pesos com base nessa informação. A equação 1 apresenta o cálculo matemático.

$$F_{rel}(x) = \frac{F_{abs}}{N} \quad (1)$$

    Onde:

    F_{rel}: frequência relativa da palavra “x”

    F_{abs}: frequência absoluta do termo “x”

    N: quantidade total de palavras no documento

- Frequência inversa de documentos

  Essa técnica, chamada de Frequência Inversa de Documentos ou TF-IDF (do inglês, Term Frequency-Inverse Document Frequency) busca normalizar termos frequentes com base em todos os documentos analisados. Essa técnica baseia-se na frequência absoluta do termo no documento, e também na frequência do mesmo termo em todos os documentos. Esse processo é capaz de aumentar a importância dos termos que aparecem em poucos documentos, e diminuir a de termos que se repetem em todos eles. A equação 2 apresenta o cálculo matemático dessa técnica.

$$TF\_IDF_{i,j} = tf_{i,j} x\ log\left(\frac{N}{df_i}\right) \quad (2)$$

  Onde;

  $TF\_IDF_{i,j}$: peso da palavra vetorizada pelo método TF_IDF

  $tf_{i,j}$:número de ocorrências da palavra i no documento j

  N: número total de documentos

  df_i: número de documentos que contém a palavra i

- Vetorização de textos

  Nessa última etapa, o objetivo da vetorização de textos é representar o documento na forma de um vetor de termos, ou seja, transforma texto em números. Cada documento transforma-se em um vetor de valores que representam a frequência dos termos, associando assim ao documento por pares de elementos na forma: (palavra 1, peso 1), (palavra 2, peso 2), ... , (palavra n, peso n).

  Vale ressaltar que esses vetores possuem não apenas as palavras que estão presentes naquele texto, mas sim todas as palavras de todos os textos em análise, variando seus valores de 0, para quando o termo não aparece no documento em análise, até 1 para termos extremamente importantes, como mostra a Tabela 2.

Tabela 2 - Exemplo de valores de peso após vetorização de texto

	25	**100**	**Acordo**	**Alto**	**Barra**	**Casa**	**Corrente**	**Dados**	**...**	**Tensão**	**Trabalho**	**Vão**
Peso	0,01	0,05	0,02	0,08	0,12	0,09	0,11	0,10	...	0,14	0,03	0,05

Uma vez realizado o pré-processamento do texto, é possível iniciar o cálculo de similaridade propriamente. A seguir estão listadas algumas das técnicas matemáticas de cálculo de similaridade (Ufrgs, 2022).

- Distância Euclidiana

É o comprimento do segmento que conecta dois pontos no espaço vetorial, como mostra a equação 3 e a Figura 4.

$$D(P,Q) = \sqrt{\sum_{i=1}^{n} (Pi - Qi)^2} \quad (3)$$

Onde:

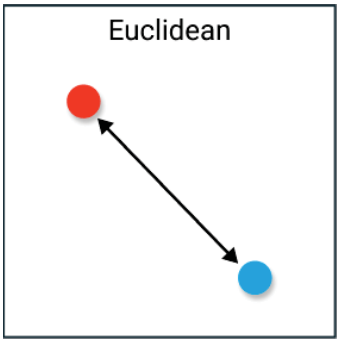
D: distância vetorial do ponto “P” ao ponto “Q”

n: número de dimensões no espaço vetorial

Pi: coordenado do ponto “P” na dimensão i

Qi: coordenado do ponto “Q” na dimensão i

Figura 4 - Distância Euclidiana



Fonte: Ufrgs (2022)

- Distância de Manhattan

A distância de Manhattan é a soma dos comprimentos da projeção da linha que combina eixos e coordenadas. Esse cálculo está representado pela equação 4 e Figura 5.

$$D(P,Q) = \sum_{i=1}^{n} |Pi - Qi| \quad (4)$$

Onde:

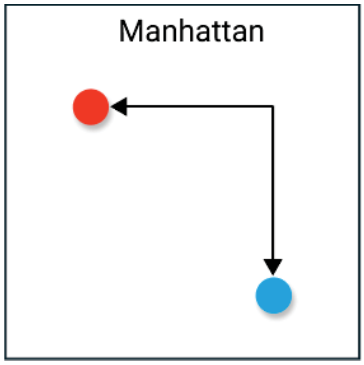
D: distância vetorial do ponto “P” ao ponto “Q”

n: número de dimensões no espaço vetorial

Pi: coordenado do ponto “P” na dimensão i

Qi: coordenado do ponto “Q” na dimensão i

Figura 5 - Distância Manhattan



Fonte: Ufrgs (2022)

- Distância de Chebyshev

  A distância de Chebyshev é definida pelo espaço vetorial em que a distância entre dois vetores é a maior de suas diferenças ao longo de qualquer dimensão de coordenada, representada pela equação 5 e Figura 6.

$$D(P,Q) = max_i(|Pi - Qi|) \quad (5)$$

Onde:

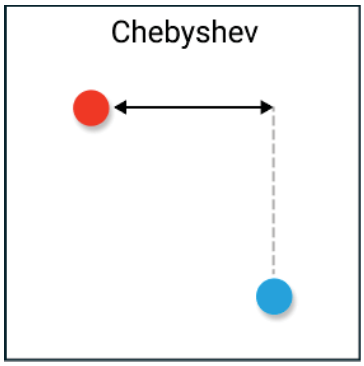
D: distância vetorial do ponto “P” ao ponto “Q”

n: número de dimensões no espaço vetorial

Pi: coordenado do ponto “P” na dimensão i

Qi: coordenado do ponto “Q” na dimensão i

Figura 6 - Distância Chebyshev



Fonte: Ufrgs (2022)

- Similaridade cossenoidal

  Trata-se de um método de calcular a similaridade de dois vetores pegando o produto escalar e dividindo-o pelas magnitudes de cada vetor, conforme mostrado equação 6 e Figura 7.

$$S(A,B) = \cos(\theta) = \frac{A * B}{\|A\|\|B\|} \tag{6}$$

  Onde:

  D: distância vetorial do ponto “A” ao ponto “B”

  $\|A\|$: comprimento do vetor “A” calculado pela equação 7

  $\|B\|$: comprimento do vetor “B” calculado pela equação 7

  θ: ângulo de abertura entre os vetores no plano cartesiano

$$\|A\| = \sqrt{x_1 + x_2 + \cdots + x_n} \tag{7}$$

  Onde:

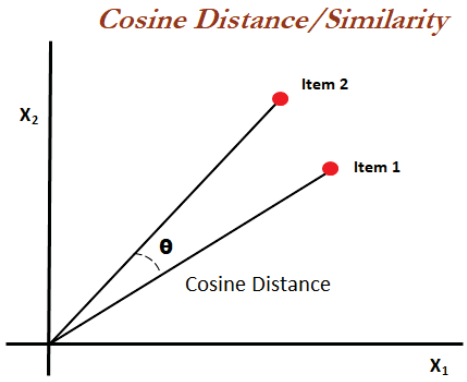
  $\|A\|$: comprimento do vetor “A”

  x_1: coordenada do vetor “A” na dimensão 1

  x_n: coordenada do vetor “A” na dimensão n

  n: número de dimensões no espaço vetorial

Figura 7 - Similaridade cossenoidal



Fonte: Oreilly (2024)

A seguir é apresentado um exemplo de cálculo de similaridade no qual são aplicados os conceitos matemáticos apresentados ao longo deste capítulo.

Para calcular a similaridade de cosseno, é importante normalizar os vetores dos documentos dividindo cada componente do vetor pelo seu comprimento Euclidiano, garantindo que possuam magnitude unitária.

Exemplo:
Considerando que a etapa de vetorização textual já foi realizada temos:
Vetor de TF-IDF do Documento 1: [2, 3, 1]
Vetor de TF-IDF do Documento 2: [1, 4, 2]

Em seguida é realizada a normalização dos vetores
Vetor de normalizado do Documento 1: [0,534, 0,803, 0,267]
Vetor de normalizado do Documento 2: [0,218, 0,872, 0,436]

A similaridade de cosseno, que mede o ângulo entre os vetores de representação dos documentos.

Multiplicamos os componentes correspondentes dos dois vetores normalizados e somamos esses produtos:
(0.534 * 0.218) + (0.803 * 0.872) + (0.267 * 0.436) ≈ 0.931

Em seguida, calculamos o produto dos comprimentos dos vetores normalizados:
sqrt((0.534^2 + 0.803^2 + 0.267^2) * (0.218^2 + 0.872^2 + 0.436^2)) ≈ 1.264

Utilizando a equação 6, temos que a similaridade por cosseno entre os vetores do exemplo é:

$$S(A,B) = \frac{0.931}{1.264} = 0.735 = 73{,}5\%$$

2.4 Considerações Finais do Capítulo

Na seção 2.1 deste capítulo, torna-se evidente a complexidade dos estudos elétricos incorporados nas diretrizes dos projetos básicos de linhas de transmissão estabelecidos pelo Operador Nacional do Sistema Elétrico (ONS). A interação dinâmica entre variáveis como carga, impedância, e a natureza variável das fontes de geração de energia requer uma análise minuciosa para garantir a estabilidade e a eficiência do sistema elétrico

A dificuldade enfrentada por profissionais na busca por similaridades nesses documentos é uma consequência da extensão e especificidade do conteúdo técnico, bem como a grande quantidade de dados, terminologias específicas e a natureza altamente técnica dos documentos tornam a identificação de padrões e a comparação manual uma tarefa complexa e propensa a erros. Diante desse cenário, destaca-se a necessidade crescente de ferramentas avançadas de processamento de linguagem natural e análise de dados, que possam aliviar o trabalho humano e proporcionar análises mais rápidas e precisas.

A aplicação da busca por similaridade com o Processamento de Linguagem Natural (NLP) no desenvolvimento de relatórios de estudos elétricos busca utilizar a vasta quantidade de relatórios já desenvolvidos e armazenados no banco de dados da Transmissora, o que representa um recurso valioso para aproveitar o conhecimento acumulado. Ao aplicar a busca por similaridade, é possível identificar o relatório mais semelhante ao novo empreendimento em termos de características técnicas, topologia de rede elétrica e outras especificidades relevantes.

Utilizar um relatório similar como base para o novo estudo, torna possível agilizar o processo de desenvolvimento, uma vez que o relatório escolhido como referência pode fornecer uma estrutura sólida e adequada para o novo projeto, facilitando a organização das informações e a inclusão dos requisitos necessários.

Por fim, a aplicação da busca por similaridade com NLP promove a padronização e a conformidade com as diretrizes estabelecidas pelo ONS. Ao utilizar relatórios anteriores que foram desenvolvidos de acordo com essas diretrizes como base, há uma garantia de consistência e alinhamento com as práticas e normas exigidas para empreendimentos de transmissão, o que reduz a margem de erro e aumenta a confiabilidade dos estudos elétricos, proporcionando um embasamento sólido para as decisões técnicas e estratégicas.

3 REVISÃO DA LITERATURA

3.1 Processo de seleção de portfólio bibliográfico

O processo de seleção do portfólio de revisão da literatura é baseado na metodologia proposta por Ensslin et al. (2010). Com base no objetivo de estudo do presente trabalho, foram selecionadas palavras-chave, sendo este o primeiro passo do processo. As palavras-chave foram divididas de acordo com o eixo de pesquisa, sendo eles aplicação e técnica. As combinações das palavras-chave esta apresentada na Tabela 3.

Tabela 3 – Eixos de pesquisa e palavras-chave para revisão da literatura

APLICAÇÃO	**TÉCNICA**
ELECTRICAL **POWER SYSTEMS**	Similarity Matching
	Text vetorization
	Cossine similarity

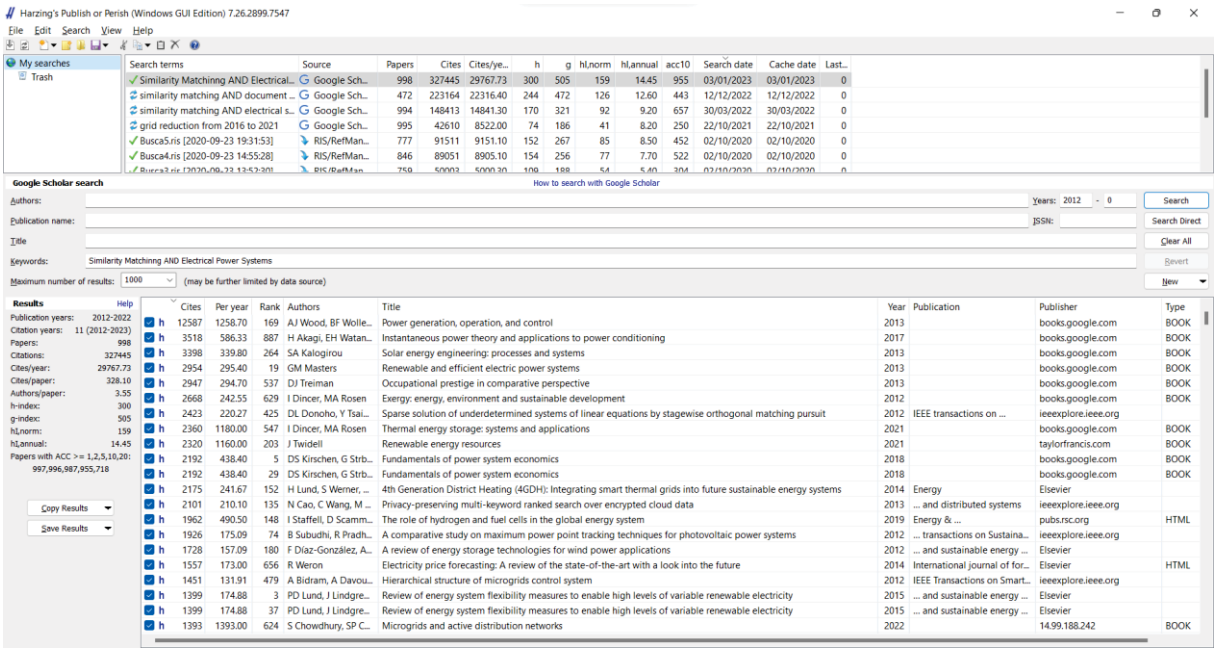
Fonte: O Autor (2022).

A definição da palavra-chave de aplicação teve como base a aplicação desejada das metodologias e técnicas voltadas para sistemas elétricos de potência, uma vez que o presente trabalho tem como foco a utilização de técnicas de inteligência artificial e computação em empreendimentos de linhas de transmissão. Já para as palavras-chave de técnica foi realizada uma análise pontual dos termos mais utilizados na literatura de forma que fosse possível afunilar o campo de busca, evitando assim palavras muito abrangentes que abordam diferentes aplicações, como por exemplo o termo “NLP” ou “inteligência artificial”.

Em seguida foi realizada a busca por artigos e trabalhos publicados nos últimos 10 anos, relacionados ao tema desse trabalho. Devido à grande variedade de trabalhos publicados e a facilidade de acesso a eles, a etapa de revisão de literatura torna-se um desafio pela quantidade de resultados encontrados. Dessa forma, o *software* utilizado para a busca pelos trabalhos foi o Publish or Perish, o qual consegue fazer o processo de busca baseado em relevância.

A base de dados selecionada na ferramenta foi o Google Scholar, o qual faz a busca das combinações das palavras-chave, listando os mil trabalhos mais citados, como mostra a Figura 8.

Figura 8 - Interface Publish or Perish



Em seguida é realizada uma filtragem nos resultados encontrados, analisando a Editora do trabalho e deixando apenas as fontes mais relevantes, como Elsevier, IEEE, Springer e Taylor & Francis. A lista de resultados reduzida pode ser salva no formato ".ris", possibilitando assim utilizá-la no *software* Mendeley, o qual consiste em um gerenciador de referências.

Adicionando os resultados a uma pasta no Mendeley, conforme mostra a Figura 9, para assim facilitar as citações futuras, foi possível realizar a segunda filtragem, a qual remove as duplicatas dos trabalhos já adicionados anteriormente.

Todo esse processo foi repetido para todas as combinações de palavras-chave, resultando na quantidade de artigos selecionados demostrada na Tabela 4.

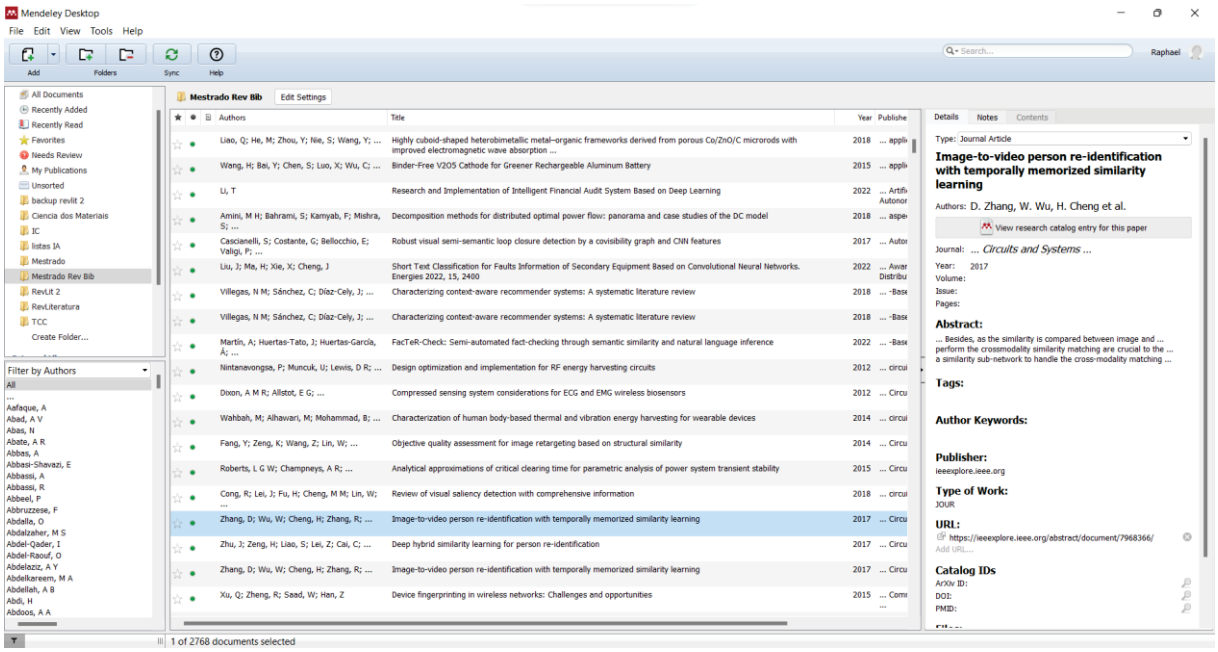
Tabela 4 - Número de trabalhos encontrados

Palavras-chave		**Antes da filtragem**		**Depois da filtragem**	
Aplicação	**Técnica**	Nº papers	Nº citações	Nº papers	Nº citações
Electrical Power Systems	Similarity Matching	997	663176	782	531168
	Text vetorization	200	1768	119	1185
	Cossine similarity	999	133179	639	64772

FONTE: O Autor (2022).

É possível notar que o total de *papers* após a filtragem foi de 2196, e retirando os títulos duplicados, este número cai para 1540 artigos.

Figura 9 - Interface do software Mendeley com referências carregadas



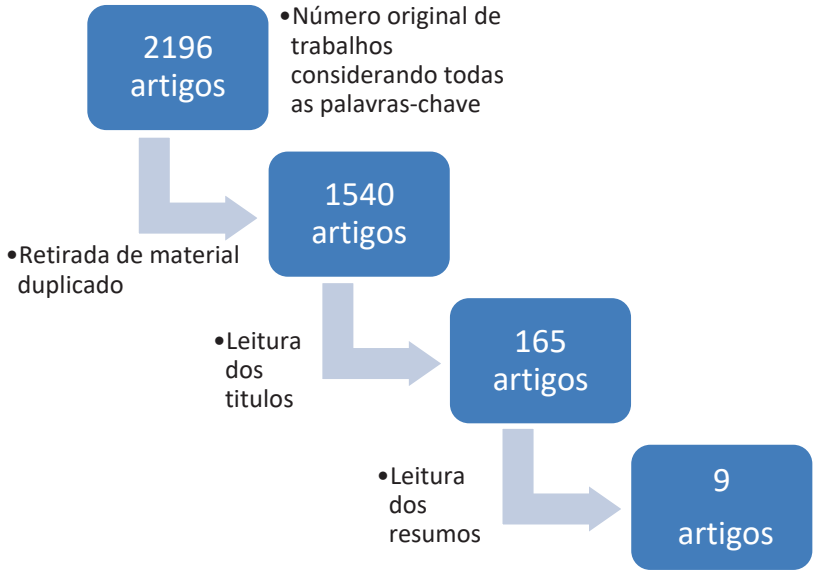
Fonte: O Autor (2022)

O próximo passo corresponde a leitura de cada um dos títulos, selecionando apenas aqueles que estão diretamente alinhados com o presente trabalho. Após essa etapa foi possível reduzir a quantidade de trabalhos para 165 títulos.

Para finalizar a etapa de filtragem de trabalhos que serão analisados, foi feita a leitura do resumo de cada um dos artigos, mantendo apenas aqueles que apresentam aspectos importantes de observação. Com isso, foram selecionados nove artigos para a serem analisados por completo.

O processo completo de busca de trabalhos para a revisão da literatura pode ser representado pelo fluxo da Figura 10.

Figura 10 - Processo de seleção dos trabalhos para a revisão bibliográfica



Fonte: O Autor (2022)

3.2 Análise do portfólio bibliográfico

Segundo Amin et al. (2019) três áreas de pesquisa com alta relevância para aplicação de técnicas de gestão de conhecimento são o processamento automático de texto difuso, extração de contexto e desambiguação. O conhecimento empresarial pode ser encontrado em trocas de mensagens de texto simples, e-mails, tickets de suporte, e outros meios voláteis, tornando a decodificação do conhecimento baseado em texto, uma tarefa desafiadora. O autor argumenta que as abordagens tradicionais de processamento de linguagem natural se concentram em uma representação abrangente do conhecimento empresarial e em quaisquer mapeamentos relevantes, mas reconhece que essas abordagens podem ser altamente complexas, não serem rentáveis e requererem manutenção frequente, especialmente em ambientes que experimentam mudanças frequentes. Diante disso, o trabalho apresenta sua proposta de aplicar redes LSTM Siamesas para medir similaridades de texto e implementar a rede neural Siamesa Manhattan LSTM (MaLSTM) para aquisição de conhecimento semiautomático de conhecimento empresarial e decodificação de recursos relevantes que permitem a construção de medidas de similaridade.

Liu et al. (2019) apresentam um método para aproveitar as informações de referência do histórico de bilhetes de manutenção na rede elétrica, afim de criar novos bilhetes de trabalho e realizar suporte de decisão multidirecional. O método inclui um

pré-processamento do texto com as informações de medidas de segurança da rede elétrica e o conhecimento do campo de energia elétrica. O método utiliza um modelo aprimorado de saco de palavras (BOW, do inglês, *bag of* words) com palavras principais e auxiliares, além do cálculo de frequência do termo chamado de frequência inversa do documento (TF-IDF), e o método de similaridade de cosseno com o objetivo de calcular a similaridade multivariável entre as informações críticas do equipamento de manutenção e as cenas históricas. Segundo os autores, o método proposto elimina os problemas de inversão de ordem das palavras e o significado de várias palavras, melhorando a eficiência e a precisão da correspondência. A eficácia do método proposto foi validada em vários casos com base em uma rede elétrica real.

