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

Automatização da Investigação Patrimonial para Recuperação de Ativos: Identificação de Grupos Econômicos

Francisco Gonçalves de Araújo Filho

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Edison Ishikawa

Brasília
2025

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Gonçalves de Araújo Filho, Francisco
G635a Automatização da Investigação Patrimonial para Recuperação de Ativos: Identificação de Grupos Econômicos / Francisco Gonçalves de Araújo Filho; orientador Edison Ishikawa. Brasília, 2025.
79 p.

Dissertação(Mestrado Profissional em Computação Aplicada) Universidade de Brasília, 2025.

1. Grupos econômicos – Identificação automatizada. 2. Recuperação de ativos – Investigações patrimoniais. 3. Grafos – Aplicações em Ciência de Dados. 4. Redes neurais em grafos (GNN) – Modelagem e predição. 5. Processamento de grandes volumes de dados – Apache Spark. I. Ishikawa, Edison, orient. II. Título.

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

Automatização da Investigação Patrimonial para Recuperação de Ativos: Identificação de Grupos Econômicos

Francisco Gonçalves de Araújo Filho

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Edison Ishikawa (Orientador)
CIC/UnB

Prof. Dr. Daniel Falcão
IDP

Prof. Dr. Márcio Victorino
UnB

Prof. Dr. Marcelo Ladeira
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 27 de agosto de 2025

Dedicatória

Dedico este trabalho aos meus pais, **Francisco Gonçalves de Araújo** (*in memoriam*) e **Ana Alves Gonçalves de Araújo**, que não pouparam esforços para oferecer aos filhos uma formação sólida e os valores que nos orientam.

Ofereço também à minha família, que me apoia, ampara e fortalece nos momentos mais desafiadores, sendo fonte constante de incentivo e inspiração para seguir adiante.

Agradecimentos

Agradeço primeiramente a Deus pela minha existência e por me conceder forças ao longo desta jornada. A **São José de Cupertino**, pela incansável intercessão em favor deste filho indigno.

Ao meu professor orientador, pela paciência, dedicação e pela valorosa orientação prestada ao longo de todo o processo. Aos professores da banca examinadora, que gentilmente dedicaram parte de seu tempo para avaliar e contribuir com este trabalho, enriquecendo-o com observações e sugestões de grande valor acadêmico.

Aos colegas do Departamento de Tecnologia da Informação do Conselho Nacional de Justiça, em especial ao **Otávio**, pelo apoio e auxílio técnico no ambiente de laboratório, sempre se dispondo a colaborar na superação de dificuldades operacionais.

Aos Magistrados e Servidores do Conselho Nacional de Justiça e do Poder Judiciário que se dispuseram a avaliar os resultados deste estudo, oferecendo percepções valiosas para a consolidação do trabalho. Em especial à **Dra. Luciana Dória de Medeiros Chaves**, cujo empenho e articulação foram essenciais para a formação do grupo de avaliadores, permitindo uma análise abrangente e qualificada dos resultados alcançados.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

De acordo com o Relatório Justiça em Números 2024, o maior gargalo do Judiciário Brasileiro é a Fase de Execução dos Processos Judiciais. O Conselho Nacional de Justiça (CNJ) disponibilizou o Sistema Nacional de Investigação Patrimonial e Recuperação de Ativos (SNIPER) para acelerar a pesquisa patrimonial por meio de consultas ao Cadastro Nacional de Pessoas Jurídicas (CNPJ), auxiliando peritos na localização de Grupos Econômicos. Com o SNIPER, a identificação de grupos econômicos *de direito* é imediata, mas a detecção de empresas que possivelmente compõem grupos econômicos *de fato* ainda demanda esforço por parte do usuário, pois essas conexões não estão formalmente documentadas. O objetivo deste trabalho é automatizar a localização de grupos econômicos *de fato*, reduzindo o tempo e o trabalho exigidos dos especialistas, além de ampliar a precisão e a abrangência das investigações patrimoniais. Para isso, empregaram-se técnicas de processamento paralelo com *Apache Spark, GraphX, GraphFrames e modelos baseados em Bidirectional Encoder Representations from Transformers (BERT) para Similaridade de Sentenças*, a fim de identificar características comuns entre Entidades Jurídicas que possam indicar a formação de um grupo econômico, tornando o processo de identificação mais rápido e efetivo. Os resultados obtidos mostraram que a solução proposta foi capaz de revelar conexões não mapeadas previamente no SNIPER, ampliando o conjunto de empresas relacionadas e evidenciando vínculos relevantes para a investigação patrimonial. A avaliação qualitativa realizada por dois especialistas em recuperação de ativos indicou ganhos em agilidade e efetividade, reforçando o potencial da ferramenta para apoiar a tomada de decisão e favorecer a aplicação de medidas jurídicas voltadas à responsabilização solidária e à desconsideração da personalidade jurídica.