Lian et al. (2022) apresentam um método capaz de integrar as informações do processo de gestão e fabricação de equipamentos elétricos e resumir dados estatísticos desde o estabelecimento do projeto, construção e produção, até a operação do equipamento, além de relacionar as informações do projeto com as informações do dispositivo. O método integra e resume os dados de várias fontes do sistema, conectando e correspondendo as informações relevantes, permitindo assim a realização da conexão do arquivo do projeto e do equipamento, e fornecendo a base para as estatísticas do processo. De acordo com o autor, o resultado de correspondência é mais rápido e eficiente utilizando similaridade cossenoidal, o esforço de trabalho é reduzido e a eficiência e precisão da correspondência pôde ser melhorada.

McNamee (2013) trata como as medidas de distância ou similaridade computacional frequentemente ignoram informações importantes contidas nas taxonomias de classificação que são utilizadas como base para essas medidas. Para resolver esse problema, o autor apresenta duas modificações que permitem que diversos métodos de pesquisa em gestão utilizem plenamente os dados de classificação hierárquica utilizando um exemplo detalhado que explora diversas medidas de similaridade tecnológica com base no sistema de classificação de patentes do USPTO (Escritório Americano de Patentes). Ele destaca que a metodologia de taxonomia oferece benefícios específicos para o contexto de patentes, como a capacidade de modelar o espaço tecnológico do documento dentro de campos específicos e a possibilidade de analisar com precisão a similaridade no nível de patente-a-patente. O autor testa o desempenho desses métodos em dois estudos diferentes, um em um nível de patente-a-patente dentro de um único campo

tecnológico e outro em um nível de organização-a-organização em diferentes indústrias chegando em resultados que mostram que os métodos taxonômicos geram distribuições mais significativas de pontuações de similaridade em ambas as amostras e as pontuações de similaridade calculadas por meio de métodos taxonômicos têm uma relação mais consistente com a probabilidade e o número de citações.

Lu e Ye (2013) propõem uma nova medida de similaridade do cosseno entre conjuntos vagos e aplicada no diagnóstico de falhas em turbinas. Para isso, foi avaliado um novo valor de medida de similaridade entre uma amostra de teste e o conhecimento prévio sobre falhas do sistema através de diagnóstico de falhas por vibração em turbinas. De acordo com os autores, a amostra de teste é considerada próxima a um tipo de falha conhecida se o valor da medida for alto, determinando o tipo de falha por vibração de acordo com o valor máximo da medida (maior que um limite). Os problemas de diagnóstico de falhas da turbina foram investigados utilizando a metodologia proposta de similaridade do cosseno, de forma que os resultados demonstraram que o método proposto não apenas diagnostica os principais tipos de falhas da turbina, mas também fornece informações úteis para análises de múltiplas falhas e tendências futuras. Dessa forma, segundo o autor, o método proposto é razoável e eficaz e oferece outra ferramenta útil para análises de falhas.

Segundo Ni et al. (2022), os engenheiros enfrentam um desafio em encontrar soluções de projetos inovadores em diferentes domínios da engenharia industrial devido à crescente demanda por novos produtos. Os documentos de patentes são ricos em conhecimento inventivo e os autores pressupõem que problemas de engenharia podem ter soluções práticas em outros domínios científicos, quando descritos de maneira semelhante. Para determinar a similaridade entre problemas de patentes, é proposta a aplicação de técnicas de aprendizado de máquina, redes neurais integrada a uma rede neural bidirecional LSTM treinada, chamada Manhattan LSTM (redes de memória de longo curto prazo, do inglês Long Short Term Memory networks), em uma abordagem chamada SAM-IDM (abordagem baseada em similaridade para fusão do método de projeto inventivo). A abordagem é então experimentada em um conjunto de dados de patentes reais dos Estados Unidos, apresentando resultados promissores em termos de correspondência de similaridade de frases e inventividade.

Arts et al. (2021) desenvolveram técnicas de processamento de linguagem natural para identificar a criação e o impacto de novas tecnologias nas patentes dos

EUA. Essas técnicas foram validadas e comparadas com métricas tradicionais com base na classificação e citações de patentes em dois estudos de caso-controle. No primeiro estudo, foram recolhidas patentes que receberam prêmios, como o Prêmio Nobel e o National Inventor Hall of Fame, que provavelmente abrangem tecnologias radicalmente novas com um grande impacto no progresso tecnológico e patenteamento. No segundo estudo, foram identificadas patentes concedidas pelo Instituto de Patentes e Marcas dos Estados Unidos, mas que foram rejeitadas tanto pelo instituto de patentes europeu como pelo japonês, indicando que essas patentes não apresentam grande novidade ou representam pequenos avanços incrementais sobre técnicas anteriores, tendo pouco impacto no progresso tecnológico.

Li e Wen (2014) apresentam uma estratégia de detecção de falhas da unidade híbrida de tratamento de ar (AHU) baseada no método de Análise de Componentes Principais (PCA) e no método de Correspondência de Padrões. O método de Correspondência de Padrões busca localizar períodos de operação similares a partir de um conjunto de dados históricos, com o objetivo de caracterizar o grau de semelhança entre a janela de dados históricos e a janela de dados atuais. O método proposto de Correspondência de Padrões - PCA utiliza dois fatores de semelhança, fatores de semelhança PCA e fatores de semelhança de distância, para identificar os dados históricos da operação da AHU que são similares aos dados atuais da operação de *snapshot*. O método foi validado pelos dados operacionais de uma AHU em um edifício real e os resultados mostram que a sensibilidade dos modelos PCA é reforçada pelo pré-processamento dos dados de formação com o Método de Correspondência de Padrões.

Segundo Akmal et al. (2014), atualmente o desenvolvimento de produtos está se tornando cada vez mais intensivo em conhecimento e as equipes de design enfrentam desafios consideráveis na utilização eficaz de quantidades crescentes de informação. Para apoiar a recuperação e reutilização de informações sobre produtos, o autor sugere que é possível utilizar o raciocínio baseado em casos (do inglês, Case Based Reasoning - CBR), no qual problemas são resolvidos através da utilização ou adaptação de soluções de problemas antigos. Para identificar casos que são mais relevantes para o problema a ser resolvido, o raciocínio baseado em casos utiliza medidas de semelhança, no entanto, a maioria das medidas de semelhança não numéricas baseia-se em fundamentos sintáticos, que muitas vezes não produzem bons resultados quando confrontados com o significado associado às palavras que

comparam. Para superar essa limitação, o trabalho faz uso das ontologias, as quais podem ser utilizadas para produzir medidas de similaridade baseadas na semântica. O artigo apresenta uma abordagem baseada em ontologias que pode determinar a semelhança entre duas classes utilizando medidas de semelhança baseadas em características que substituem características por atributos. A abordagem proposta é avaliada em relação a outras semelhanças existentes e a eficácia da abordagem é ilustrada com um estudo de caso sobre problemas de concepção de um produto.

3.3 Considerações finais do capítulo

Com base na busca de trabalhos acadêmicos realizada neste capítulo, fica clara a inovação da metodologia proposta no presente trabalho, de forma que a análise apresentada na Tabela 5 demonstra que nenhum dos trabalhos identificados como mais relevantes para serem utilizados como referência possuem uma aplicação para a elaboração de relatórios de estudos elétricos.

Tabela 5 - Tabela comparativa da revisão da literatura

Critérios	Amin et al. (2019)	Liu et al. (2019)	Lian et al. (2022)	McNamee (2013)	Lu; Ye (2013)	Ni et al. (2022)	Arts et al. (2021)	Li; Wen (2014)	Akmal et al. (2014)	Presente trabalho
Gestão de conhecimentos prévios	X	X	X			X	X	X	X	X
Busca por similaridade	X	X	X	X	X	X		X	X	X
Processamento natural de linguagem (NLP)	X	X	X	X	X	X	X		X	X
Similaridade por cosseno		X	X	X	X					X
Aplicação em engenharia elétrica		X	X	X	X	X		X		X
Aplicação em estudos elétricos										X

Fonte: O Autor (2022)

Percebe-se que a maioria dos casos de aplicação da busca por similaridade são vinculados aos casos de manutenção industrial e detecção de falhas de equipamentos, o que torna a aplicação da metodologia para a etapa de documentação de projetos uma aplicação não tanto difundida.

Apesar das diferentes áreas que a busca por similaridade pode ser aplicada, os conceitos computacionais e matemáticos são iguais, o que torna a etapa de revisão da literatura imprescindível para o desenvolvimento do presente trabalho.

Por fim, a partir da análise dos trabalhos, foi possível perceber que a técnica de busca por similaridade por cosseno tem aplicação atual em diferentes ramos da pesquisa e se mostrou útil e validada para estudos que envolvem a gestão de conhecimentos prévios.

4 MATERIAL E MÉTODOS

4.1 MATERIAL

4.1.1 Python

A linguagem de programação chamada de Python teve sua origem no começo dos anos 1990, desenvolvida pelo matemático holandês Guido van Rossun. O modelo da linguagem trata-se de desenvolvimento comunitário, sem fins lucrativos, com o objetivo de ser legível, de fácil manutenção e com suporte avançado a mecanismos de reutilização de software. (PEREIRA, 2021)

Em termos técnicos o Python considerado uma linguagem de programação de alto nível, dinâmica, modular, multiplataforma e orientada a objetos. por ser uma linguagem de sintaxe relativamente simples e de fácil compreensão tornou-se muito popular entre profissionais da indústria de tecnologia, engenheiros, matemáticos e pesquisadores.

Uma das principais vantagens em utilizar esta linguagem de programação está no grande número de bibliotecas disponíveis e validadas, o que a torna muito difundida e útil em uma grande variedade de setores. (KRIGER, 2022)

A seguir são apresentadas algumas bibliotecas que podem ser utilizadas para a busca por similaridade, tanto para o pré-processamento dos dados, quanto para o cálculo da similaridade entre relatórios. A utilização das bibliotecas partiu do princípio de validações constantes e análise das suas funções, de forma que foi necessário o estudo do funcionamento e processamento que ocorrem ao chamá-las no script de testes.

4.1.1.1 Bibliotecas

As bibliotecas listadas a seguir são aquelas que apresentam funções úteis para o presente trabalho e consequentemente foram utilizadas em algumas versões do código para comparação e validação. Dessa forma, cada uma delas apresenta uma breve apresentação de suas funções e, quando relevante, uma explicação de seu funcionamento.

- Glob

  A biblioteca Python "glob" é utilizada para realizar buscas de arquivos e diretórios com base em padrões de nome. Ela permite que você encontre arquivos que correspondam a um determinado padrão de nome, utilizando caracteres curinga, como asteriscos (*) e pontos de interrogação (?).

  Por exemplo, para encontrar todos os arquivos com extensão ".txt" em um diretório, pode usar o padrão "*.txt" ao chamar a função "glob". Ela retornará uma lista de caminhos para todos os arquivos que correspondam a esse padrão. (THE PYTHON SOFTWARE FOUNDATION, 2023)

- Re - Operações de expressão regular (Regular expression operations)

  A biblioteca Python "re" é utilizada para trabalhar com expressões regulares, as quais são sequências de caracteres que definem um padrão de busca. Com a biblioteca "re", é possível realizar operações como encontrar padrões em strings, substituir partes de uma string por outra, dividir strings com base em padrões e muito mais.

  A biblioteca "re" é utilizada para tarefas como validação de formatos de strings, extração de informações específicas de uma string e transformações de texto com base em padrões. (THE PYTHON SOFTWARE FOUNDATION, 2023b)

- Os

  Este módulo fornece uma maneira de usar a funcionalidade dependente do sistema operacional, podendo acessar e manipular recursos do sistema, como arquivos, diretórios, variáveis de ambiente, entre outros. É possível, por exemplo, ler ou escrever um arquivo usando “open()”, manipular caminhos, através do módulo “os.path”, e ler todas as linhas em todos os arquivos na linha de comando pelo módulo “fileinput” . Além disso é também possível criar arquivos e diretórios temporários e para manipular arquivos e diretórios de alto nível. (THE PYTHON SOFTWARE FOUNDATION, 2023c)

- Gensim

  Gensim é uma biblioteca Python para modelagem de tópicos, indexação de documentos e recuperação de similaridade, tendo como público-alvo a comunidade de processamento de linguagem natural (NLP) e recuperação de informações (IR, do inglês, information retrieval). (PyPI, 2023)

  A principal funcionalidade do "gensim" é permitir a criação e manipulação de representações vetoriais de documentos de texto. Essas representações são utilizadas para extrair informações relevantes dos documentos e realizar tarefas como a identificação de tópicos e a busca por documentos similares.

  - Corpora

    Este módulo implementa o conceito de dicionário – um mapeamento entre palavras e seus IDs inteiros. (Řehůřek, 2022a)

    A função principal é doc2bow, que converte uma coleção de palavras em sua representação de pacote de palavras: uma lista de (word_id, word_frequency).

    Em Gensim, o objeto de dicionário é utilizado para criar um pacote de palavras (BoW, do inglês, bag of words), que também é utilizado como input para a modelação de tópicos e outros modelos. O dicionário contém o mapeamento de todas as palavras, ou seja, os símbolos da sua identificação inteira única. Pode-se criar um dicionário a partir de uma lista de frases e a partir de um ou mais ficheiros de texto.

  - TfidfModel

    É o termo de Frequência Inversa de Documentos (em inglês, Frequency-Inverse Document Frequency) que é também um modelo de pacote de palavras. Se diferencia da frequência absoluta porque reduz o peso dos IDs, ou seja, das palavras que aparecem frequentemente nos documentos.

No momento da transformação, é necessária uma representação vetorial e é devolvida outra representação vetorial. O vector de saída terá a mesma dimensão, porém o peso das palavras raras será elevado. Em termos resumidos, este módulo converte vetores com valor inteiro em vetores com valor efetivo.

O modelo computacional que realiza esses cálculos segue dois passos simples:

- Multiplicação da componente local e global
  Neste primeiro passo, o modelo multiplicará um componente local como a TF (Frequência de Termo) com um componente global como a IDF (Frequência de Documento Inversa).
- Normalizar o resultado
  Uma vez feita a multiplicação, na próxima etapa o modelo TF-IDF normalizará o resultado em função do comprimento da unidade.

- Similarities

  Este módulo calcula a similaridade cossenoidal de uma verificação dinâmica em relação a um conjunto de documentos estáticos ("o índice"), ou seja, calcula as semelhanças entre uma coleção de documentos no Modelo Espacial Vetorial.

  A escalabilidade é adquirida dividindo o índice em partes menores, cada uma das quais encaixa na memória central. Os fragmentos em si são armazenados como ficheiros em disco e mapeados de volta, conforme necessário. (API, 2016)

  Os documentos são divididos (internamente, de forma transparente) em fragmentos de documentos de tamanho reduzido cada um, e cada fragmento é convertido numa matriz, para chamadas mais rápidas. Cada fragmento é armazenado em disco sob output_prefix.shard_number. (ŘEHŮŘEK, 2022b)

- Scikit-learn (sklearn)

  Trata-se de uma biblioteca de aprendizado de máquina e mineração de dados, a qual oferece uma ampla gama de algoritmos e ferramentas para lidar com tarefas de classificação, regressão, agrupamento, redução de dimensionalidade, pré-processamento de dados.

  Essa biblioteca possui ferramentas para avaliação de modelos, como métricas de desempenho, validação cruzada e seleção de modelos, além de oferecer recursos de pré-processamento de dados, como codificação de variáveis categóricas, normalização de dados, redução de dimensionalidade e seleção de recursos.

  - cosine_similarity

    A função "cosine_similarity" está integrada ao submódulo "pairwise" do pacote "metrics", sendo usada para calcular a similaridade cossenoidal entre dois conjuntos de vetores. Ao passar dois conjuntos de vetores como argumentos, cada conjunto de vetores pode ser representado como uma matriz bidimensional, onde cada linha representa um vetor. A função retornará uma matriz de similaridade, onde cada elemento [i, j] representa a similaridade do cosseno entre o vetor i do primeiro conjunto e o vetor j do segundo conjunto.

  - TfidfVectorizer

    Essa função que se encontra dentro do módulo "feature_extraction.text" é uma implementação do vetorizador TF-IDF usado para converter uma coleção de documentos de texto em uma matriz de recursos TF-IDF, onde cada documento é representado por um vetor que representa a importância relativa de cada termo dentro do documento e em toda a coleção.

- Pdfminer

  Com a biblioteca “pdfminer” é possível extrair texto, metadados e outras informações de documentos em formato PDF permitindo a análise e extração de conteúdo de arquivos PDF de forma programática. A biblioteca oferece diferentes níveis de abstração, desde a extração básica de texto até a análise mais detalhada da estrutura do PDF, sendo possível

personalizar a forma como o texto é extraído, como lidar com espaços em branco, quebras de linha, estilos e outros elementos do PDF.

- Pandas

  A biblioteca Pandas foi criada para a manipulação e análise de dados, e em particular, oferece estruturas de dados e operações para manipulação de tabelas numéricas e séries temporais. (RAVULAKOLLU, 2023)

  - DataFrame

    DataFrames armazenam dados no formato familiar de tabela de linhas e colunas, muito semelhante a uma folha de cálculo ou base de dados, facilitando muitas tarefas analíticas, tais como encontrar as médias por coluna num conjunto de dados.

    Também se pode pensar nessa estrutura como uma colecção de séries, como múltiplas colunas combinadas formam uma tabela, múltiplas séries formam um DataFrame.

    Pandas DataFrame é uma estrutura de dados bidimensional, de tamanho mutável, com estrutura de dados tabulares, com eixos etiquetados (linhas e colunas), que consiste em três componentes principais, sendo eles dados, linhas e colunas. (MODE, 2022)

- Matplotlib.pyplot

  Matplotlib é uma biblioteca de extensão numérica de interface baseada em estado para um módulo que fornece uma estrutura do tipo MATLAB. Há várias plotagens que podem ser usadas em Pyplot como Line Plot, Contour, Histogram, Scatter, 3D Plot, etc.

  O API explícito (orientado para objetos) é recomendado para gráficos complexos, embora o pyplot ainda seja normalmente utilizado para criar a figura e muitas vezes os eixos da figura. (MATPLOTLIB, 2023)

- Scipy.cluster

  - Hierarchy

    O módulo Hierarchy fornece funções para agrupamento hierárquico e de aglomeração. As suas características incluem a geração de clusters hierárquicos a partir de matrizes de distância,

> cálculo de estatísticas sobre clusters, corte de ligações para gerar clusters planos, e visualização de clusters com dendrogramas. (THE SCIPY, 2023)

4.1.2 Relatórios Estudos elétricos

Os relatórios de projetos básicos de linhas de transmissão são documentos técnicos fundamentais para o planejamento, projeto, construção e manutenção de sistemas elétricos de transmissão de energia, uma vez que fornecem informações detalhadas sobre o trajeto da linha de transmissão, o tipo de terreno e a distância entre as torres, o tipo de material utilizado para construir as torres e a fiação, bem como os equipamentos necessários para operar e manter a linha.

Os relatórios de projetos básicos de linhas de transmissão enfrentam dificuldades durante sua elaboração, uma vez que é necessário equilibrar a capacidade de transmissão e a qualidade da energia com os custos de construção e manutenção da linha, levando em conta a topografia e as condições climáticas da região onde a linha será instalada.

A elaboração dos relatórios de projetos básicos de linhas de transmissão considera as particularidades das redes elétricas regionais e a complexidade das interconexões entre as linhas de transmissão, sendo documentos técnicos imprescindíveis para garantir a eficiência e segurança na construção e operação de linhas de transmissão de energia elétrica. Dessa forma é importante que os profissionais responsáveis pela elaboração desses relatórios possuam expertise técnica e experiência prática na área, a fim de garantir a confiabilidade e a qualidade desses documentos.

Nos relatórios estão contemplados os estudos elétricos apresentados na fundamentação teórica, podendo seguir diferentes classificações de acordo com os equipamentos abordados em cada um dos documentos, seguindo as seguintes categorizações:

- Linha de transmissão
    - Energização
    - Religamento monopolar e tripolar
    - Extinção de arco
    - Rejeição de carga

- Disjuntor
    - Tensão de Restabelecimento Transitória (TRT)

- Chaves
    - Correntes induzidas em lâmina de terra
- Transformador
    - Energização
- Banco de Capacitor
    - Energização

Considerando que o presente trabalho faz parte de um projeto de Pesquisa e Desenvolvimento desenvolvido em conjunto com a COPEL, foi possível ter acesso a um conjunto de 72 documentos de relatórios de estudos elétricos, os quais, apesar de suas particularidades de acordo com o tipo de estudo, possuem a seguinte estrutura padrão:

- Introdução e objetivos
- Critérios
- Base de dados
- Configuração/Casos analisados
- Resultados
- Conclusões

Com base nessa estrutura, a documentação dos relatórios é preenchida com os dados referentes a cada empreendimento referente ao estudo elétrico, seguindo as diretrizes estabelecidas pelo ONS.

4.2 MÉTODOS

4.2.1 Metodologia de desenvolvimento da estratégia de busca por similaridade

O processo de desenvolvimento da ferramenta incorporou diversas etapas de aprendizado e desenvolvimento técnico, tanto na área da engenharia necessária para

a realização dos estudos presentes nos relatórios quanto no desenvolvimento computacional e suas particularidades para a implementação de uma ferramenta que utiliza princípios da semântica textual nos documentos.

Dessa forma, o fluxo de estudo e aplicação dos conhecimentos seguiu um caminho de construção de conceitos, iniciando pela compreensão dos estudos elétricos que compõem os relatórios e que são exigidos pelo ONS. Em seguida foram iniciados os estudos voltados para as técnicas computacionais necessárias para a implementação da busca por similaridade, o que abrange o aprendizado da linguagem de programação mais adequada, e os conceitos matemáticos necessários para a composição dos métodos de NLP já existentes.

O processo da elaboração da metodologia foi desenvolvido após a compreensão dos conceitos isolados, e uma visão clara do objetivo do presente trabalho, iniciando assim a aplicação prática das técnicas estudadas.

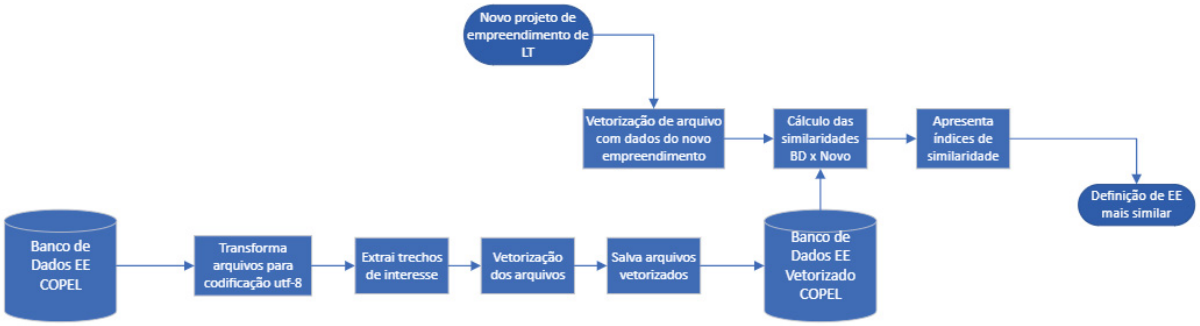
A necessidade de testes iniciais na programação surgiu assim que a operação se tornou possível, utilizando diferentes documentos e bancos de dados para a realização. Parte dos documentos utilizados para os testes foi disponibilizado pela COPEL, de forma gradual, uma vez que os primeiros testes contaram com poucos documentos, e assim progredindo com o decorrer das atividades.

A etapa de validação dos resultados se deu através de conversas com os profissionais da COPEL e análises manuais dos documentos.

4.2.2 Fluxograma geral

A metodologia do processo de busca por similaridade de relatórios de estudos elétrico desenvolvidos pela COPEL é apresentada na Figura 11.

Figura 11 - Fluxograma da metodologia de busca por similaridade



Fonte: o autor (2024).

O processo inicia com as informações de um novo empreendimento de linha de transmissão, com informações como localização geográfica, comprimento da linha, nível de tensão, entre outros. O arquivo com essas informações deve ser incluído no diretório onde se localizam os demais relatórios já gerados, a fim de fazer a busca comparativa entre eles.

Uma vez que o script é executado, ocorre a vetorização do documento com as informações do novo empreendimento, além de realizar a vetorização dos demais documentos da pasta, a fim de criar um banco de dados somente de documentos já vetorizados.

Por fim é realizado o cálculo das similaridades entre os documentos antigos e a nova referência, a qual contém as informações do novo empreendimento. Nessa etapa são apresentados os índices de similaridade e consequentemente a definição do estudo elétrico já realizado mais similar.

5 TESTES E ANÁLISE DOS RESULTADOS

5.1 Caso teste

Para a implementação do cálculo de similaridade entre documentos textuais, foi desenvolvido um software em linguagem Python, no qual são implementadas as metodologias matemáticas e processuais de pré-processamento e cálculo da similaridade entre textos.

A fim de realizar testes de desempenho do programa, foram consideradas seis frases simples, nas quais buscou-se saber previamente quais os níveis de similaridade entre elas, ou seja, criadas propositalmente semelhantes ou distintas com o objetivo de validação da compreensão semântica da metodologia implementada.

A seguir estão descritas as seis frases:

A - Hoje de manhã fez sol mas amanhã pode chover
B - Hoje de manhã fez chuva mas amanhã pode fazer sol
C - O carro está molhado porque choveu de manhã
D - Hoje de manhã fez sol mas amanhã pode não fazer sol
E - O carro está na rua
F - Tem um prédio novo na rua

Tendo cada uma dessas frases armazenadas em documentos de texto no formato .txt com codificação UTF-8, localizados na pasta denominada “datasets_2”, o primeiro passo do software realiza a obtenção desses documentos para dentro do ambiente de processamento, como mostra a Figura 12.

Figura 12 - Seção software de similaridade - importa documentos

```
# Importa biblioteca
import glob

# Pasta com documentos
folder = "datasets_2/"

# Lista dos documentos .txt ordenados
files = glob.glob(folder + "*.txt")
files.sort()
```

Fonte: o autor (2024).

A Figura 13 mostra que após ter acesso aos documentos, inicia-se a etapa de pré-processamento dos textos. São criados vetores de armazenamento de conteúdo e título, sendo o título definido como o nome do arquivo em que o texto está salvo. Em seguida, é feita a análise dos caracteres dos textos, onde são retirados aqueles que não se tratam de caracteres alfanuméricos. Uma vez feita essa seleção, os textos e títulos são armazenados nos seus respectivos vetores, e adicionalmente identifica-se o índice do documento que será a referência do cálculo de similaridade, no caso, o documento “A”.

Figura 13 - Seção software de similaridade - acesso aos textos e remoção de caracteres não alfanuméricos

```
# Importa bibliotecas
import re, os

# Initializa objeto que contem textos e titulos
txts = []
titles = []

for n in files:
    # Open each file
    f = open(n, encoding='utf-8-sig')
    # Remove caracteres não-alpha-numericos
    data = re.sub('[\W_]+', ' ', f.read())
    # Store the texts and titles of the books in two separate lists
    txts.append(data)
    titles.append(os.path.basename(n).replace('.txt', ''))

# Lista de todos os titulos
for i in range(len(titles)):
    # Guarda o index do primeiro documento
    if titles[i] == 'A':
        ori = i
```

Fonte: o autor (2024).

Uma das etapas do pré-processamento trata-se da remoção de palavras pouco significativas do texto, chamadas de “stop-words”. Com isso o programa possui uma lista de palavras pré definida, a qual contém termos que não foram considerados como essenciais para a estrutura de significado dos textos. Assim, para que se possa fazer uma busca adequada desses termos que estão escritas em letras minúsculas na lista, todo o texto também é convertido para letras minúsculas e cada palavra separada na forma de um token, e dessa forma sendo possível fazer a análise de busca individual de cada palavra pouco significativa, como apresentado na Figura 14.

Figura 14 - Seção software de similaridade - remoção das palavras pouco significativas

```
# Define lista de 'stop-words'
stoplist = set('para a de o e to em no na é qual algum são que nós eu que pode nosso como'.split())

# Converte texto para ministulo
txts_lower_case = [i.lower() for i in txts]

# Transforma texto em tokens
txts_split = [i.split() for i in txts]

# Remove tokens que fazem parte da lista de 'stop-words'
texts = [[word for word in txt if word not in stoplist] for txt in txts_split]
```

Fonte: o autor (2024).

Para iniciar a vetorização dos documentos, é necessária a criação do conceito de dicionário, onde são armazenadas todas as palavras existentes no texto, através do modelo de pacote de palavras (bag of words), o qual para cada termo é associado um valor que representa a quantidade de vezes que a palavra se repete. Esse processo está descrito pela Figura 15.

Figura 15 - Seção software de similaridade - método de pacote de palavras

```
# Importa bibliotecas
from gensim import corpora

# Cria dicionario para tokens
dictionary = corpora.Dictionary(texts)

# Cria modelo de pacote de palavras (bag-of-words) para cada documento, usando o dicionário criado
bows = [dictionary.doc2bow(text2) for text2 in texts]
```

Fonte: o autor (2024).

Com o dicionário pronto, é possível utilizar o método de frequência do termo-inverso da frequência nos documentos para assim ter o modelo vetorizado de cada um dos documentos em análise, apresentado na Figura 16.

Figura 16 - Seção software de similaridade - modelo TF-IDF

```
# Carrega função gensim para gerar modelo tf-idf
from gensim.models import TfidfModel

# Gera o modelo tf-idf
model = TfidfModel(bows)
```

Fonte: o autor (2024).

Agora, já possuindo os textos vetorizados, finaliza-se a etapa de pré-processamento de dados, e é possível realizar o cálculo da similaridade entre os mesmos. A utilização da biblioteca calcula a semelhança cossenoidal em relação a

um conjunto de documentos, armazenando a matriz de índices na memória, como apontado na Figura 17

Figura 17 - Seção software de similaridade - cálculo da semelhança cossenoidal

```
# Importa biblioteca
import pandas as pd
# Carrega biblioteca para cálculo de similaridade
from gensim import similarities

# Cálculo da matriz de similaridade
sims = similarities.MatrixSimilarity(model[bows])

# Tranforma resultado em dataframe
sim_df = pd.DataFrame(list(sims))

# Adiciona os titulos dos documentos ao dataframe
sim_df.columns = titles
sim_df.index = titles

# Print the resulting matrix
sim_df
```

Fonte: o autor (2024).

Com o cálculo da similaridade finalizado, são ordenados os documentos, do mais semelhante ao menos semelhante, utilizando os valores dos ínices com relação ao documento selecionado como referência no início do software, no caso o documento “A”. Para a apresentação dos resultados, foi escolhida a plotagem de um gráfico de barras, no qual a similaridade é indicada através dos índices individuais dos textos com relação à referência, como destacado na Figura 18.

Figura 18 - Seção software de similaridade - organização e apresentação dos resultados

```
# Importa biblioteca
import matplotlib.pyplot as plt

# Seleciona a coluna correspondente ao primeiro documento
v = sim_df['A']

# Organiza pelos valores do scores
v_sorted = v.sort_values()

# Plota grafico de barra horizontal
v_sorted.plot.barh()

# nomeia eixo x
plt.xlabel('Similarity')
```

Fonte: o autor (2024).

Finalmente, com a disposição dos dados de similaridade de todos os textos entre si, é calculado e plotado o dendrograma, o qual apresenta um diagrama de

árvore que são exibidos os grupos formados por agrupamentos de observações em cada passo e seus níveis de similaridade, realizado como mostra a Figura 19.

Figura 19 - Seção software de similaridade - plotagem do dendrograma

```
# Importa biblioteca
from scipy.cluster import hierarchy

# Calcula os clusters das matrizes de similaridade
Z = hierarchy.linkage(sim_df, 'ward')

# Apresenta resultado através de dendograma horizontal
plt.figure()
dn = hierarchy.dendrogram(Z, leaf_font_size=8, orientation="left")
```

Fonte: o autor (2024).

5.2 Análise dos resultados – Caso Teste

O software de cálculo de similaridade entre texto apresenta três tipos de resultados, os quais, a partir da análise individual, é possível chegar a diferentes conclusões.

Primeiramente é gerada a matriz de índices de similaridade pareada, ou seja, qual o valor da similaridade, entre 0 e 1, sendo 1 o próprio documento e 0 nenhuma semelhança, entre cada um dos documentos em análise.

A Figura 20 apresenta esta matriz gerada a partir do caso teste.

Figura 20 - Matriz de similaridade

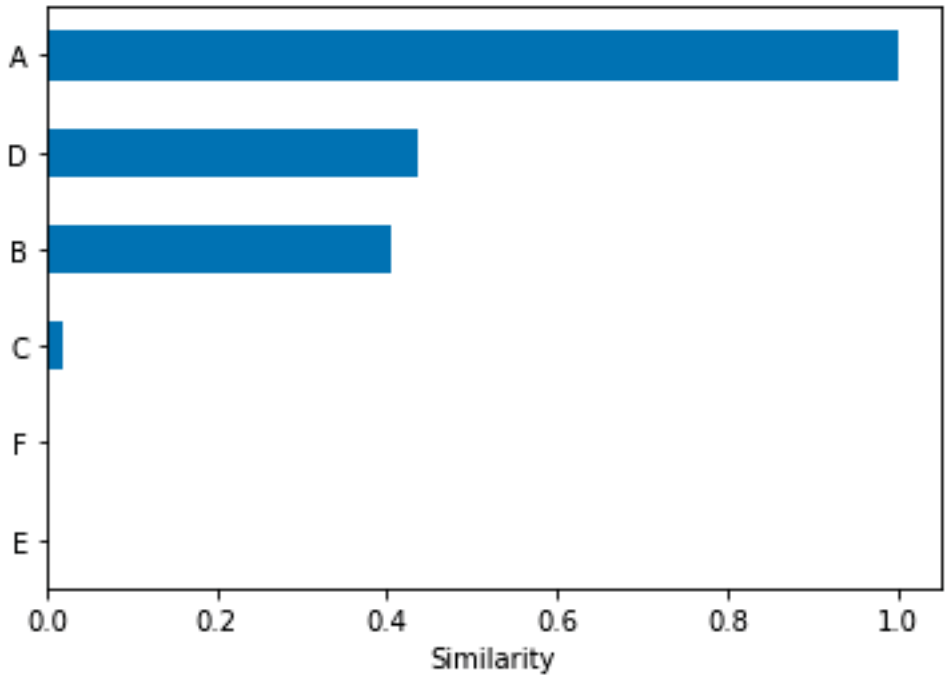
	A	B	C	D	E	F
A	1.000000	0.404076	0.019575	0.436758	0.000000	0.000000
B	0.404076	1.000000	0.017803	0.554571	0.000000	0.000000
C	0.019575	0.017803	1.000000	0.016209	0.363049	0.000000
D	0.436758	0.554571	0.016209	1.000000	0.000000	0.000000
E	0.000000	0.000000	0.363049	0.000000	1.000000	0.169226
F	0.000000	0.000000	0.000000	0.000000	0.169226	1.000000

Fonte: o autor (2024).

Em seguida é gerado o gráfico de barras horizontal, no qual são apresentadas as similaridades de cada um dos textos com relação ao texto de referência selecionado no início do programa. Nesta etapa torna-se mais fácil a análise a acuracidade dos resultados, como apresentado Figura 21.

No momento de criação das frases do caso teste, buscou-se criar arquivos propositalmente semelhantes com o primeiro, e outros com baixíssima similaridade, utilizando como referência o arquivo de título “A”.

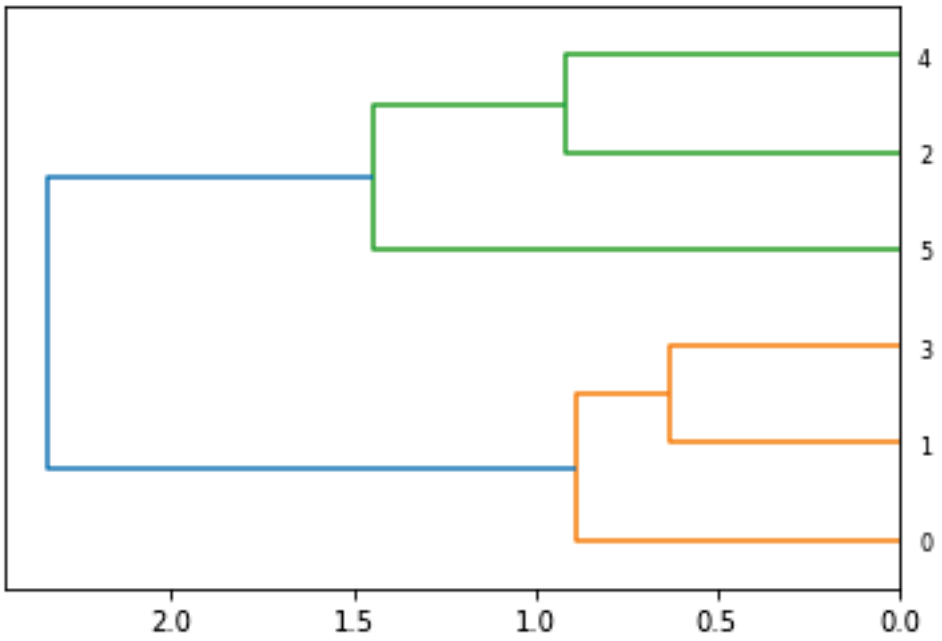
Figura 21 - Gráfico de similaridade com relação à frase A



Fonte: o autor (2024).

Por fim é gerado o dendrograma da Figura 22, o qual se trata de uma representação gráfica de uma estrutura hierárquica que exibe a relação de similaridade ou dissimilaridade entre elementos em um conjunto de dados, agrupando-os em clusters aninhados. A nomenclatura no dendrograma foi alterada por se tratar de uma primeira versão da ferramenta, seguindo a seguinte relação direta sequencial iniciando com “A” representado por “0”, “B” representado por “1”, e assim respectivamente até “F” ser representado pelo número “5”.

Figura 22 - Dendrograma de similaridade entre as frases

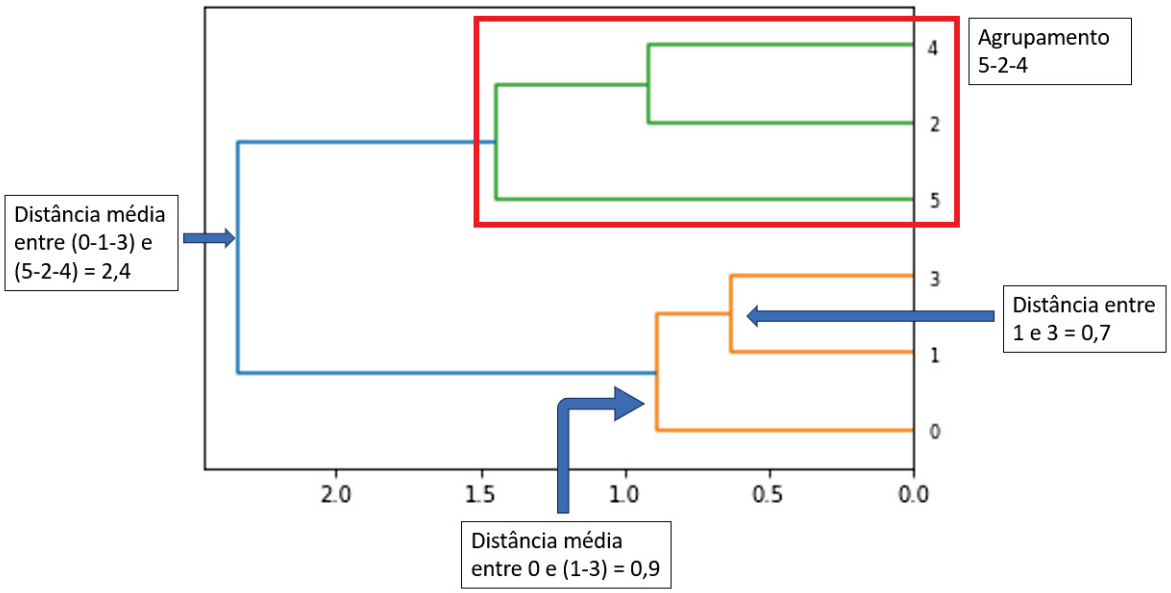


Fonte: o autor (2024).

A interpretação de um dendrograma envolve diferentes aspectos, conforme a Figura 23:

- Distância: A altura das ligações no dendrograma indica a dissimilaridade entre os elementos ou agrupamentos. Quanto maior a altura, maior a dissimilaridade.
- Agrupamentos: Os agrupamentos são formados a partir dos nós internos do dendrograma. Os elementos que estão mais próximos no dendrograma são considerados mais similares entre si.
- Corte: O dendrograma pode ser cortado em uma determinada altura para formar um número específico de clusters. O ponto de corte pode ser escolhido com base em critérios como a dissimilaridade máxima desejada ou o número desejado de clusters.

Figura 23 - Aspectos do dendrograma



Fonte: o autor (2024).

5.3 Caso com 13 relatórios

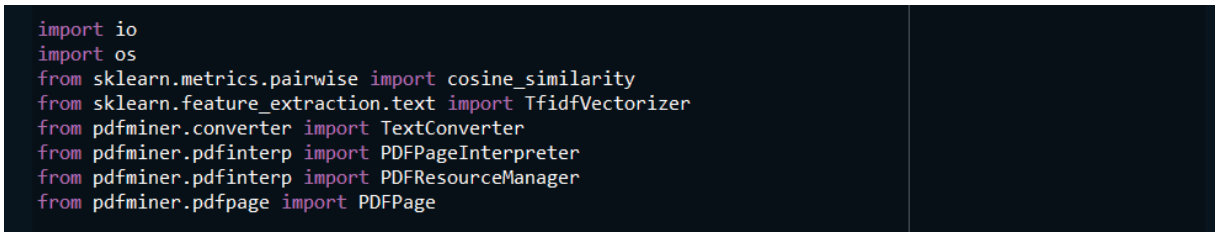
Para a segunda etapa de testes, dessa vez com acesso à parte do banco de dados de relatórios antigos da COPEL, foram implementadas mudanças de bibliotecas e funções essenciais para o desenvolvimento da busca. Essa mudança se deu pelo fato da realização de estudos mais aprofundados a respeito de suas funcionalidades e capacidades.

Uma das mudanças fundamentais do novo script é o uso da biblioteca “sklearn” substituindo a biblioteca “gensim”, usada nos testes anteriores. Essa alteração foi implementada para fins comparativos de desempenho e facilidade de utilização, uma vez que a biblioteca “sklearn” já conta com ferramentas de pré-processamento de dados e validação cruzada, além de ser mais otimizada para conjuntos de dados de tamanho moderado a grande.

Uma vez realizadas as alterações de código, foram utilizados 13 relatórios de estudos elétricos disponibilizados pela COPEL para a realização dos testes iniciais, e para isso os arquivos foram salvos no mesmo diretório do programa Python.

O script inicia com a seção de importação das bibliotecas que serão utilizadas, conforme mostra a Figura 24.

Figura 24 - Bibliotecas importadas para o cálculo de similaridade



```
import io
import os
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.feature_extraction.text import TfidfVectorizer
from pdfminer.converter import TextConverter
from pdfminer.pdfinterp import PDFPageInterpreter
from pdfminer.pdfinterp import PDFResourceManager
from pdfminer.pdfpage import PDFPage
```

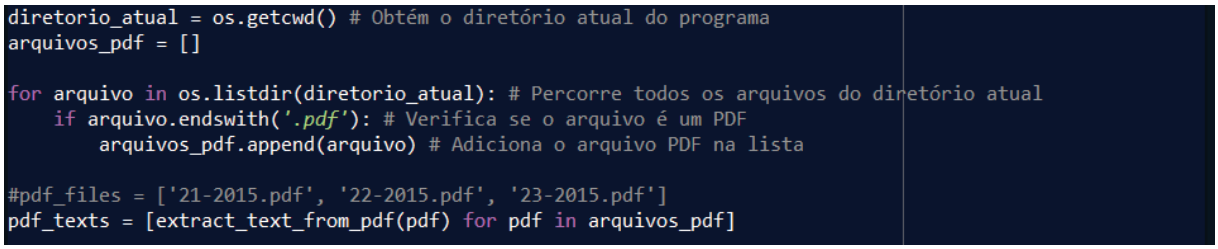
Fonte: o autor (2024).

Importações:

- io: Biblioteca para lidar com operações de entrada e saída.
- os: Biblioteca para interagir com o sistema operacional.
- sklearn.metrics.pairwise.cosine_similarity: Função para calcular a similaridade do cosseno entre vetores.
- sklearn.feature_extraction.text.TfidfVectorizer: Classe para converter uma coleção de documentos em uma matriz de recursos TF-IDF.
- pdfminer.converter.TextConverter: Classe para converter um arquivo PDF em texto.
- pdfminer.pdfinterp.PDFPageInterpreter: Classe para interpretar o conteúdo de uma página PDF.
- pdfminer.pdfinterp.PDFResourceManager: Classe para gerenciar recursos em um arquivo PDF.
- pdfminer.pdfpage.PDFPage: Classe para representar uma página PDF.

Em seguida é obtida a lista de arquivos PDF no diretório atual na seção da Figura 25.

Figura 25 - Código para obter arquivos PDF do diretório atual



```
diretorio_atual = os.getcwd() # Obtém o diretório atual do programa
arquivos_pdf = []

for arquivo in os.listdir(diretorio_atual): # Percorre todos os arquivos do diretório atual
    if arquivo.endswith('.pdf'): # Verifica se o arquivo é um PDF
        arquivos_pdf.append(arquivo) # Adiciona o arquivo PDF na lista

#pdf_files = ['21-2015.pdf', '22-2015.pdf', '23-2015.pdf']
pdf_texts = [extract_text_from_pdf(pdf) for pdf in arquivos_pdf]
```

Fonte: o autor (2024).

Com a lista de todos os arquivos PDF salvos no diretório, a próxima etapa realiza a extração do texto dos arquivos, porém ainda sem a sua vetorização, somente a criação de vetores de strings, apresentado na Figura 26

Figura 26 - Função para extrair texto de arquivos PDF

```
def extract_text_from_pdf(pdf_path):
    resource_manager = PDFResourceManager()
    fake_file_handle = io.StringIO()
    converter = TextConverter(resource_manager, fake_file_handle)
    page_interpreter = PDFPageInterpreter(resource_manager, converter)

    with open(pdf_path, 'rb') as fh:
        for page in PDFPage.get_pages(fh, caching=True, check_extractable=True):
            page_interpreter.process_page(page)

        text = fake_file_handle.getvalue()

    # close open handles
    converter.close()
    fake_file_handle.close()

    return text
```

Fonte: o autor (2024).

Esta função recebe o caminho de um arquivo PDF como parâmetro e retorna o texto extraído desse arquivo, utilizando a biblioteca “pdfminer” para abrir o arquivo PDF, percorrer suas páginas e extrair o texto de cada página.

Uma vez que se obtém o conteúdo dos documentos, é possível fazer sua vetorização através da classe “TfidfVectorizer”, a qual é usada para converter a lista pdf_texts em uma matriz de recursos TF-IDF.

Por fim, a vetorização dos textos torna possível o cálculo de similaridade entre os mesmos, através da função “cosine_similarity”, representado na Figura 27.

Figura 27 - Vetorização e cálculo de similaridade

```
tfidf = TfidfVectorizer().fit_transform(pdf_texts)
similarity_matrix = cosine_similarity(tfidf)
print(similarity_matrix)
```

Fonte: o autor (2024).

5.4 Análise dos Resultados – Caso inicial com 13 relatórios

Por fim, os resultados são apresentados em formatos de matriz, as quais possuem os valores de similaridade entre os documentos, um arquivo Excel com as

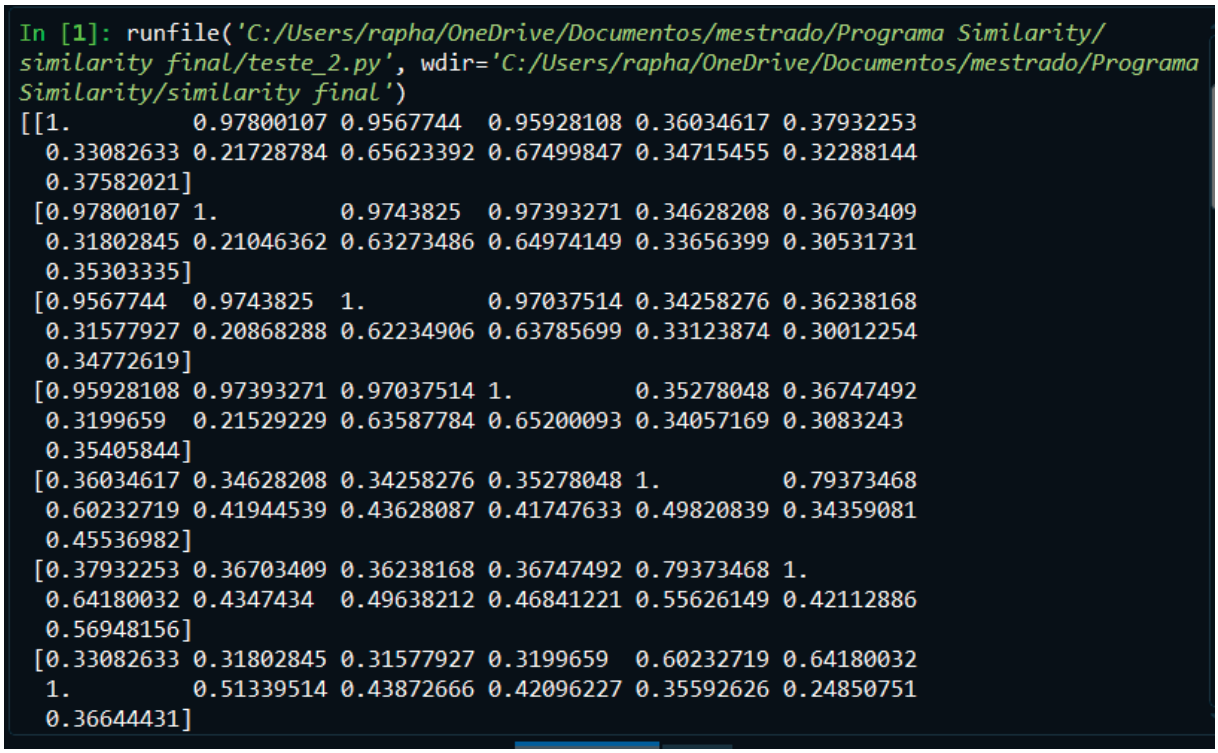
porcentagens de similaridades e o dendrograma, como ilustrado pela Figura 28, Figura 29 e Figura 31. Percebe-se que a análise dos resultados diretamente no prompt Python dificulta a interpretação, sendo esse o principal motivo da criação de uma planilha Excel para apresentação e análise dos resultados.

Para facilitar a leitura das informações das matrizes e do dendrograma, foi criada uma tabela de legenda relacionando cada um dos documentos a um número de zero a 12, conforme apresentado na Tabela 6.

Tabela 6 - Relação número com título do relatório de EE

Número	*Título do Relatório*
0	21-2015.pdf
1	22-2015.pdf
2	23-2015.pdf
3	24-2015.pdf
4	COPEL VPEE 04-2019Coordenação de Isolamento SE MEDIANEIRA NORTE.pdf
5	COPEL VPEE 07-2016 Coordenação de Isolamento SE Realeza Sul
6	COPEL VPEE 11-2016 Coordenação de Isolamento SE Andirá Lesta.pdf
7	COPEL VPEE 12-2015 Coordenação de Isolamento SE Curitiba Norte.pdf
8	RZA 1.pdf
9	RZA.pdf
10	SE Realeza Sul Estudo fluxo potência barram LT BXI 2016.pdf
11	VPEE 10-2015-VPEE 46-2014 rev 1 SE Realeza Sul Estudos a frequencia fundamental
12	VPEE 18-2016 – Evolução da Assimetria e das Correntes de Curto-Circuito LT 230 kV BXI_RZS.pdf

Figura 28 - Resultados de similaridade no prompt



```
In [1]: runfile('C:/Users/rapha/OneDrive/Documentos/mestrado/Programa Similarity/
similarity final/teste_2.py', wdir='C:/Users/rapha/OneDrive/Documentos/mestrado/Programa
Similarity/similarity final')
[[1.         0.97800107 0.9567744  0.95928108 0.36034617 0.37932253
  0.33082633 0.21728784 0.65623392 0.67499847 0.34715455 0.32288144
  0.37582021]
 [0.97800107 1.         0.9743825  0.97393271 0.34628208 0.36703409
  0.31802845 0.21046362 0.63273486 0.64974149 0.33656399 0.30531731
  0.35303335]
 [0.9567744  0.9743825  1.         0.97037514 0.34258276 0.36238168
  0.31577927 0.20868288 0.62234906 0.63785699 0.33123874 0.30012254
  0.34772619]
 [0.95928108 0.97393271 0.97037514 1.         0.35278048 0.36747492
  0.3199659  0.21529229 0.63587784 0.65200093 0.34057169 0.3083243
  0.35405844]
 [0.36034617 0.34628208 0.34258276 0.35278048 1.         0.79373468
  0.60232719 0.41944539 0.43628087 0.41747633 0.49820839 0.34359081
  0.45536982]
 [0.37932253 0.36703409 0.36238168 0.36747492 0.79373468 1.
  0.64180032 0.4347434  0.49638212 0.46841221 0.55626149 0.42112886
  0.56948156]
 [0.33082633 0.31802845 0.31577927 0.3199659  0.60232719 0.64180032
  1.         0.51339514 0.43872666 0.42096227 0.35592626 0.24850751
  0.36644431]
```

Fonte: o autor (2024).

Figura 29 - Resultados de similaridade em arquivo Excel

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	100%	98%	96%	96%	36%	38%	33%	22%	66%	67%	35%	32%	38%
1	98%	100%	97%	97%	35%	37%	32%	21%	63%	65%	34%	31%	35%
2	96%	97%	100%	97%	34%	36%	32%	21%	62%	64%	33%	30%	35%
3	96%	97%	97%	100%	35%	37%	32%	22%	64%	65%	34%	31%	35%
4	36%	35%	34%	35%	100%	79%	60%	42%	44%	42%	50%	34%	46%
5	38%	37%	36%	37%	79%	100%	64%	43%	50%	47%	56%	42%	57%
6	33%	32%	32%	32%	60%	64%	100%	51%	44%	42%	36%	25%	37%
7	22%	21%	21%	22%	42%	43%	51%	100%	26%	25%	27%	19%	24%
8	66%	63%	62%	64%	44%	50%	44%	26%	100%	90%	47%	43%	55%
9	67%	65%	64%	65%	42%	47%	42%	25%	90%	100%	44%	43%	52%
10	35%	34%	33%	34%	50%	56%	36%	27%	47%	44%	100%	65%	60%
11	32%	31%	30%	31%	34%	42%	25%	19%	43%	43%	65%	100%	54%
12	38%	35%	35%	35%	46%	57%	37%	24%	55%	52%	60%	54%	100%

Fonte: o autor (2024).

Os resultados obtidos pela matriz de similaridades demonstram os níveis de similaridade dos relatórios, onde é possível verificar, como esperado, que a diagonal principal da matriz é composta por similaridades de 100%, uma vez que a comparação se trata do documento com o próprio.

Ainda através da matriz de similaridade, percebe-se que a representação numérica da similaridade entre relatórios facilita a identificação de padrões do banco de dados, tendo destaque os agrupamentos demonstrados na Figura 30.

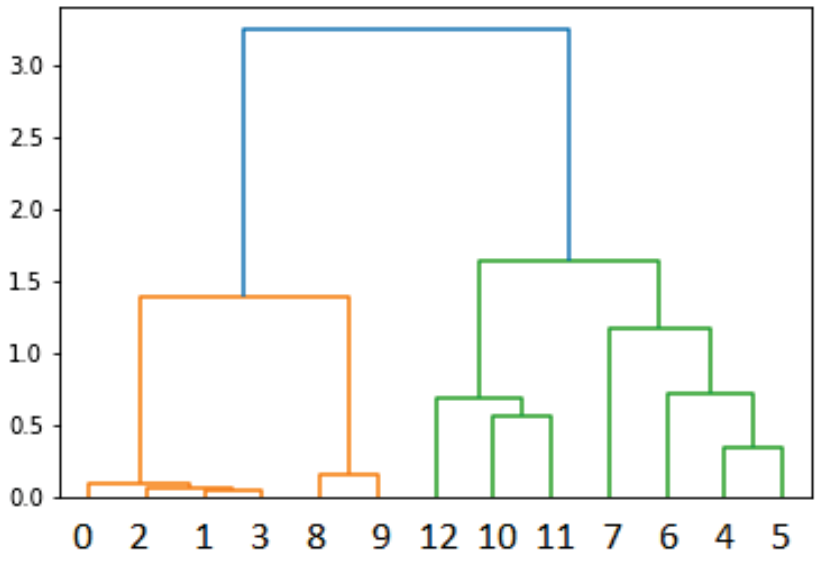
Figura 30 - Agrupamentos na matriz de similaridades

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	100%	98%	96%	96%	36%	38%	33%	22%	66%	67%	35%	32%	38%
1	98%	100%	97%	97%	35%	37%	32%	21%	63%	65%	34%	31%	35%
2	96%	97%	100%	97%	34%	36%	32%	21%	62%	64%	33%	30%	35%
3	96%	97%	97%	100%	35%	37%	32%	22%	64%	65%	34%	31%	35%
4	36%	35%	34%	35%	100%	79%	60%	42%	44%	42%	50%	34%	46%
5	38%	37%	36%	37%	79%	100%	64%	43%	50%	47%	56%	42%	57%
6	33%	32%	32%	32%	60%	64%	100%	51%	44%	42%	36%	25%	37%
7	22%	21%	21%	22%	42%	43%	51%	100%	26%	25%	27%	19%	24%
8	66%	63%	62%	64%	44%	50%	44%	26%	100%	90%	47%	43%	55%
9	67%	65%	64%	65%	42%	47%	42%	25%	90%	100%	44%	43%	52%
10	35%	34%	33%	34%	50%	56%	36%	27%	47%	44%	100%	65%	60%
11	32%	31%	30%	31%	34%	42%	25%	19%	43%	43%	65%	100%	54%
12	38%	35%	35%	35%	46%	57%	37%	24%	55%	52%	60%	54%	100%

Fonte: o autor (2024).

Apesar da facilidade de identificação dos agrupamentos na matriz numérica, a forma mais rápida e intuitiva de análise de similaridades dos relatórios presentes no banco de dados é através do dendrograma, apresentado na Figura 31. Nele é possível identificar dois agrupamentos principais, identificados com as cores laranjada e verde, em que as similaridades são maiores entre os documentos. Um ponto de destaque para essa análise foi a capacidade de agrupar todos os arquivos referentes a estudos de coordenação de isolamento no mesmo grupo (7-6-4-5), mais uma vez demonstrando e validando o desempenho da ferramenta.

Figura 31 - Dendrograma do cálculo de similaridade



Fonte: o autor (2024).

Por fim, para realizar uma análise do custo computacional da busca das similaridades dos 13 relatórios, foi utilizado um notebook com as seguintes especificações:

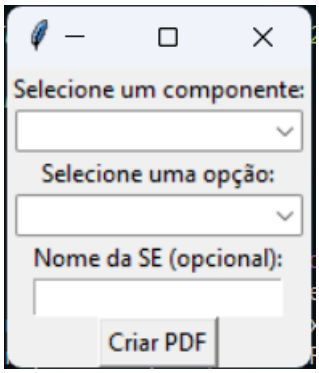
- Processador: Intel Core i7-8565U
- Memória RAM: 8 GB
- Memória HD: 1 TB
- Sistema Operacional: Windows 11 versão 22H2
- Versão Python: 3.8.3

Assim o tempo para finalização da busca e geração dos gráficos e matrizes foi de 6 minutos e 45 segundos.

5.5 Caso com 72 relatórios

Para esta etapa de testes, foi incorporada à ferramenta uma interface para a inserção dos dados do relatório que se deseja utilizar como referência na busca de trabalhos similares. A interface está ilustrada na Figura 32.

Figura 32 - Interface de entrada de dados



Fonte: o autor (2024).

Para cada seleção de componente de estudo, são abertas diferentes opções de parâmetros, variando de acordo com a estrutura da Tabela 7.

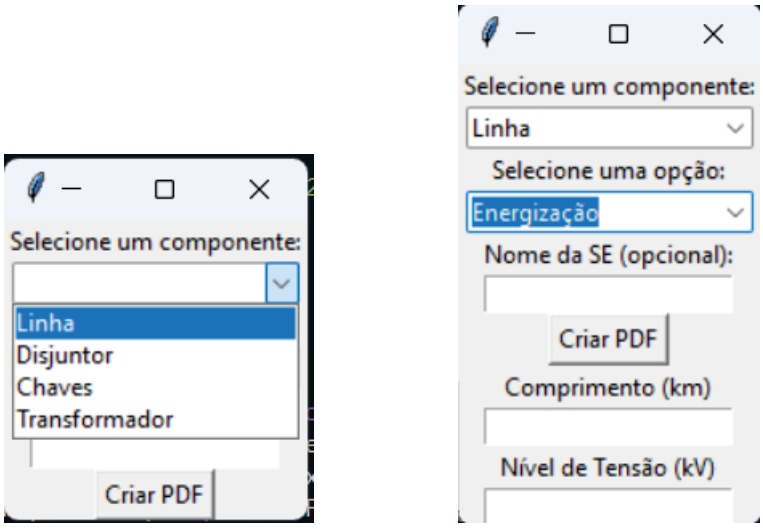
Tabela 7 - Estrutura de informações da interface

Componente	Opção	Comprimento	Nível de Tensão	Nível de Potência
Linha	Energização	SIM	SIM	NÃO
	Religamento mono e tri	SIM	SIM	NÃO
	Extinção de arco	SIM	SIM	NÃO
	Rejeição de Carga	SIM	SIM	NÃO
Disjuntor	TRT	NÃO	SIM	NÃO
Chaves	Correntes induzidas em lâmina de terra	NÃO	SIM	NÃO
Transformador	Energização	NÃO	SIM	SIM
Banco de Capacitores	Energização	NÃO	SIM	SIM

Fonte: o autor (2024).

A Figura 33 ilustra um exemplo de interação com a interface em que se deseja utilizar como referência um estudo energização de linha de transmissão, e dessa forma sendo necessário informa as informações referentes a esse tipo de documento.

Figura 33 - Exemplos de interação com a interface



Fonte: o autor (2024).

Uma vez inseridas as informações necessárias, é gerado um documento PDF com essas informações, o qual é adicionado automaticamente ao banco de dados para que faça parte do processo de vetorização e cálculos de similaridade com os demais. Para facilitar a identificação do documento de referência na visualização dos resultados o processo de armazenamento do arquivo intitula o mesmo com o nome de "_INPUT".

Para o caso final de estudo para o presente trabalho apresenta um banco de dados com 72 arquivos de relatórios de estudos elétricos disponibilizados pela COPEL, no qual foram realizados dois tipos de buscas.

A primeira busca considerou o banco de dados único com todos os 72 documentos, no qual são calculadas as similaridades de todos os arquivos com cada um dos relatórios, resultando em uma matriz de dimensão 72x72.

Já para a segunda busca, foi realizada a categorização dos documentos de acordo com a estrutura de seleção da interface, de modo que, uma vez selecionado componente de interesse, apenas os relatórios categorizados para aquela seleção serão incluídos na busca.

O intuito de realizar a categorização do banco de dados (BD) é realizar agrupamentos prévios dos arquivos, evitando possíveis erros de identificação da ferramenta e um grande impacto na economia de custo computacional, uma vez que toda a etapa de pré-processamento textual e cálculos das similaridades estará limitado a apenas uma fração do banco de dados.

Por outro lado, o agrupamento prévio decorrente da categorização do BD pode causar redução na análise de similaridade, uma vez que está sujeito à interpretação humana, e consequentemente excluir da análise documentos com informações que podem ser relevantes para o desenvolvimento, mesmo se tratando de diferentes componentes ou opções de estudo.

5.6 Caso final com 72 relatórios e banco de dados não categorizado

O primeiro cenário de análise para a busca por similaridade considera o banco de dados com os 72 arquivos de relatórios unificado, de forma que as comparações na busca não são diferenciadas de acordo com o componente principal de cada relatório. A Tabela 8 apresenta o nome dos 72 arquivos disponibilizados pela Copel para a realização da busca.

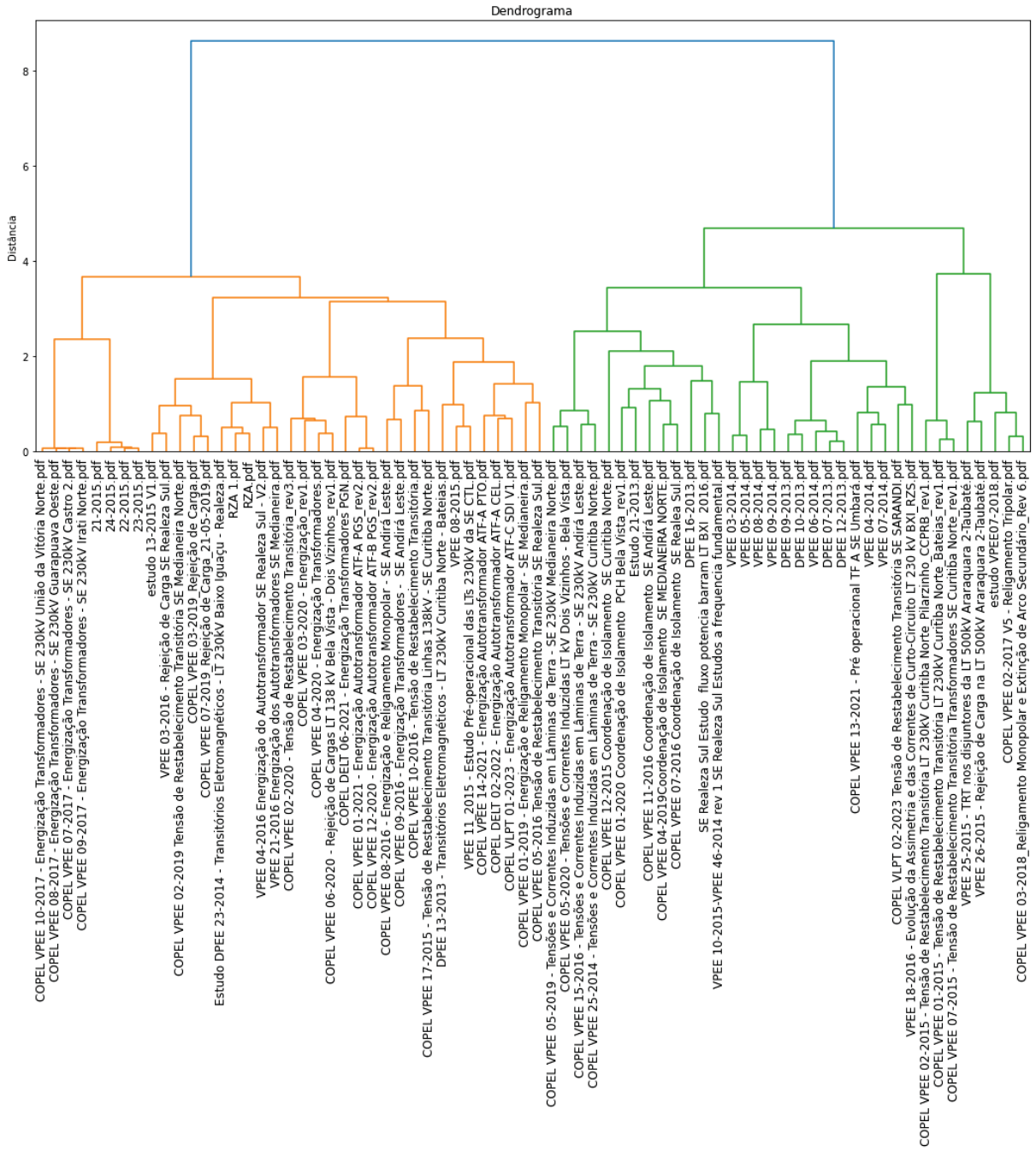
Tabela 8 - Relatórios disponibilizados pela Copel para realização do estudo de caso

NUMERAÇÃO	TÍTULO DO RELATÓRIO
0	21-2015.pdf
1	22-2015.pdf
2	23-2015.pdf
3	24-2015.pdf
4	COPEL DELT 02-2022 - Energização Autotransformador ATF-A CEL.pdf
5	COPEL DELT 06-2021 - Energização Transformadores PGN.pdf
6	COPEL VLPT 01-2023 - Energização Autotransformador ATF-C SDI V1.pdf
7	COPEL VLPT 02-2023 Tensão de Restabelecimento Transitória SE SARANDI.pdf
8	COPEL VPEE 01-2015 - Tensão de Restabelecimento Transitória LT 230kV Curitiba Norte_Bateias_rev1.pdf
9	COPEL VPEE 01-2019 - Energização e Religamento Monopolar - SE Medianeira.pdf
10	COPEL VPEE 01-2020 Coordenação de Isolamento PCH Bela Vista_rev1.pdf
11	COPEL VPEE 01-2021 - Energização Autotransformador ATF-A PGS_rev2.pdf
12	COPEL VPEE 02-2015 - Tensão de Restabelecimento Transitória LT 230kV Curitiba Norte_Pilarzinho_CCPRB_rev1.pdf
13	COPEL VPEE 02-2017 V5 - Religamento Tripolar.pdf
14	COPEL VPEE 02-2019 Tensão de Restabelecimento Transitória SE Medianeira Norte.pdf
15	COPEL VPEE 02-2020 - Tensão de Restabelecimento Transitória_rev3.pdf
16	COPEL VPEE 03-2018_Religamento Monopolar e Extinção de Arco Secundário_Rev 6.pdf
17	COPEL VPEE 03-2019_Rejeição de Carga.pdf
18	COPEL VPEE 03-2020 - Energização_rev1.pdf
19	COPEL VPEE 04-2019Coordenação de Isolamento SE MEDIANEIRA NORTE.pdf
20	COPEL VPEE 04-2020 - Energização Transformadores.pdf
21	COPEL VPEE 05-2016 Tensão de Restabelecimento Transitória SE Realeza Sul.pdf
22	COPEL VPEE 05-2019 - Tensões e Correntes Induzidas em Lâminas de Terra - SE 230kV Medianeira Norte.pdf
23	COPEL VPEE 05-2020 - Tensões e Correntes Induzidas LT kV Dois Vizinhos - Bela Vista.pdf
24	COPEL VPEE 06-2020 - Rejeição de Cargas LT 138 kV Bela Vista - Dois Vizinhos_rev1.pdf
25	COPEL VPEE 07-2015 - Tensão de Restabelecimento Transitória Transformadores SE Curitiba Norte_rev1.pdf
26	COPEL VPEE 07-2016 Coordenação de Isolamento SE Realea Sul.pdf
27	COPEL VPEE 07-2017 - Energização Transformadores - SE 230kV Castro 2.pdf
28	COPEL VPEE 07-2019_Rejeição de Carga_21-05-2019.pdf
29	COPEL VPEE 08-2016 - Energização e Religamento Monopolar - SE Andirá Leste.pdf
30	COPEL VPEE 08-2017 - Energização Transformadores - SE 230kV Guarapuava Oeste.pdf
31	COPEL VPEE 09-2016 - Energização Transformadores - SE Andirá Leste.pdf
32	COPEL VPEE 09-2017 - Energização Transformadores - SE 230kV Irati Norte.pdf
33	COPEL VPEE 10-2016 - Tensão de Restabelecimento Transitória.pdf
34	COPEL VPEE 10-2017 - Energização Transformadores - SE 230kV União da Vitória Norte.pdf
35	COPEL VPEE 11-2016 Coordenação de Isolamento SE Andirá Leste.pdf
36	COPEL VPEE 12-2015 Coordenação de Isolamento SE Curitiba Norte.pdf

37	COPEL VPEE 12-2020 - Energização Autotransformador ATF-B PGS_rev2.pdf
38	COPEL VPEE 13-2021 - Pré operacional TF A SE Umbará.pdf
39	COPEL VPEE 14-2021 - Energização Autotransformador ATF-A PTO.pdf
40	COPEL VPEE 15-2016 - Tensões e Correntes Induzidas em Lâminas de Terra - SE 230kV Andirá Leste.pdf
41	COPEL VPEE 17-2015 - Tensão de Restabelecimento Transitória Linhas 138kV - SE Curitiba Norte.pdf
42	COPEL VPEE 25-2014 - Tensões e Correntes Induzidas em Lâminas de Terra - SE 230kV Curitiba Norte.pdf
43	DPEE 07-2013.pdf
44	DPEE 09-2013.pdf
45	DPEE 10-2013.pdf
46	DPEE 12-2013.pdf
47	DPEE 13-2013 - Transitórios Eletromagnéticos - LT 230kV Curitiba Norte - Bateias.pdf
48	DPEE 16-2013.pdf
49	estudo 13-2015 V1.pdf
50	Estudo 21-2013.pdf
51	Estudo DPEE 23-2014 - Transitórios Eletromagnéticos - LT 230kV Baixo Iguaçu - Realeza.pdf
52	estudo VPEE07-2018.pdf
53	RZA 1.pdf
54	RZA.pdf
55	SE Realeza Sul Estudo fluxo potencia barram LT BXI 2016.pdf
56	VPEE 03-2014.pdf
57	VPEE 03-2016 - Rejeição de Carga SE Realeza Sul.pdf
58	VPEE 04-2014.pdf
59	VPEE 04-2016 Energização do Autotransformador SE Realeza Sul - V2.pdf
60	VPEE 05-2014.pdf
61	VPEE 06-2014.pdf
62	VPEE 07-2014.pdf
63	VPEE 08-2014.pdf
64	VPEE 08-2015.pdf
65	VPEE 09-2014.pdf
66	VPEE 10-2015-VPEE 46-2014 rev 1 SE Realeza Sul Estudos a frequencia fundamental.pdf
67	VPEE 11_2015 - Estudo Pré-operacional das LTs 230kV da SE CTL.pdf
68	VPEE 18-2016 - Evolução da Assimetria e das Correntes de Curto-Circuito LT 230 kV BXI_RZS.pdf
69	VPEE 21-2016 Energização dos Autotransformadores SE Medianeira.pdf
70	VPEE 25-2015 - TRT nos disjuntores da LT 500kV Araraquara 2-Taubaté.pdf
71	VPEE 26-2015 - Rejeição de Carga na LT 500kV Araraquara 2-Taubaté.pdf

A Figura 34 apresenta o dendrograma resultante da análise completa do BD, a qual levou 15 minutos e 26 segundos para finalizar o processamento. Este custo computacional se dá considerando as mesmas configurações do computador utilizado nos casos anteriores.

Figura 34 - Dendrograma da busca por similaridade do BD não categorizado

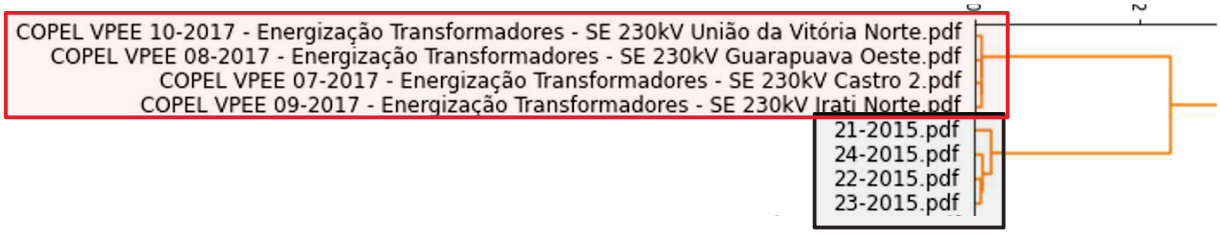


Fonte: o autor (2024).

Ao realizar uma análise dos agrupamentos resultantes no dendrograma, é possível perceber que, de maneira geral, os documentos com maior similaridade são relacionados através dos seus componentes de estudo, como por exemplo os agrupamentos primários da Figura 35, em que o agrupamento destacado em vermelho relaciona relatórios que abordam a energização de transformadores, enquanto no

agrupamento destacado em preto todos os relatórios abordam religamentos em linhas de transmissão.

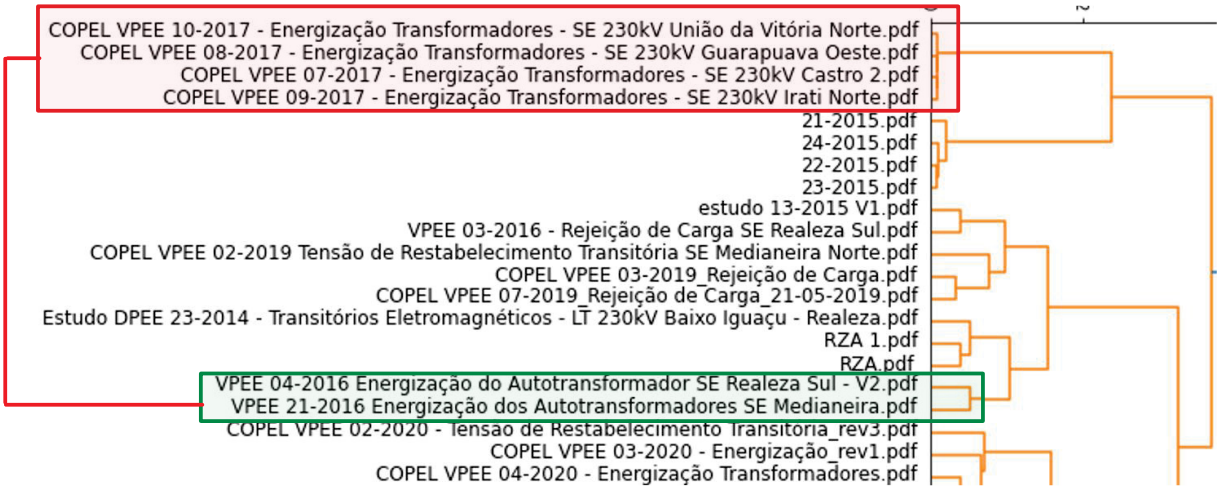
Figura 35 – Exemplo de agrupamentos primários da busca por similaridade não categorizada



Fonte: o autor (2024).

Dessa forma, em uma análise de agrupamentos primários, tem-se resultados satisfatórios de categorização automática do BD, porém quando a análise se estende para os agrupamentos secundários, percebe-se que o agrupamento dos arquivos de transformadores deveria estar conectado à outros agrupamentos de relatórios que abordam a energização de transformadores antes da comparação direta de similaridade com os relatórios de linhas de transmissão, conforme demonstrado na Figura 36.

Figura 36 - Exemplo de similaridade entre agrupamentos de relatórios que abordam estudos correlatos



Fonte: o autor (2024).

Outra similaridade verificada nos agrupamentos primários do dendrograma são relacionados à localização geográfica das subestações presentes nos relatórios, o que muitas vezes pode ser útil também para o desenvolvimento de novos projetos, como por exemplo o agrupamento dos arquivos 53 (RZA 1.pdf) e 54 (RZA.pdf), os quais possuem os seguintes títulos:

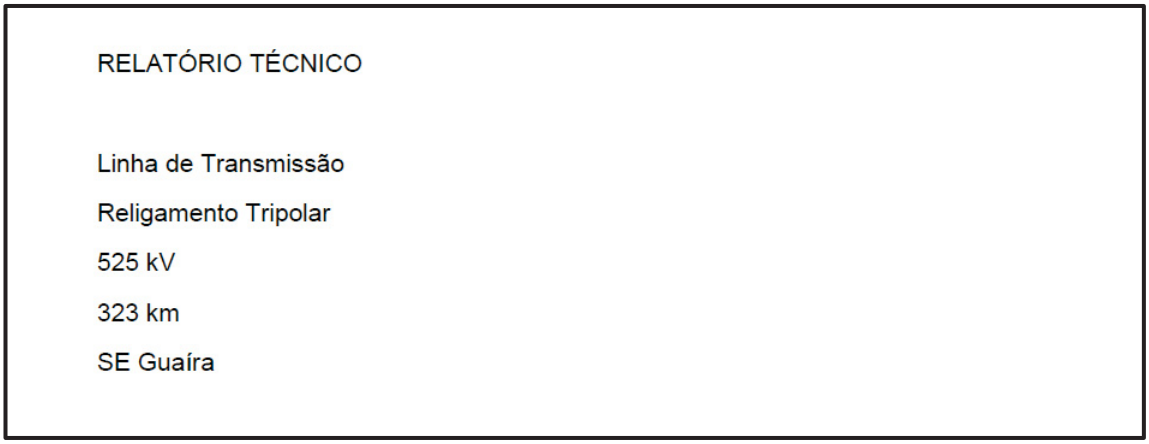
- Relatório 53 (RZA 1.pdf): RELATÓRIO TÉCNICO REFERENTE À NOVA INSTALAÇÃO DA REDE BÁSICA - PROJETO BÁSICO - ANÁLISE DE TRANSITÓRIOS ELETROMAGNÉTICOS – TENSÃO DE RESTABELECIMENTO TRANSITÓRIA DOS DISJUNTORES DA LT 230 KV BAIXO IGUAÇU - REALEZA
- Relatório 54 (RZA.pdf): RELATÓRIO TÉCNICO REFERENTE À NOVA INSTALAÇÃO DA REDE BÁSICA - PROJETO BÁSICO - ANÁLISE DE TRANSITÓRIOS ELETROMAGNÉTICOS ENERGIZAÇÃO E RELIGAMENTO – LT 230 kV REALEZA – BAIXO IGUAÇU

Neste caso, verifica-se que os relatórios 53 e 54 abordam, respectivamente, estudos de tensão de restabelecimento transitória de disjuntores, e energização e religamentos, ambos na mesma linha de transmissão entre as subestações de Realeza e Baixo Iguaçu. Esse tipo de similaridade é válido uma vez que através dela é possível identificar trabalhos de modelagem de trechos do sistema elétrico já realizados.

Através desses exemplos percebe-se que ambos os tipos de similaridade identificados pela ferramenta (com relação ao componente de estudo e em relação à localização) são importantes, porém ao mesmo tempo fica evidente a necessidade de definir qual dos parâmetros de similaridade a busca deve seguir, pois a falta de categorização do banco de dados deixa com que esses dois critérios se misturem e consequentemente provoque uma redução na acuracidade dos resultados.

A fim de validar a capacidade de identificação de similaridade com um arquivo de referência “_INPUT”, o qual recebeu a numeração 72. Neste arquivo foram incluídas as informações mais características do relatório 0 (21-2015.pdf) a fim de verificar o valor de similaridade entre eles e se a ferramenta cometeria o erro de indicar outro arquivo como sendo mais similar. A Figura 37 demonstra as informações incluídas no arquivo de referência.

Figura 37 - Informações do arquivo de referência "_INPUT" para o BD não categorizado



Fonte: o autor (2024).

A verificação do arquivo indicado como mais similar com o documento "_INPUT" se dá através da verificação da matriz de similaridade, onde para cada um dos arquivos presentes no banco de dados são dados valores de similaridade com relação a todos os demais. Dessa forma, como o novo arquivo de referência recebeu a última numeração (72), para a verificação da sua similaridade com os demais relatórios basta analisar a última coluna da matriz de similaridades, apresentada na Tabela 9.

Tabela 9 - Valores de similaridade do BD não categorizado com o arquivo _INPUT

Arquivo	72	Arquivo	72	Arquivo	72	Arquivo	72
0	24%	**21**	9%	**42**	12%	**63**	10%
1	24%	**22**	13%	**43**	13%	**64**	19%
2	18%	**23**	15%	**44**	15%	**65**	23%
3	17%	**24**	18%	**45**	14%	**66**	11%
4	9%	**25**	5%	**46**	14%	**67**	18%
5	19%	**26**	15%	**47**	16%	**68**	15%
6	12%	**27**	6%	**48**	11%	**69**	15%
7	15%	**28**	10%	**49**	14%	**70**	10%
8	6%	**29**	12%	**50**	14%	**71**	10%
9	12%	**30**	6%	**51**	21%	**72**	100%
10	16%	**31**	11%	**52**	9%		
11	19%	**32**	6%	**53**	15%		
12	8%	**33**	10%	**54**	16%		
13	9%	**34**	6%	**55**	14%		
14	13%	**35**	8%	**56**	8%		
15	16%	**36**	6%	**57**	11%		
16	10%	**37**	19%	**58**	14%		
17	10%	**38**	17%	**59**	14%		
18	15%	**39**	10%	**60**	21%		
19	12%	**40**	11%	**61**	13%		
20	18%	**41**	11%	**62**	16%		

Fonte: o autor (2024).

Os resultados demonstram que, mesmo sem a categorização do BD, a ferramenta de busca por similaridade foi capaz de identificar o relatório 0 com o maior valor de similaridade. No entanto em conjunto com o relatório 0, que recebeu o valor de 24%, o relatório 1 também recebeu o mesmo valor de similaridade.

Verificando o conteúdo o relatório 1, percebe-se que a mesma indicação de similaridade com relação ao arquivo de referência se mostrou válida, uma vez que o estudo possui as mesmas características técnicas do relatório 0 e a única alteração de localização se dá por ser uma linha de transmissão que une a subestação Guaíra (conforme solicitado no arquivo de referência) com a subestação Sarandi CD.

5.7 Caso final com 72 relatórios e banco de dados categorizado por componente de estudo

A partir dos resultados do banco de dados não categorizado viu-se a necessidade da realização de uma categorização do BD, com o objetivo de evitar que diferentes parâmetros de similaridade sejam levados em consideração nos momentos de análise e geração do dendrograma.

Para a categorização do banco de dados, foi realizada, em conjunto com um especialista responsável pelo desenvolvimento dos relatórios da COPEL, uma classificação dos relatórios que consideram os seguintes componentes principais:

- Linha de transmissão
- Transformadores
- Disjuntores
- Chaves
- Bancos de capacitores

Dessa forma, para cada componente foi criado um novo diretório, contendo os documentos referentes à cada categoria, de forma que, ao iniciar a execução da ferramenta de busca por similaridade, o programa muda seu diretório de seleção de arquivos de acordo com a seleção na interface. A seguir serão apresentados os resultados para a busca categorizada para fins de comparação de com a busca não realizada apresentada anteriormente.

5.7.1 Linhas de Transmissão

A seção de linhas de transmissão considera os relatórios que abordam, como tema principal, os seguintes estudos:

- Energização
- Religamento Monopolar
- Religamento Tripolar
- Extinção de Arco
- Rejeição de Carga

A Tabela 10 apresenta os relatórios categorizados como estudos de linhas de transmissão:

Tabela 10 - Relatórios categorizados para Linhas de transmissão

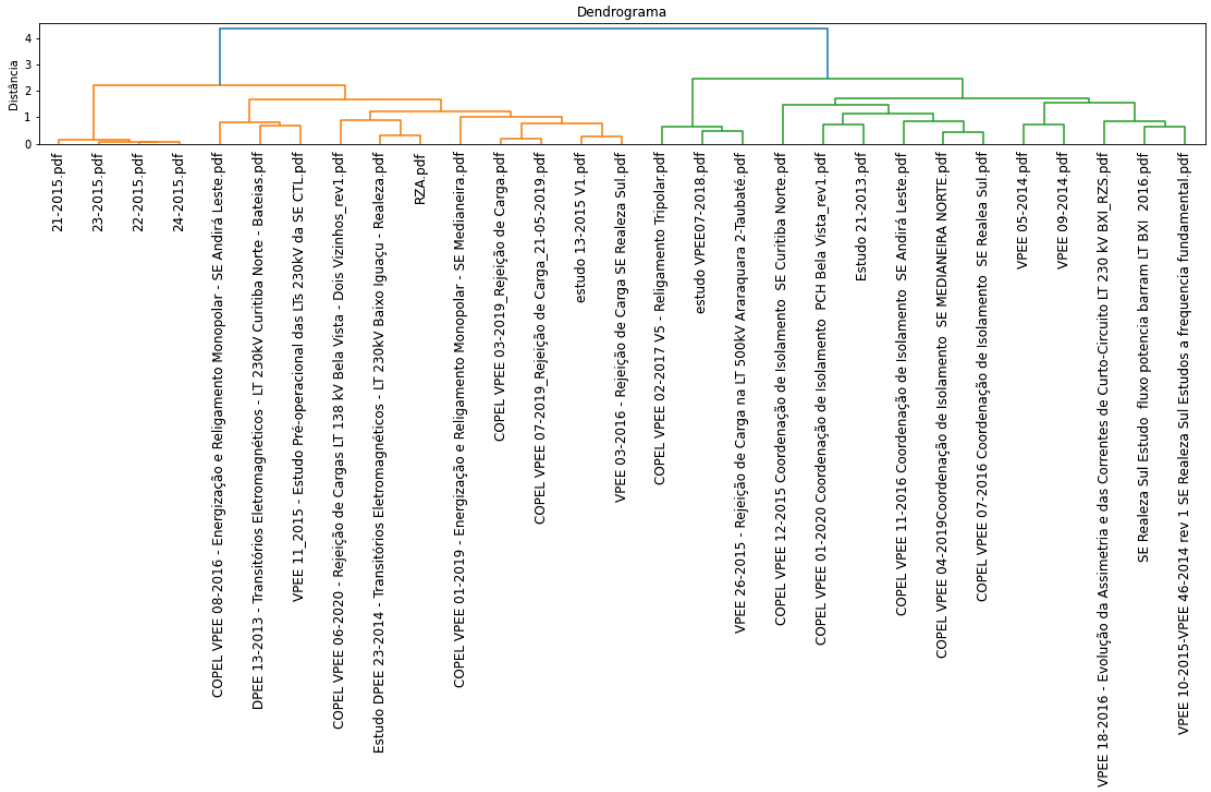
Numeração	***Título do relatório***
0	21-2015.pdf
1	22-2015.pdf
2	23-2015.pdf
3	24-2015.pdf
4	COPEL VPEE 01-2019 - Energização e Religamento Monopolar - SE Medianeira.pdf
5	COPEL VPEE 01-2020 Coordenação de Isolamento PCH Bela Vista_rev1.pdf
6	COPEL VPEE 02-2017 V5 - Religamento Tripolar.pdf
7	COPEL VPEE 03-2019_Rejeição de Carga.pdf
8	COPEL VPEE 04-2019Coordenação de Isolamento SE MEDIANEIRA NORTE.pdf
9	COPEL VPEE 06-2020 - Rejeição de Cargas LT 138 kV Bela Vista - Dois Vizinhos_rev1.pdf
10	COPEL VPEE 07-2016 Coordenação de Isolamento SE Realea Sul.pdf
11	COPEL VPEE 07-2019_Rejeição de Carga_21-05-2019.pdf
12	COPEL VPEE 08-2016 - Energização e Religamento Monopolar - SE Andirá Leste.pdf
13	COPEL VPEE 11-2016 Coordenação de Isolamento SE Andirá Leste.pdf
14	COPEL VPEE 12-2015 Coordenação de Isolamento SE Curitiba Norte.pdf
15	DPEE 13-2013 - Transitórios Eletromagnéticos - LT 230kV Curitiba Norte - Bateias.pdf
16	estudo 13-2015 V1.pdf
17	Estudo 21-2013.pdf
18	Estudo DPEE 23-2014 - Transitórios Eletromagnéticos - LT 230kV Baixo Iguaçu - Realeza.pdf
19	estudo VPEE07-2018.pdf
20	RZA.pdf
21	SE Realeza Sul Estudo fluxo potencia barram LT BXI 2016.pdf
22	VPEE 03-2016 - Rejeição de Carga SE Realeza Sul.pdf
23	VPEE 05-2014.pdf
24	VPEE 09-2014.pdf

25	VPEE 10-2015-VPEE 46-2014 rev 1 SE Realeza Sul Estudos a frequencia fundamental.pdf
26	VPEE 11_2015 - Estudo Pré-operacional das LTs 230kV da SE CTL.pdf
27	VPEE 18-2016 - Evolução da Assimetria e das Correntes de Curto-Circuito LT 230 kV BXI_RZS.pdf
28	VPEE 26-2015 - Rejeição de Carga na LT 500kV Araraquara 2-Taubaté.pdf

Fonte: o autor (2024).

Uma vez finalizada a categorização foi primeiramente realizada a busca por similaridade entre os documentos, tendo como resultado o dendrograma da Figura 38.

Figura 38 - Dendrograma da busca por similaridade - categoria Linha de Transmissão

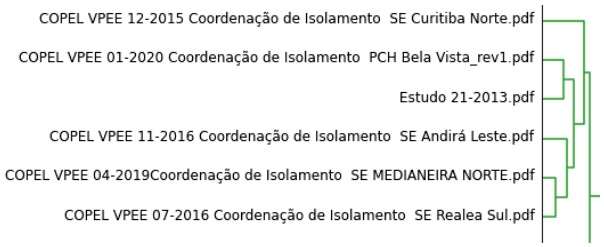


Fonte: o autor (2024).

Neste resultado é possível observar alguns agrupamentos já esperados, como por exemplo, o agrupamento da

Figura 39, em que todos os relatórios que abordam a coordenação de isolamento estão presentes. Outro agrupamento esperado, se trata dos arquivos numerados de 0 a 3, referentes a relatórios de estudos de linhas de transmissão de uma mesma região, a qual aborda as subestações de Guaíra e Sarandi CD.

Figura 39 -Agrupamento de relatórios de coordenação de isolamento de linhas de transmissão



Fonte: o autor (2024).

Uma vez realizada a busca somente entre os relatórios já existentes foi utilizada a interface de criação de um novo documento para o banco de dados, com as informações de um novo empreendimento hipotético para fins de validação. Para facilitar a verificação da busca em relação a valores de referência, foram utilizadas informações próximas às de um dos relatórios já existentes, com o objetivo de que o resultado apresente, na matriz de similaridades, a maior semelhança entre o arquivo com as informações inseridas e o documento escolhido.

O documento escolhido para utilizar como base para as informações de entrada foi o arquivo numerado 0: 21-2015.pdf, o qual possui o seguinte título: RELATÓRIO TÉCNICO REFERENTE À NOVA INSTALAÇÃO DA REDE BÁSICA - RELATÓRIO R2 - ANÁLISE DE TRANSITÓRIOS ELETROMAGNÉTICOS LT 525 kV GUAÍRA – FOZ DO IGUAÇU CD

Dessa forma o arquivo com as informações inseridas como referência da busca está ilustrado na Figura 40.

Figura 40 - Informações do arquivo de referência "_INPUT" para a categoria de linhas de transmissão

RELATÓRIO TÉCNICO

Linha de Transmissão

Religamento Tripolar

525 kV

323 km

SE Guaíra

Fonte: o autor (2024).

A partir da análise da matriz de similaridade, da Tabela 11,resultante da busca em que inclui o arquivo “_INPUT”, numerado com 29, temos que o maior nível de similaridade, confirme esperado, é com os arquivos numerados 0 e 1. Essa mesma similaridade com dois relatórios pode ser considerada válida, uma vez que, como comentado anteriormente, os arquivos numerados de 0 a 3 apresentam estudos de uma mesma região, sendo o título do documento numerado 1 (22-2015.pdf): RELATÓRIO TÉCNICO REFERENTE À NOVA INSTALAÇÃO DA REDE BÁSICA - RELATÓRIO R2 - ANÁLISE DE TRANSITÓRIOS ELETROMAGNÉTICOS LT 525 kV GUAÍRA – SARANDI CD.

Dessa forma fica claro do motivo da similaridade ter sido a mesma em ambos os relatórios, pois as informações do arquivo de referência pedem um estudo com a SE Guaíra e nível de tensão de 525 kV, sendo estes parâmetros válidos para ambos os relatórios.

Tabela 11 - Valores de similaridade do BD de linha de transmissão com o arquivo _INPUT

	29
0	24%
1	24%
2	19%
3	18%
4	12%
5	19%
6	11%
7	11%
8	15%
9	20%
10	18%
11	10%
12	13%
13	11%
14	8%
15	17%
16	15%
17	17%
18	22%
19	11%
20	16%
21	18%
22	11%
23	20%
24	23%
25	14%

26	19%
27	18%
28	12%
29	100%

Fonte: o autor (2024).

Para este caso em que se deseja encontrar os relatórios antigos mais similares a um documento de referência utilizamos sempre somente a matriz de similaridade, uma vez que o objetivo do dendrograma é criar agrupamentos e relações entre os arquivos do banco de dados como um todo, e não necessariamente a posição do arquivo _INPUT estará conectada ao arquivo mais similar. Isso ocorre devido os valores de similaridade do arquivo de referência com os demais ser na faixa de 20%, enquanto as similaridades dos relatórios entre si têm valores na faixa de 80% e 90%.

Para as próximas seções foram realizadas abordagens semelhantes à demonstrada no BD categorizado para linha de transmissão, sempre realizando inicialmente uma análise somente dos arquivos referentes à categoria, e, em seguida, incluindo um arquivo de referência utilizando as informações de um dos relatórios já existentes, a fim de facilitar a validação dos resultados.

5.7.2 Transformadores

Para a categorização de transformadores, foram utilizados os relatórios que abordam o seguinte tema:

- Energização de transformadores
- Energização de autotransformadores

A Tabela 12 apresenta o nome dos arquivos vinculados a categoria.

Tabela 12 - Relatórios categorizados para Transformadores

Numeração	***Título do relatório***
0	COPEL DELT 02-2022 - Energização Autotransformador ATF-A CEL.pdf
1	COPEL DELT 06-2021 - Energização Transformadores PGN.pdf
2	COPEL VLPT 01-2023 - Energização Autotransformador ATF-C SDI V1.pdf
3	COPEL VPEE 01-2021 - Energização Autotransformador ATF-A PGS_rev2.pdf
4	COPEL VPEE 03-2020 - Energização_rev1.pdf
5	COPEL VPEE 07-2017 - Energização Transformadores - SE 230kV Castro 2.pdf
6	COPEL VPEE 08-2016 - Energização e Religamento Monopolar - SE Andirá Leste.pdf

7	COPEL VPEE 08-2017 - Energização Transformadores - SE 230kV Guarapuava Oeste.pdf
8	COPEL VPEE 09-2016 - Energização Transformadores - SE Andirá Leste.pdf
9	COPEL VPEE 09-2017 - Energização Transformadores - SE 230kV Irati Norte.pdf
10	COPEL VPEE 10-2017 - Energização Transformadores - SE 230kV União da Vitória Norte.pdf
11	COPEL VPEE 12-2020 - Energização Autotransformador ATF-B PGS_rev2.pdf
12	COPEL VPEE 13-2021 - Pré operacional TF A SE Umbará.pdf
13	COPEL VPEE 14-2021 - Energização Autotransformador ATF-A PTO.pdf
14	estudo 13-2015 V1.pdf
15	Estudo DPEE 23-2014 - Transitórios Eletromagnéticos - LT 230kV Baixo Iguaçu - Realeza.pdf
16	RZA.pdf
17	VPEE 03-2014.pdf
18	VPEE 04-2014.pdf
19	VPEE 04-2016 Energização do Autotransformador SE Realeza Sul - V2.pdf
20	VPEE 07-2014.pdf
21	VPEE 08-2014.pdf
22	VPEE 08-2015.pdf
23	VPEE 11_2015 - Estudo Pré-operacional das LTs 230kV da SE CTL.pdf
24	VPEE 21-2016 Energização dos Autotransformadores SE Medianeira.pdf

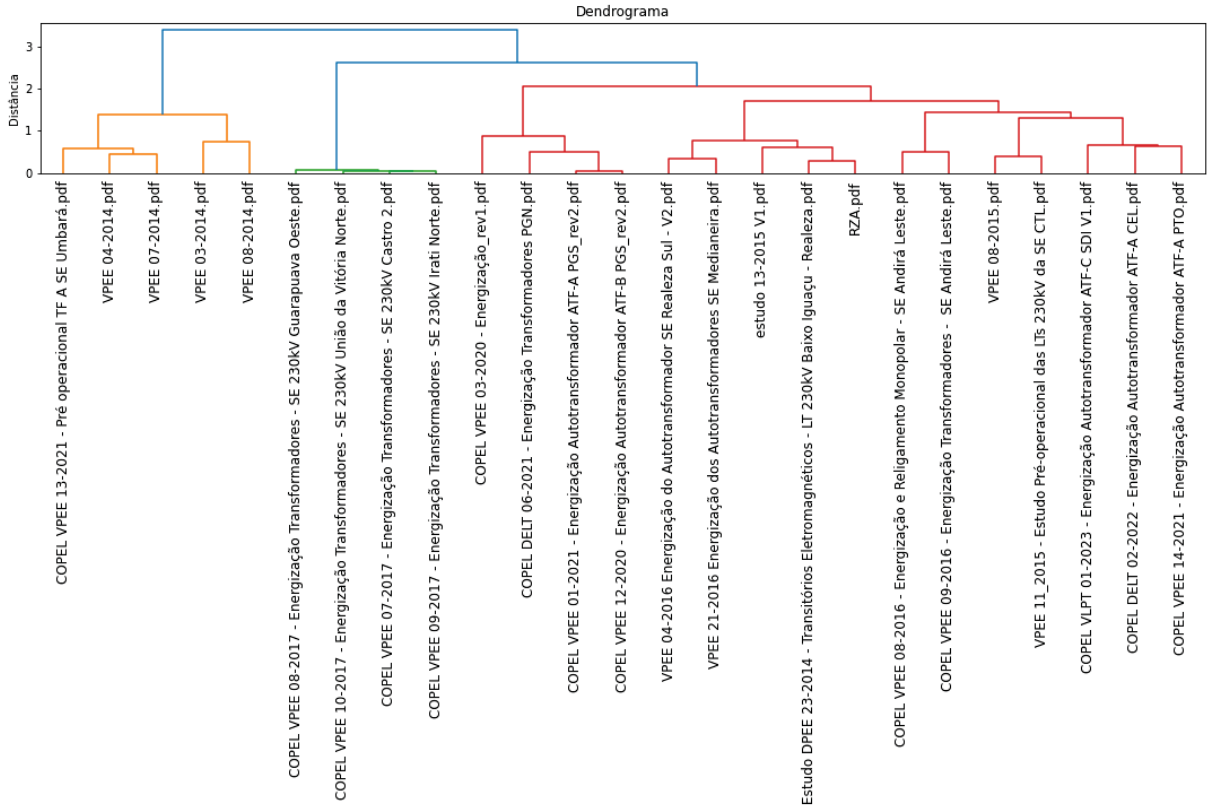
Fonte: o autor (2024).

Percebe-se que alguns dos relatórios categorizados anteriormente em linha de transmissão, aparecem novamente na categorização de transformadores. Isso ocorre pois em alguns casos, como por exemplo o relatório com arquivo nomeado “Estudo DPEE 23-2014 - Transitórios Eletromagnéticos - LT 230kV Baixo Iguaçu - Realeza.pdf”, o mesmo documento aborda tanto os estudos referentes à linha de transmissão como à energização dos transformadores.

Dessa forma é realizada a busca por similaridade entre os relatórios, tendo como resultado o dendrograma da

Figura 41.

Figura 41 - Dendrograma da busca por similaridade - categoria Transformadores

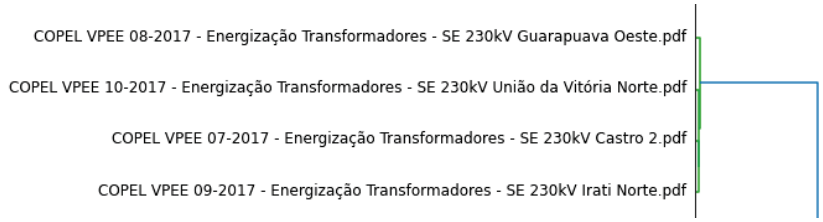


Fonte: o autor (2024).

Os principais pontos de destaque com relação aos agrupamentos apresentados são:

- Agrupamentos de energização de autotransformadores separados de agrupamentos de energização de transformadores
- Agrupamento dos relatórios numerados como 17, 18, 20 e 21 com o arquivo numerado como 12. Essa junção se dá por duas principais características desses relatórios que, apesar de estarem em padrões de escrita diferentes, a ferramenta identificou o nível de potência semelhante (150 MVA).
- O agrupamento dos relatórios numerados de 5, 7, 9 e 10 representa a similaridade regional do sistema elétrico da COPEL em que as subestações Castro 2, Guarapuava Oeste, Irati Norte e União da Vitória Norte estão localizadas, bem como estudos em transformadores de mesmo nível de tensão 230/138/13,8 kV. Este agrupamento está destacado na Figura 42.

Figura 42 - Agrupamento de relatórios de transformadores regionalizados



Fonte: o autor (2024).

Em seguida foi selecionado o documento número 1 (COPEL DELT 06-2021 - Energização Transformadores PGN.pdf), com título “ESTUDO PRÉ-OPERACIONAL TRANSITÓRIOS ELETROMAGNÉTICOS DE ENERGIZAÇÃO DOS TRANSFORMADORES 230/34,5/13,8 KV – TF-1 E TF-2 DA SUBESTAÇÃO PONTA GROSSA NORTE”, para utilização dos dados como referência de busca. A Figura 43 indica o trecho do arquivo “_INPUT”, no qual as informações foram inseridas.

Figura 43 - Informações do arquivo de referência "_INPUT" para a categoria de transformadores

RELATÓRIO TÉCNICO

Transformador

Energização de Transformadores

230/34,5/13,8 kV

50 MVA

SE Ponta Grossa Norte

Fonte: o autor (2024).

O resultado da busca por similaridade com o arquivo de referência está apresentado na Tabela 13, onde os valores representam a similaridade com relação ao arquivo “_INPUT”, o qual recebeu a numeração 25.

Tabela 13 - Valores de similaridade do BD de transformadores com o arquivo _INPUT

	25
0	19%
1	39%
2	16%
3	37%
4	28%
5	18%
6	10%
7	19%
8	11%
9	18%
10	18%
11	37%
12	19%
13	15%
14	20%
15	23%
16	25%
17	6%
18	20%
19	24%
20	23%
21	7%
22	18%
23	14%
24	25%
25	100%

Fonte: o autor (2024).

Analisando os valores de similaridade, percebe-se que, com as informações de entrada retiradas do relatório 1, temos o maior valor correspondente ao mesmo, atingindo a faixa de 39%.

Esse resultado está alinhado com a validação esperada da ferramenta, bem como representa a sua utilidade quando se estende a análise para os outros arquivos com maior similaridade, como é o caso dos relatórios 3 e 11, ambos com similaridade de 37%. O relatório 3 (COPEL VPEE 01-2021 - Energização Autotransformador ATF-A PGS_rev2.pdf) com o título “ESTUDO PRÉ-OPERACIONAL ANÁLISE DE TRANSITÓRIOS ELETROMAGNÉTICOS ENERGIZAÇÃO AUTOTRANSFORMADOR 230/138/13,8 KV – ATF-A – SUBESTAÇÃO PONTA GROSSA SUL REVISÃO 2” e o relatório 11 (COPEL VPEE 12-2020 - Energização Autotransformador ATF-B PGS_rev2.pdf) intitulado “ESTUDO PRÉ-OPERACIONAL

ANÁLISE DE TRANSITÓRIOS ELETROMAGNÉTICOS ENERGIZAÇÃO AUTOTRANSFORMADOR 230/138/13,8 KV – ATF-B – SUBESTAÇÃO PONTA GROSSA SUL REVISÃO 2", ambos possuem estudos de transformadores na subestação Ponta Grossa Sul, a qual se localiza geograficamente próxima à subestação Ponta Grossa Norte, porém o transformador em estudo no relatório 3 possui potência de 15 MVA, diferentemente da potência de 50 MVA inserida na referência e nos outros dois relatórios.

5.7.3 Disjuntores

A categorização dos relatórios de disjuntores utiliza como parâmetro apenas um tipo de estudo:

- Tensão de restabelecimento transitória (TRT)

Com base nos estudos de TRT presente nos relatórios, foi realizada a categorização do banco de dados, sendo composto pelos arquivos apresentados na Tabela 14.

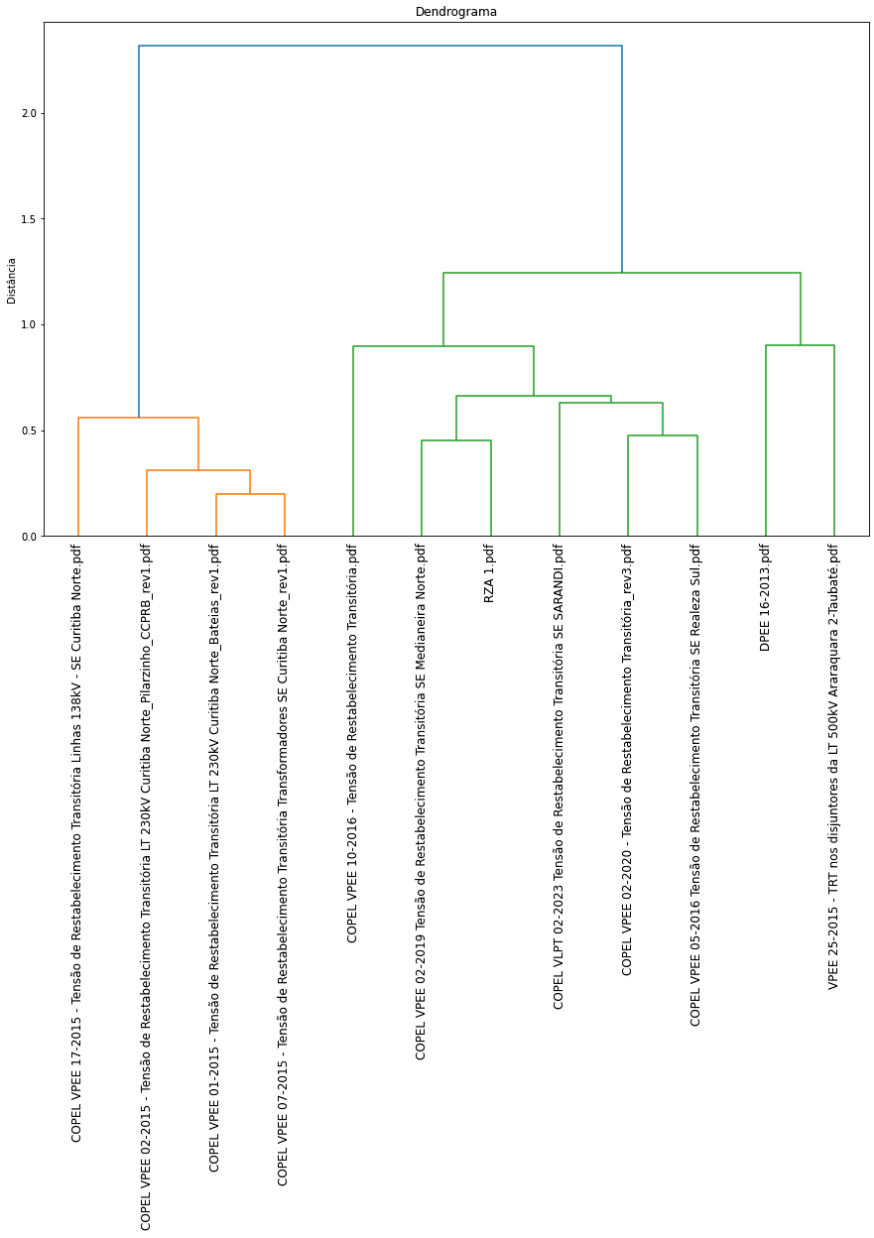
Tabela 14 - Relatórios categorizados para Disjuntores

Numeração	***Título do relatório***
0	COPEL VLPT 02-2023 Tensão de Restabelecimento Transitória SE SARANDI.pdf
1	COPEL VPEE 01-2015 - Tensão de Restabelecimento Transitória LT 230kV Curitiba Norte_Bateias_rev1.pdf
2	COPEL VPEE 02-2015 - Tensão de Restabelecimento Transitória LT 230kV Curitiba Norte_Pilarzinho_CCPRB_rev1.pdf
3	COPEL VPEE 02-2019 Tensão de Restabelecimento Transitória SE Medianeira Norte.pdf
4	COPEL VPEE 02-2020 - Tensão de Restabelecimento Transitória_rev3.pdf
5	COPEL VPEE 05-2016 Tensão de Restabelecimento Transitória SE Realeza Sul.pdf
6	COPEL VPEE 07-2015 - Tensão de Restabelecimento Transitória Transformadores SE Curitiba Norte_rev1.pdf
7	COPEL VPEE 10-2016 - Tensão de Restabelecimento Transitória.pdf
8	COPEL VPEE 17-2015 - Tensão de Restabelecimento Transitória Linhas 138kV - SE Curitiba Norte.pdf
9	DPEE 16-2013.pdf
10	RZA 1.pdf
11	VPEE 25-2015 - TRT nos disjuntores da LT 500kV Araraquara 2-Taubaté.pdf

Fonte: o autor (2024).

Uma vez categorizado o BD de disjuntores, foi realizada a busca por similaridade sem documento de referência, a fim de criar os agrupamentos no dendrograma, como mostrado na Figura 44.

Figura 44 - Dendrograma da busca por similaridade - categoria Disjuntores

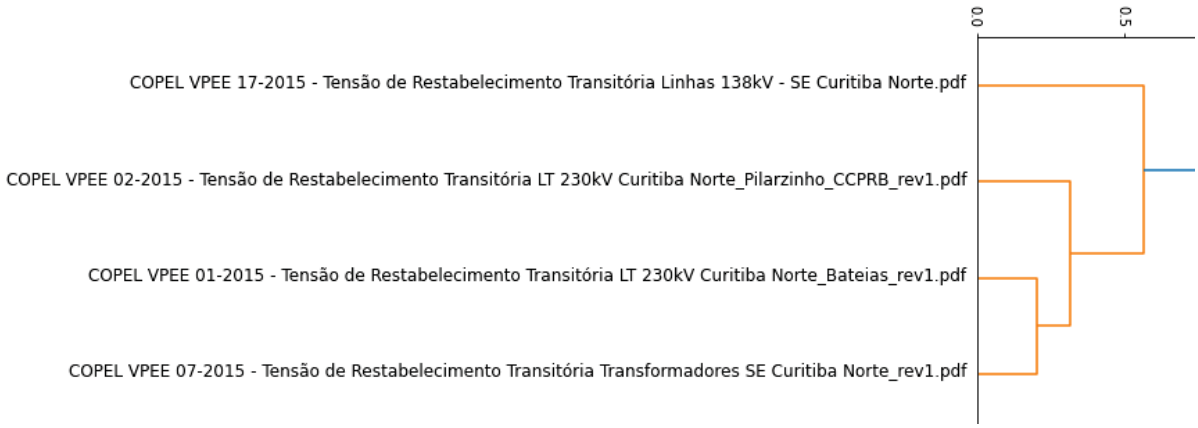


Fonte: o autor (2024).

A análise dos agrupamentos do dendrograma demonstra que, para o caso de estudos de disjuntores, considerando que a maior parte dos estudos foram realizados para o mesmo nível de tensão, de 230 kV, a localização geográfica do empreendimento foi um dos principais pontos de similaridade encontrado nos

documentos, como fica destacado no agrupamento dos relatórios 1, 2, 6 e 8. Neste caso observa-se que os 3 relatórios mais similares (1, 2 e 6) possuem o nível de tensão de 230 kV, enquanto o relatório 8 possui nível de tensão de 138 kV, porém ainda assim foi agrupado com os outros três, uma vez que também está localizado na região da subestação Curitiba Norte, conforme destacado na Figura 45.

Figura 45 - Agrupamento de relatórios de disjuntores com localização da SE Curitiba Norte



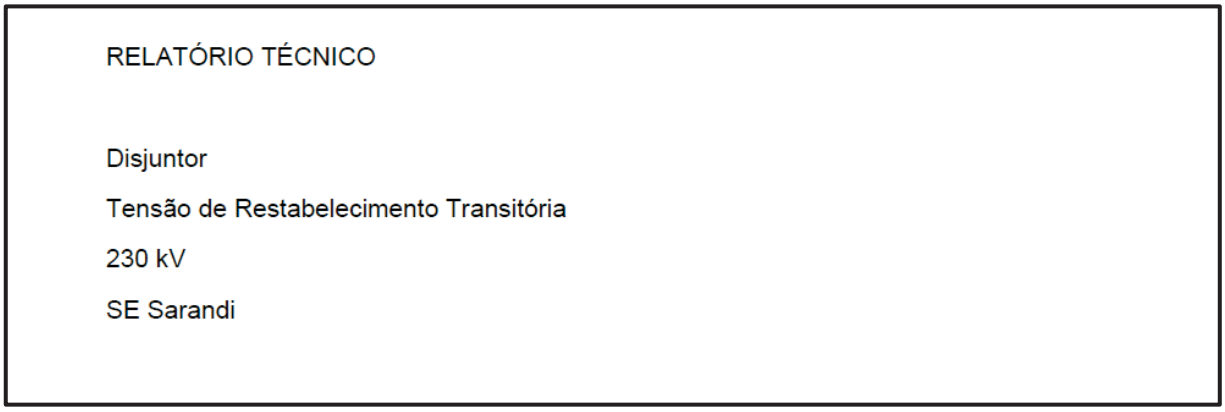
Fonte: o autor (2024).

Outro ponto de análise no dendrograma é o agrupamento dos relatórios 9 e 11, já mais distanciados dos demais, uma vez que são estudos que contemplam disjuntores com nível de tensão de 525 kV e 500 kV respectivamente, e nesse caso, mesmo com localizações geográficas distintas, a particularidade do nível de tensão foi suficiente para o seu agrupamento.

Em seguida, para realizar a análise contendo um documento de referência com as informações do estudo desejado, foram utilizadas as características apresentadas no relatório número 0 (COPEL VLPT 02-2023 Tensão de Restabelecimento Transitória SE SARANDI.pdf), com o título “RELATÓRIO TÉCNICO REFERENTE À ANÁLISE DE TRANSITÓRIOS ELETROMAGNÉTICOS – TENSÃO DE RESTABELECIMENTO TRANSITÓRIA DOS DISJUNTORES DE 230 KV E DOS GERAIS DE 138 KV DA SUBESTAÇÃO SARANDI”.

A Figura 46 apresenta o recorte do arquivo “_INPUT” e suas informações.

Figura 46 - Informações do arquivo de referência "_INPUT" para a categoria de disjuntores



Fonte: o autor (2024).

Executando mais uma vez a busca por similaridade, porém incluindo no BD o arquivo “_INPUT”, numerado como 12, são obtidos os resultados da Tabela 15.

Tabela 15 - Valores de similaridade do BD de disjuntores com o arquivo _INPUT

	12
0	42%
1	13%
2	17%
3	23%
4	28%
5	16%
6	13%
7	20%
8	21%
9	18%
10	22%
11	14%
12	100%

Fonte: o autor (2024).

Os resultados obtidos para a busca, tendo como referência o arquivo “_INPUT” demonstram novamente que a subestação de estudo é o principal fator de similaridade, uma vez que a maior parte dos estudos abordam disjuntores de 230 kV. Percebe-se também que os relatórios 9 e 11, que analisam disjuntores da ordem de 500 kV e localizações distintas da referência, apesar de apresentarem baixa similaridade, não foram os documentos com os menores valores

5.7.4 Chaves seccionadoras

Os relatórios categorizados para o BD de chaves seccionadoras contemplam o seguinte estudo:

- Tensões e Correntes induzidas em lâmina de terra

Os relatórios selecionados que contemplam o estudo estão apresentados na Tabela 16.

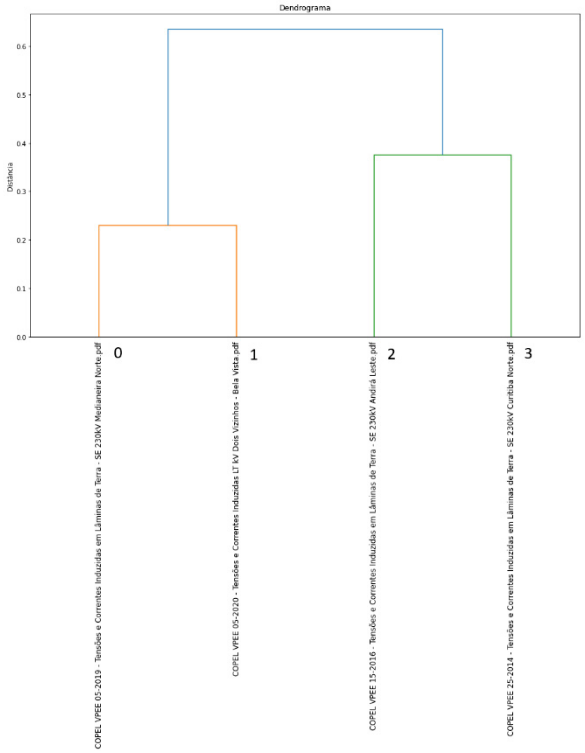
Tabela 16 - Relatórios categorizados para Chaves Seccionadoras

Numeração	***Título do relatório***
0	COPEL VPEE 05-2019 - Tensões e Correntes Induzidas em Lâminas de Terra - SE 230kV Medianeira Norte.pdf
1	COPEL VPEE 05-2020 - Tensões e Correntes Induzidas LT kV Dois Vizinhos - Bela Vista.pdf
2	COPEL VPEE 15-2016 - Tensões e Correntes Induzidas em Lâminas de Terra - SE 230kV Andirá Leste.pdf
3	COPEL VPEE 25-2014 - Tensões e Correntes Induzidas em Lâminas de Terra - SE 230kV Curitiba Norte.pdf

Fonte: o autor (2024).

Utilizando a ferramenta de busca por similaridade no banco de arquivos dos relatórios de chaves seccionadoras, realiza-se a análise do dendrograma apresentado na Figura 47.

Figura 47 - Dendrograma da busca por similaridade - categoria Chave Seccionadora



Fonte: o autor (2024).

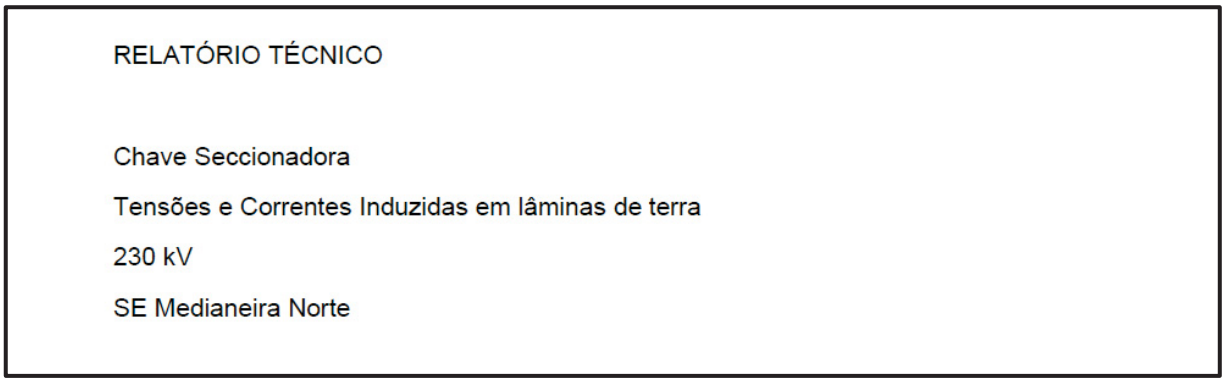
Para o caso da análise dos relatórios de estudos de chaves seccionadoras, devido à pequena quantidade de relatórios disponíveis para o presente trabalho, verifica-se que a distância entre os documentos no dendrograma é muito baixa, dificultando assim o agrupamento, como fica evidente em que a distância média entre os dois agrupamentos é de apenas 0,6.

Como consequência da baixa quantidade de arquivos e a grande similaridade entre os mesmos, temos o agrupamento do arquivo de número 0 (COPEL VPEE 05-2019 - Tensões e Correntes Induzidas em Lâminas de Terra - SE 230kV Medianeira Norte.pdf) o qual possui como título “PROJETO BÁSICO ANÁLISE DAS TENSÕES E CORRENTES INDUZIDAS EM LÂMINAS DE TERRA DE SECCIONADORAS DA SE 230KV MEDIANEIRA NORTE”, com o relatório de número 1 (COPEL VPEE 05-2020 - Tensões e Correntes Induzidas LT kV Dois Vizinhos - Bela Vista.pdf), de título “PROJETO BÁSICO ANÁLISE DAS TENSÕES E CORRENTES INDUZIDAS EM LÂMINAS DE TERRA DE SECCIONADORAS - LT 138 KV BELA VISTA - DOIS VIZINHOS”. Esse agrupamento é inesperado, pois são relatórios de níveis de tensão

distintos, considerando que os outros dois relatórios, de número 2 e 3, do outro agrupamento, ambos possuem nível de tensão de 230 kV, esperando assim que o relatório de número 0 fosse agrupado com estes.

Em seguida foi gerado o arquivo de referência “_INPUT”, de numeração 4, com as informações coincidentes com o relatório 0, conforme a seção da Figura 48.

Figura 48 - Informações do arquivo de referência "_INPUT" para a categoria de chave seccionadora



Fonte: o autor (2024).

Os resultados obtidos a partir da busca por similaridade incluindo o documento de referência são apresentados na Tabela 17 .

Tabela 17 - Valores de similaridade do BD de chaves seccionadoras com o arquivo _INPUT

	4
0	39%
1	20%
2	29%
3	28%
4	100%

Fonte: o autor (2024).

A análise dos resultados dos valores de similaridade, quando incluído um documento de referência aponta dois destaques principais:

- O relatório 0, do qual foram retiradas as informações para o arquivo “_INPUT” apresentou a maior similaridade, como esperado, representando que o nome da subestação foi o principal ponto de similaridade, uma vez que os relatórios 2 e 3 possuem as mesmas características técnicas.
- O relatório 1 apresentou a menor similaridade, tendo em vista o estudo da chave seccionadora possuir outro nível de tensão.

Apesar dos resultados apresentarem boa validação com relação ao documento de referência, vale ressaltar que a utilização da ferramenta de busca por similaridade em bancos de dados com poucos documentos não se torna relevante, uma vez que é viável a verificação manual de cada um dos relatórios.

5.7.5 Bancos de Capacitores

Para a categorização do último banco de dados referente ao estudos em bancos de capacitores, foi utilizado apenas um tipo de abordagem:

- Energização de banco de capacitores

Assim como para a categorização do BD de chaves seccionadoras, o BD de bancos de capacitores possui poucos relatórios que apresentam os estudos, sendo eles apresentados na Tabela 18.

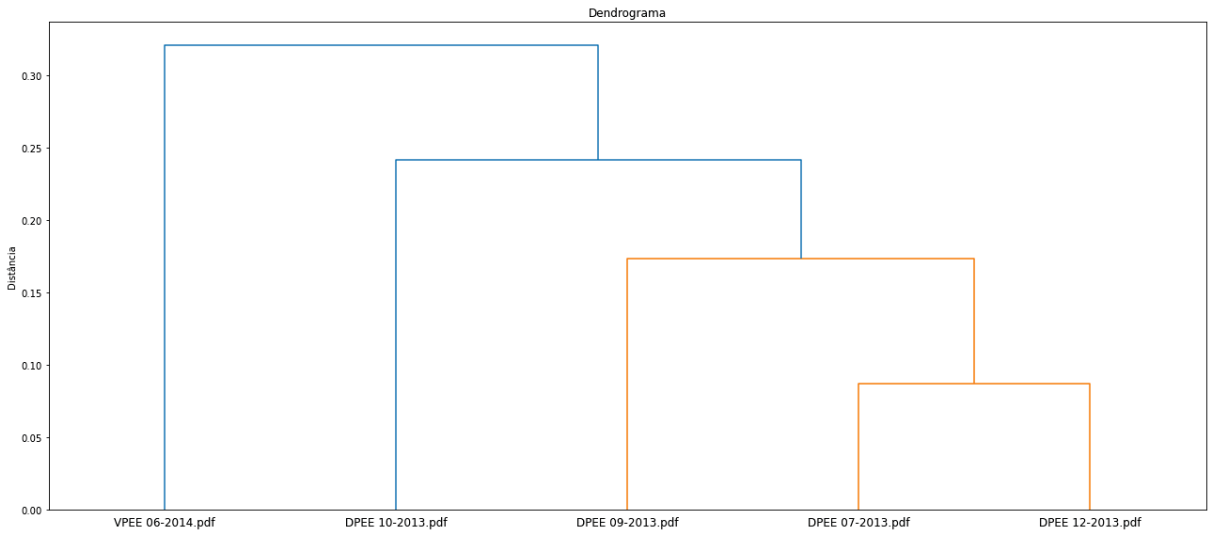
Tabela 18 - Relatórios categorizados para Bancos de Capacitores

0	1
0	DPEE 07-2013.pdf
1	DPEE 09-2013.pdf
2	DPEE 10-2013.pdf
3	DPEE 12-2013.pdf
4	VPEE 06-2014.pdf

Fonte: o autor (2024).

Após a categorização, é executada a busca por similaridade entre os relatórios, com o objetivo de verificar as similaridades entre os documentos através do dendrograma da Figura 49.

Figura 49 - Dendrograma da busca por similaridade - categoria Banco de Capacitores



Fonte: o autor (2024).

A análise do dendrograma demonstra a capacidade da ferramenta de criar um único agrupamento de todos os arquivos, porém identificar a maior semelhança entre o relatório 0 (DPEE 07-2013.pdf) com título “ESTUDO DE TRANSITÓRIOS ELETROMAGNÉTICOS DE ENERGIZAÇÃO DOS BANCOS DE CAPACITORES DE 2 X 30 MVAR E DE 15 MVAR DA SUBESTAÇÃO MARINGÁ” e o relatório 3 (DPEE 12-2013.pdf) com título “ESTUDO DE TRANSITÓRIOS ELETROMAGNÉTICOS DE ENERGIZAÇÃO DOS BANCOS DE CAPACITORES DE 30 MVAR E DE 15 MVAR DA SUBESTAÇÃO LONDRINA”, os quais de tratam do mesmo tipo de estudo realizados em subestações diferentes.

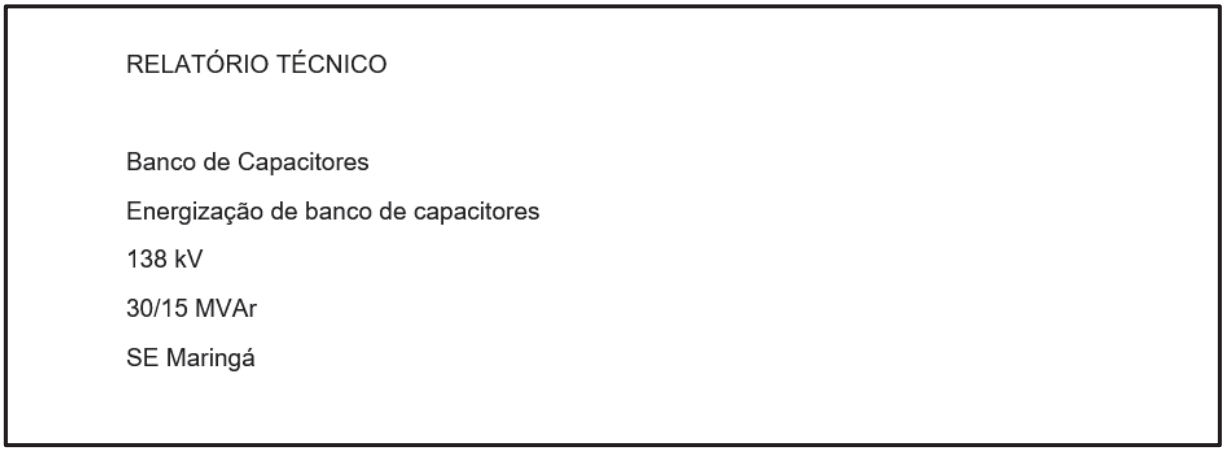
Em seguida temos o relatório 1 (DPEE 09-2013.pdf) como sendo o mais próximo, o qual aborda apenas banco de capacitores de 15 MVAr da subestação Sarandi, seguido pelo relatório 2 (DPEE 10-2013.pdf) com estudos apenas de banco de capacitores de 30 MVAr. Por fim o relatório 4 (VPEE 06-2014.pdf), por abordar bancos de capacitores de 50 MVAr da subestação de Cerquilho, foi considerado o documento mais distante dos demais.

Por se tratar de um banco de dados com poucos arquivos é possível verificar e validar esses resultados de forma manual.

O arquivo “_INPUT” gerado para a busca com documento de referência incluiu os dados do relatório 0 (DPEE 07-2013.pdf) com título “ESTUDO DE TRANSITÓRIOS ELETROMAGNÉTICOS DE ENERGIZAÇÃO DOS BANCOS DE CAPACITORES DE

2 X 30 MVAR E DE 15 MVAR DA SUBESTAÇÃO MARINGÁ”, como mostra a Figura 50.

Figura 50 - Informações do arquivo de referência "_INPUT" para a categoria de banco de capacitores



Fonte: o autor (2024).

Incluindo o arquivo “_INPUT”, recebendo a numeração de arquivo 5, ao BD é realizada novamente a busca por similaridade, obtendo os valores expressos na Tabela 19.

Tabela 19 - Valores de similaridade do BD de banco de capacitores com o arquivo _INPUT

	5
0	65%
1	61%
2	57%
3	63%
4	57%
5	100%

Fonte: o autor (2024).

Os resultados da segunda busca apontam corretamente para a maior similaridade com o relatório 0, e seguido pelos relatórios 3 e 1, em conformidade com os resultados antes da inclusão do documento de referência. Já os relatórios que mais divergem, por questões da potência do banco de capacitor abordado no estudo. Ao realizar uma análise do resultado conjunto, verifica-se que os valores de similaridade são muito próximos, o que representa que o banco de dados é composto por documentos muito padronizados de forma que, mesmo com características variadas, os valores de similaridade são próximos.

5.8 Discussão dos resultados

Após a análise individual de cada um dos bancos de dados categorizados, os quais se diferenciaram em termos de assuntos abordados, temas de estudos, e quantidade de relatórios disponíveis para as buscas, foi possível verificar um incremento na acuracidade dos resultados obtidos tendo como foco que, para os especialistas responsáveis pelo desenvolvimento dos estudos na Copel, o parâmetro principal para ganhos de eficiência na realização e novos estudos é o componente de estudo dos relatórios utilizados como referência.

Através da realização das buscas por similaridade sem o documento com as informações de referência para a busca, ou seja, somente com os arquivos dos relatórios já desenvolvidos anteriormente, o principal ganho de produtividade se dá no mapeamento do banco de dados ilustrado pelo dendrograma. Esse mapeamento se mostrou eficaz nas situações em que o especialista já possui um ou mais documentos onde já tem o conhecimento de serem utilizados como base para um novo estudo. Dessa forma, a partir desse documento já pré-identificado, o especialista consegue facilmente selecionar os relatórios mais similares ao mesmo, o que torna sua análise de referência mais aprofundada e com possibilidades de identificação não selecionadas a priori.

Já para os casos em que o especialista não possui nenhum relatório em mente para servir como base em sua busca, a facilidade de criação do arquivo “_INPUT” através da interface da ferramenta, em que, com apenas algumas poucas informações essenciais do novo empreendimento, é possível, em um tempo relativamente curto, identificar conjuntos de documentos que possuem informações relevantes que impedem o retrabalho exaustivo, o qual é comumente uma consequência da falta de conhecimento da base de dados da corporação.

Para uma comparação dos resultados das buscas com o BD categorizado e não categorizado, é possível tomar como base a Figura 34 em que a leitura do dendrograma com 72 arquivos chega a um limite de análise, a qual fica dificultada devido a quantidade de agrupamentos em um mesmo gráfico. Além disso, a verificação dos agrupamentos sinaliza a dificuldade em correlacionar todos os documentos seguindo um só parâmetro.

Já a categorização do banco de dados, apesar de se mostrar mais eficiente computacionalmente, uma vez que analisa uma quantidade menor de arquivos, e

também mais assertiva, tendo em vista que é forçada a realizar a busca já seguindo o parâmetro dos componentes de estudo, está sujeita à equívocos humanos durante a categorização manual dos documentos em cada BD. Outro ponto de desvantagem na categorização dos relatórios é a perda da capacidade da ferramenta em verificar similaridades não identificadas entre documentos de categorias distintas, tendo assim a possibilidade de perda de informações relevantes, principalmente com relação a similaridade de localizações.

6 CONCLUSÕES E TRABALHOS FUTUROS

A análise dos estudos elétricos necessários para novos empreendimentos de linhas de transmissão em conjunto com a elaboração da ferramenta de busca por similaridade durante o período em que o presente trabalho foi desenvolvido, demonstrou a variedade de opções de segmentos que o tema poderia se encaminhar, bem como a complexidade dos estudos referentes aos componentes que compõem um projeto básico de linhas de transmissão.

Tendo como exemplo as diferentes técnicas para o cálculo de similaridade, e as diferentes ferramentas disponíveis para validações e testes, a decisão em utilizar a linguagem de programação Python tornou possível a compreensão mais aprofundada de técnicas de programação, bem como a otimização de tarefas complexas, como a implementação de técnicas de NLP, através de bibliotecas que, uma vez estudadas e avaliadas, apresentaram bons resultados iniciais.

O estudo das técnicas de Processamento de Linguagem Natural (NLP) revelou-se fundamental para a busca por similaridade de documentos no contexto da engenharia elétrica. Ao longo deste estudo, ficou evidente que as abordagens baseadas em NLP oferecem um arcabouço robusto e versátil para compreender e comparar a semântica subjacente aos documentos técnicos, otimizando a identificação de padrões, estruturas e informações relevantes. A aplicação dessas técnicas não apenas permitiu uma análise mais precisa e eficiente de vastos conjuntos de relatórios, mas também desempenhou um papel crucial na evolução de ferramenta para a recuperação de informações e na própria eficiência de projetos.

Por fim, a avaliação e validação dos resultados obtidos com a ferramenta desenvolvida no presente trabalho apresentou conceitos relacionados à gestão de conhecimentos, similar ao trabalho de Liu et al. (2019), porém com a aplicação a relatórios de projetos ao invés de bilhetes de manutenção. Ainda de forma similar a este trabalho, a utilização das mesmas técnicas de vetorização textual e cálculo de similaridade se mostraram eficazes para a aplicação em questão.

Já com relação ao trabalho de McNamee (2013), em que são aplicadas metodologias de taxonomia especificas para patentes, pode ser considerado um tema para trabalhos futuros, no qual são aplicadas as mesmas técnicas de busca, porém com a possibilidade de personalização da vetorização textual para o cenário de

relatórios de estudos elétricos, com o objetivo de oferecer benefícios na precisão das buscas por similaridade entre os documentos.

A utilização da ferramenta de busca por similaridade em bancos de dados de diferentes tamanhos mostrou que existe um intervalo na quantidade de documentos analisados que pode ser considerado como ótimo, uma vez que com uma quantidade muito grande, acima de 70 arquivos, o custo computacional cresce exponencialmente e os agrupamentos secundários podem apresentar erros, enquanto com quantidade muito pequenas, abaixo de 5 documentos, demonstrou resultados insatisfatórios para agrupamentos, uma vez que a análise de padrões é muito reduzida.

A partir da apresentação e validação dos resultados de todos os casos de estudo desenvolvidos no presente trabalho, foi possível concluir que a ferramenta de gestão de conhecimento por meio de busca por similaridade pode oferecer muitos benefícios aos especialistas usuários, uma vez que através dela torna-se possível uma visão mais abrangente e precisa a respeito do conhecimento técnico armazenado nos bancos de dados das corporações.

Outro benefício identificado com o uso de ferramentas de busca por similaridade é a capacidade de padronização dos novos documentos uma vez que o uso de relatórios anteriores, os quais já foram desenvolvidos de acordo com as diretrizes, garantem que uma consistência e alinhamento para os demais.

REFERÊNCIAS

AKMAL, S.; SHIH, L. H.; BATRES, R. Ontology-based similarity for product information retrieval. **Computers in Industry**, 2014. Elsevier. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0166361513001590>. .

AMIN, K.; LANCASTER, G.; KAPETANAKIS, S.; ALTHOFF, K. D.; ... Advanced similarity measures using word embeddings and siamese networks in CBR. **… SAI Intelligent Systems …**, 2019. Springer. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-030-29513-4_32>. .

API, N. gensim.similarities.Similarity. Disponível em: <https://tedboy.github.io/nlps/generated/generated/gensim.similarities.Similarity.html>. .

ARTS, S.; HOU, J.; GOMEZ, J. C. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. **Research Policy**, 2021. Elsevier. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0048733320302195>. .

DORNELES, C. F. Similaridade de Dados. , 2022.

ENSSLIN, L.; ENSSLIN, S. R.; LACERDA, R. T. O.; TASCA, J. E. Processo de Selação de Portifólio Bibliográfico. **Processo técnico com patente de registro pendente junto ao INPI**, 2010.

FREDIGO, A.; HACKS, L.; NUNES, F.; et al. Text Mining - Data Mining. , p. 12, 2013. Disponível em: <http://www.inf.ufsc.br/~luis.alvares/INE5644/G2_texto.pdf>. .

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. 3º ed. Stanford, 2023.

KRIGER, D. O QUE É PYTHON, PARA QUE SERVE E POR QUE APRENDER? Disponível em: <https://kenzie.com.br/blog/o-que-e-python/>. .

LI, S.; WEN, J. Application of pattern matching method for detecting faults in air handling unit system. **Automation in Construction**, 2014. Elsevier. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0926580514000545>. .

LIAN, J.; ZHANG, B.; ZHANG, J.; et al. Research on the Application of Natural Language Processing in the Virtual Comment's Classification. **Artificial Intelligence in …**, 2022. Springer. Disponível em: <https://link.springer.com/chapter/10.1007/978-981-16-9423-3_22>. .

LIU, T.; LI, S.; GU, X.; et al. Historical Similar Ticket Matching and Extraction used for Power Grid Maintenance Work Ticket Decision Making. **… Conference on Data …**, 2019. ieeexplore.ieee.org. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9270166/>. .

LU, Z. K.; YE, J. Cosine similarity measure between vague sets and its application of fault diagnosis. **Research Journal of Applied Sciences, Engineering …**, 2013. Citeseer. Disponível em: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4dc4a08449d17d43ffe5763cff7d5edf7dbd534f>. .

MATPLOTLIB. matplotlib.pyplot. Disponível em: <https://matplotlib.org/stable/api/pyplot_summary.html>. .

MCNAMEE, R. C. Can't see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example. **Research policy**, 2013. Elsevier. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0048733313000115>. .

MODE. Pandas tutorials. Disponível em: <https://mode.com/python-tutorial/libraries/pandas/>. .

NI, X.; SAMET, A.; CAVALLUCCI, D. Similarity-based approach for inventive design solutions assistance. **Journal of Intelligent Manufacturing**, 2022. Springer. Disponível em: <https://link.springer.com/article/10.1007/s10845-021-01749-4>. .

Oreilly. Statistics for Machine Learning by Pratap Dangeti, 2023. Disponível em: <https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/eb9cd609-e44a-40a2-9c3a-f16fc4f5289a.xhtml>. .

PEREIRA, V. Python: O que é, quais as vantagens e onde aplicar? Disponível em: <https://king.host/blog/2021/03/python/#:~:text=É possível afirmar que Python,coleta e análise de dados>. .

PYPI. gensim 4.3.1. Disponível em: <https://pypi.org/project/gensim/>. .

RAVULAKOLLU, N. Introduction to Python Pandas for Beginners. Disponível em: <https://www.almabetter.com/bytes/articles/introduction-to-python-pandas-for-beginners>. .

ŘEHŮŘEK, R. corpora.dictionary – Construct word<->id mappings. Disponível em: <https://radimrehurek.com/gensim/corpora/dictionary.html>. .

ŘEHŮŘEK, R. similarities.docsim – Document similarity queries. Disponível em: <https://radimrehurek.com/gensim/similarities/docsim.html>. .

THE PYTHON SOFTWARE FOUNDATION. glob — Unix style pathname pattern expansion. Disponível em: <https://docs.python.org/3/library/glob.html>. .

THE PYTHON SOFTWARE FOUNDATION. re — Regular expression operations. Disponível em: <https://docs.python.org/3/library/re.html>. .

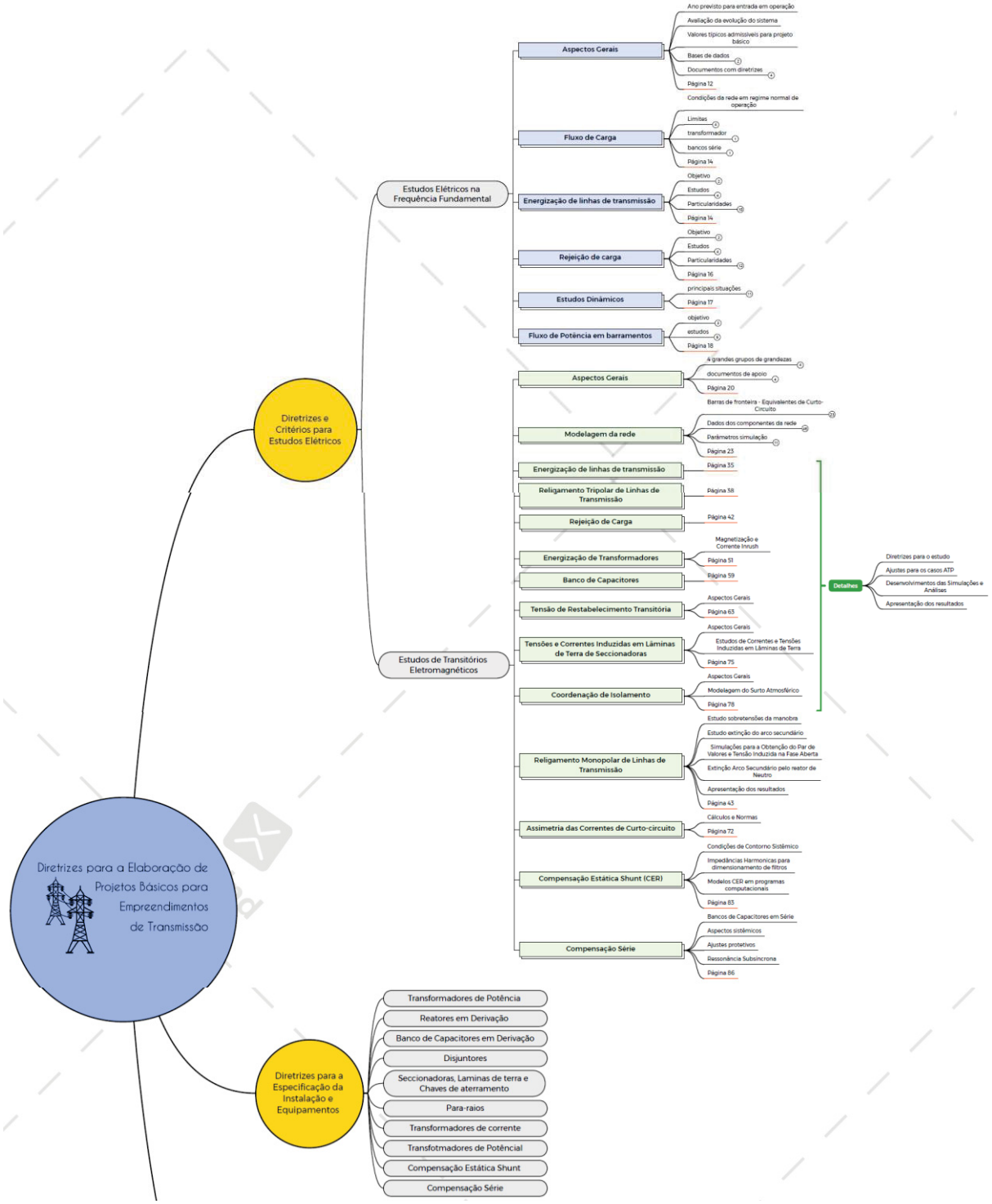
THE PYTHON SOFTWARE FOUNDATION. os — Miscellaneous operating system interfaces. Disponível em: <https://docs.python.org/3/library/os.html>. .

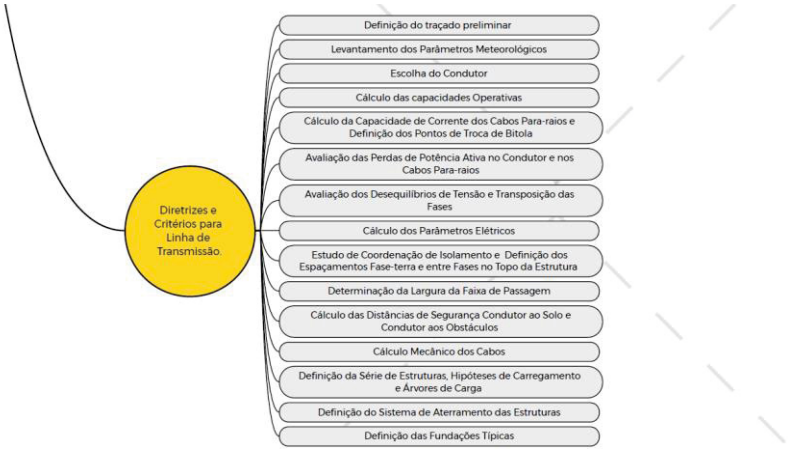
THE SCIPY. Clustering package (scipy.cluster). Disponível em: <https://docs.scipy.org/doc/scipy/reference/cluster.html#module-scipy.cluster>. .

UFRGS. Cálculo de similaridade. Disponível em: <https://www.ufrgs.br/wiki-r/index.php?title=Cálculo_de_similaridade#:~:text=É o comprimento se seguimento,(distância entre dois pontos).>. .

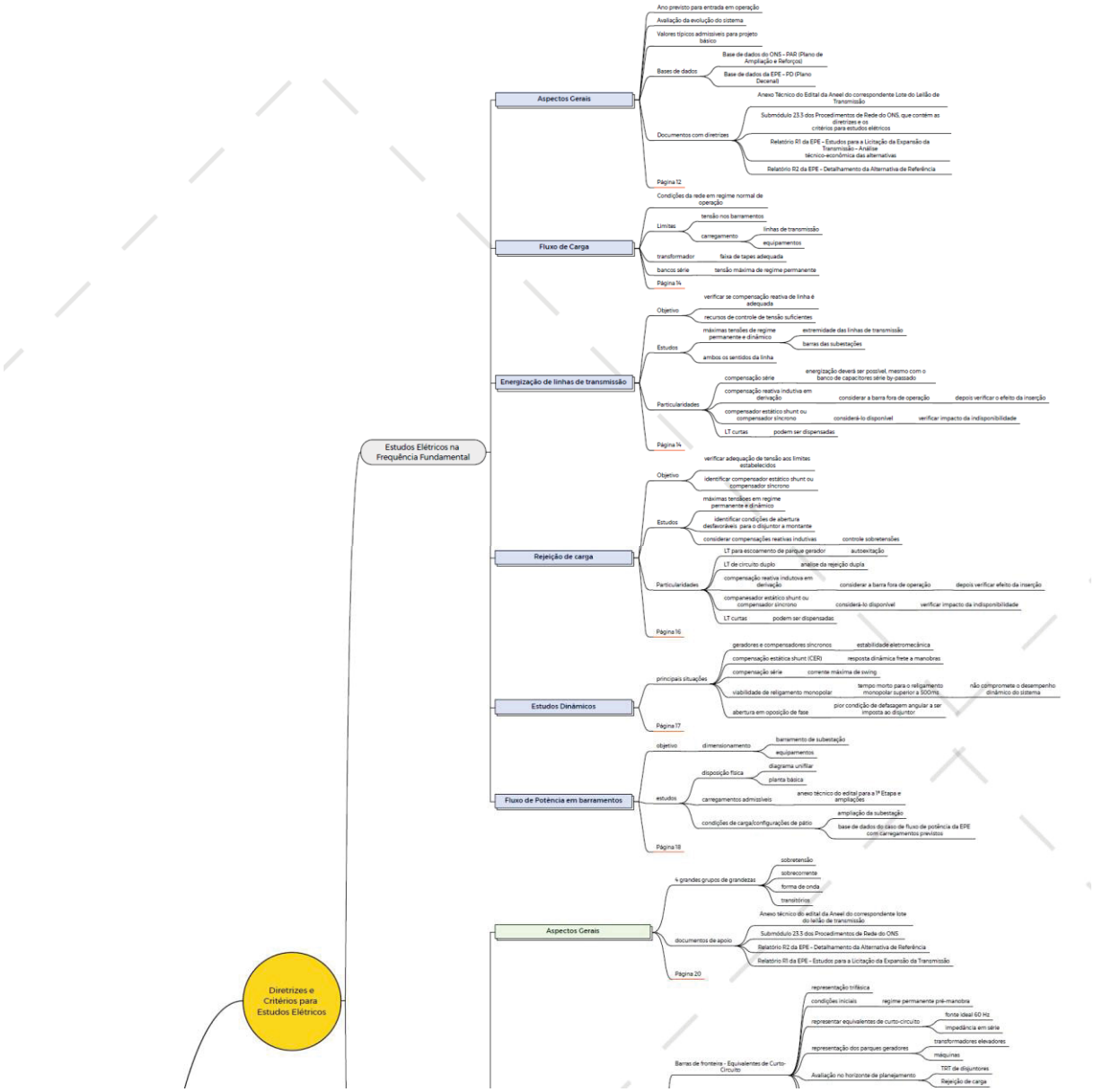
ANEXOS

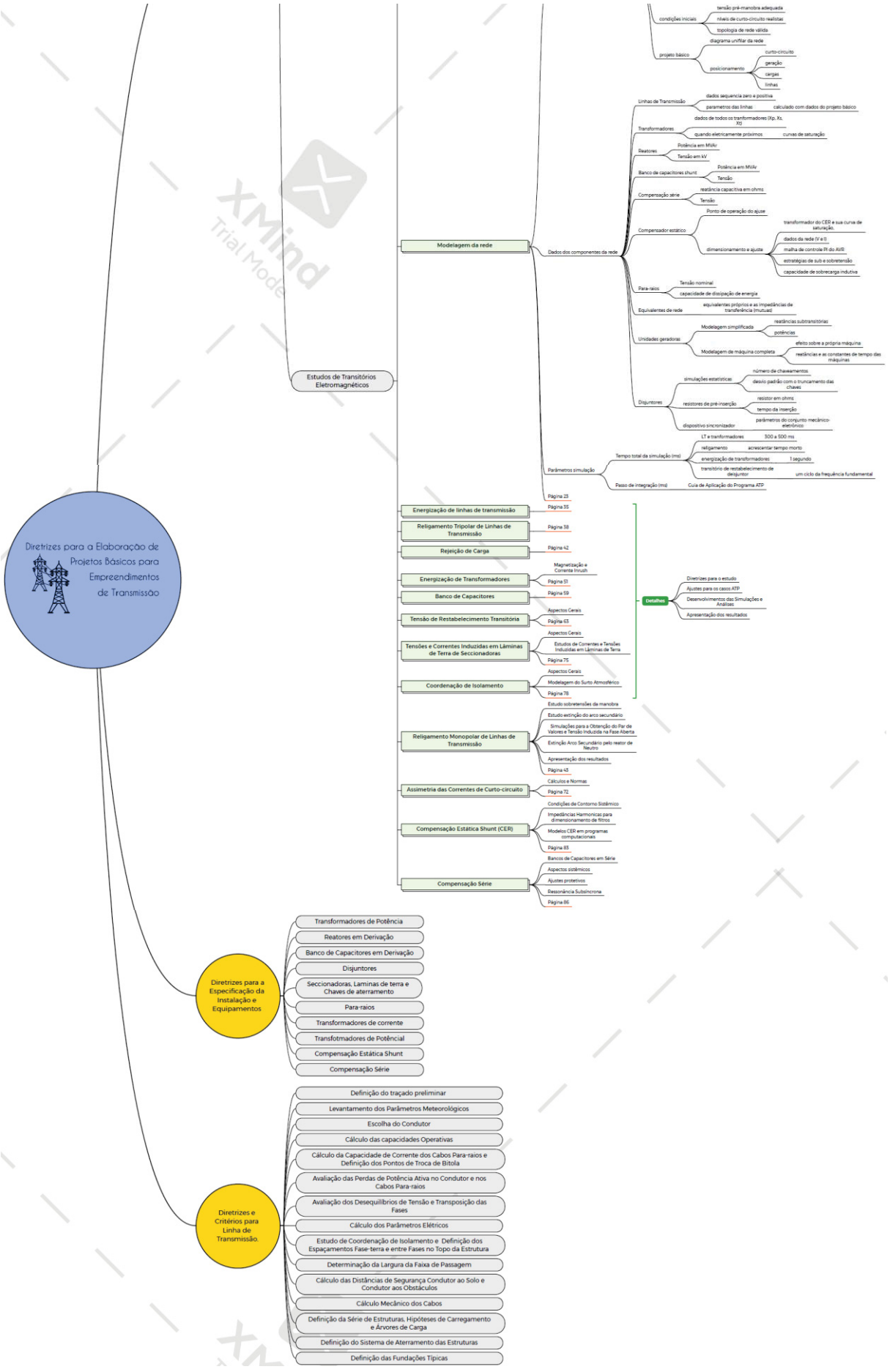
ANEXO 1 – Sumário Diretrizes ONS reduzido




Diretrizes e Critérios para Linha de Transmissão.
Definição do traçado preliminar
Levantamento dos Parâmetros Meteorológicos
Escolha do Condutor
Cálculo das capacidades Operativas
Cálculo da Capacidade de Corrente dos Cabos Para-raios e Definição dos Pontos de Troca de Bitola
Avaliação das Perdas de Potência Ativa no Condutor e nos Cabos Para-raios
Avaliação dos Desequilíbrios de Tensão e Transposição das Fases
Cálculo dos Parâmetros Elétricos
Estudo de Coordenação de Isolamento e Definição dos Espaçamentos Fase-terra e entre Fases no Topo da Estrutura
Determinação da Largura da Faixa de Passagem
Cálculo das Distâncias de Segurança Condutor ao Solo e Condutor aos Obstáculos
Cálculo Mecânico dos Cabos
Definição da Série de Estruturas, Hipóteses de Carregamento e Árvores de Carga
Definição do Sistema de Aterramento das Estruturas
Definição das Fundações Típicas

ANEXO 2 – Sumário Diretrizes ONS completo




Diretrizes para a Elaboração de Projetos Básicos para Empreendimentos de Transmissão
Modelagem da rede
Dados dos componentes da rede
condições iniciais
tensão pré-manobra adequada
níveis de curto-circuito realistas
topologia de rede válida
projeto básico
diagrama unifilar da rede
posicionamento
curto-circuito
geração
cargas
linhas
Linhas de Transmissão
dados sequencia zero e positiva
parametros das linhas
calculado com dados do projeto básico
Transformadores
dados de todos os tranformadores (Rp, Xs, Xt)
quando eletricamente próximos
curvas de saturação
Reatores
Potência em MVAr
Tensão em kV
Banco de capacitores shunt
Potência em MVAr
Tensão
Compensação série
reatância capacitiva em ohms
Tensão
Compensador estático
Ponto de operação do ajuse
dimensionamento e ajuste
transformador do CER e sua curva de saturação.
dados da rede (V e I)
malha de controle PI do AVR
estratégias de sub e sobretensão
capacidade de sobrecarga indutiva
Para-raios
Tensão nominal
capacidade de dissipação de energia
Equivalentes de rede
equivalentes próprios e as impedâncias de transferência (mutuas)
Unidades geradoras
Modelagem simplificada
reatâncias subtransitórias
potências
Modelagem de máquina completa
efeito sobre a própria máquina
reatâncias e as constantes de tempo das máquinas
Disjuntores
simulações estatísticas
número de chaveamentos
desvio padrão com o truncamento das chaves
resistores de pré-inserção
resistor em ohms
tempo da inserção
dispositivo sincronizador
parâmetros do conjunto mecânico-eletrônico
Parâmetros simulação
Tempo total da simulação (ms)
LT e tranformadores
300 a 500 ms
religamento
acrescentar tempo morto
energização de transformadores
1 segundo
transitório de restabelecimento de disjuntor
um ciclo da frequência fundamental
Passo de integração (ms)
Cula de Aplicação do Programa ATP
Estudos de Transitórios Eletromagnéticos
Página 23
Energização de linhas de transmissão
Página 35
Religamento Tripolar de Linhas de Transmissão
Página 38
Rejeição de Carga
Página 42
Energização de Transformadores
Magnetização e Corrente Inrush
Página 51
Banco de Capacitores
Página 59
Tensão de Restabelecimento Transitória
Aspectos Gerais
Página 63
Tensões e Correntes Induzidas em Lâminas de Terra de Seccionadoras
Aspectos Gerais
Estudos de Correntes e Tensões Induzidas em Lâminas de Terra
Página 75
Coordenação de Isolamento
Aspectos Gerais
Modelagem do Surto Atmosférico
Página 78
Detalhes
Diretrizes para o estudo
Ajustes para os casos ATP
Desenvolvimentos das Simulações e Análises
Apresentação dos resultados
Religamento Monopolar de Linhas de Transmissão
Estudo sobretensões da manobra
Estudo extinção do arco secundário
Simulações para a Obtenção do Par de Valores e Tensão Induzida na Fase Aberta
Extinção Arco Secundário pelo reator de Neutro
Apresentação dos resultados
Página 43
Assimetria das Correntes de Curto-circuito
Cálculos e Normas
Página 72
Compensação Estática Shunt (CER)
Condições de Contorno Sistêmico
Impedâncias Harmonicas para dimensionamento de filtros
Modelos CER em programas computacionais
Página 83
Compensação Série
Bancos de Capacitores em Série
Aspectos sistêmicos
Ajustes protetivos
Ressonância Subsincrona
Página 86
Diretrizes para a Especificação da Instalação e Equipamentos
Transformadores de Potência
Reatores em Derivação
Banco de Capacitores em Derivação
Disjuntores
Seccionadoras, Laminas de terra e Chaves de aterramento
Para-raios
Transformadores de corrente
Transfotmadores de Potêncial
Compensação Estática Shunt
Compensação Série
Diretrizes e Critérios para Linha de Transmissão.
Definição do traçado preliminar
Levantamento dos Parâmetros Meteorológicos
Escolha do Condutor
Cálculo das capacidades Operativas
Cálculo da Capacidade de Corrente dos Cabos Para-raios e Definição dos Pontos de Troca de Bitola
Avaliação das Perdas de Potência Ativa no Condutor e nos Cabos Para-raios
Avaliação dos Desequilíbrios de Tensão e Transposição das Fases
Cálculo dos Parâmetros Elétricos
Estudo de Coordenação de Isolamento e Definição dos Espaçamentos Fase-terra e entre Fases no Topo da Estrutura
Determinação da Largura da Faixa de Passagem
Cálculo das Distâncias de Segurança Condutor ao Solo e Condutor aos Obstáculos
Cálculo Mecânico dos Cabos
Definição da Série de Estruturas, Hipóteses de Carregamento e Árvores de Carga
Definição do Sistema de Aterramento das Estruturas
Definição das Fundações Típicas