Palavras-chave: Mineração de Dados, Processamento de Linguagem Natural, Análise de Grafos, Similaridade de Sentenças, Apache Spark, BERT, Grupos Econômicos, Recuperação de Ativos

Abstract

According to the Justice in Numbers 2024 report, the main bottleneck in the Brazilian Judiciary lies in the Enforcement Phase of judicial proceedings. The Conselho Nacional de Justiça (CNJ) has made available the Sistema Nacional de Investigação Patrimonial e Recuperação de Ativos (SNIPER) system to expedite asset tracing through Cadastro Nacional de Pessoas Jurídicas (CNPJ) queries, assisting experts in identifying corporate groups. While the SNIPER enables the immediate identification of *de jure* corporate groups, detecting companies that may constitute *de facto* corporate groups still requires substantial user effort, as these connections are not formally documented. This work aims to automate the identification of *de facto* corporate groups, reducing the time and work required from specialists, while also enhancing the accuracy and coverage of asset recovery investigations. To achieve this, parallel processing techniques were employed using *Apache Spark, GraphX, GraphFrames*, and models based on Bidirectional Encoder Representations from Transformers (BERT) for sentence similarity, in order to identify common characteristics among legal entities that may indicate the formation of a corporate group, making the identification process faster and more effective. The results showed that the proposed solution was able to uncover connections not previously mapped in the SNIPER, expanding the set of related companies and highlighting relevant links for asset tracing. A qualitative assessment carried out by two asset recovery specialists indicated gains in both speed and effectiveness, reinforcing the tool's potential to support decision-making and to facilitate the application of legal measures related to joint liability and the disregard of legal personality.

Keywords: Data Mining, Natural Language Processing, Graph Analysis, Sentence Similarity, Apache Spark, BERT, Economic Groups, Asset Recovery

Sumário

1 Introdução **1**

1.1 Descrição do Problema 4
1.2 Pergunta de Pesquisa 4
1.3 Hipótese 5
1.4 Justificativa da Hipótese 5
1.5 Objetivos 5

2 Trabalhos Relacionados e Fundamentação Teórica **7**

3 Metodologia **9**

3.1 Fases no Desenvolvimento dos Artefatos 10

4 Arquitetura Proposta **13**

4.1 Entendendo o Domínio 13
4.2 Entendendo o Dado 15
4.2.1 Estrutura de Grafo 18
4.3 Preparação do Dado 20
4.3.1 Resumo dos Arquivos Utilizados 21
4.3.2 ETL 21
4.4 Integração metodológica: da análise dos dados à modelagem da solução . . 23

5 Artefato 1 - Ferramenta Exploratória **25**

5.1 Modelagem do Dado 26
5.1.1 Extração 26
5.1.2 Transformação 26
5.1.3 Gerando Embeddings de Nome Empresarial 28
5.2 Formulação Proposta 30
5.2.1 Primeiro Estágio 31
5.2.2 Segundo Estágio 33
5.2.3 Terceiro Estágio 34

5.3 Experimentos e Resultados 34
5.3.1 Definição da Linha de Base 35
5.3.2 Seleção da Amostra e Alocação de Neyman 36
5.3.3 Resultados e Análise 37
5.3.4 Validação Estatística dos Resultados 38
5.4 Conclusão Parcial 39

6 Artefato 2 - Interface Investigatória 41
6.1 Metodologia de Implementação Técnica 42
6.1.1 Camada de Banco de Dados 43
6.1.2 Camada de Serviços 44
6.1.3 Interface Web e Visualização dos Resultados 45
6.2 Avaliação do Artefato 2 46
6.2.1 Avaliação da Heurística 47
6.2.2 Avaliação da Interface e do Artefato 47
6.3 Resultados 48
6.3.1 Desempenho da Heurística 48
6.3.2 Percepção sobre a Ferramenta 49
6.3.3 Net Promoter Score (NPS) 50
6.4 Conclusão Parcial 50

7 Considerações Finais 52
7.1 Análise de Resultados 53
7.2 Conclusões 54
7.3 Contribuições 55
7.4 Trabalhos Futuros 56

Referências 58

Anexo 61

I Formulário de Avaliação da Interface Investigatória 62

Lista de Figuras

3.1 Etapas Design Science Research (DSR). Fonte: adaptado de [1]. 10

4.1 Modelo de Dados adaptado de CNPJ. Fonte: o própio autor. 15
4.2 Resumo dos arquivos utilizados na pesquisa.Fonte: o própio autor. 21
4.3 Processo de ETL. Fonte: o própio autor. 22
4.4 Modelo de Referência CRISP-DM. Fonte: adaptado de [2]. 24

5.1 Fluxo Operacional da Heurística Exploratória. Fonte: o próprio autor. . . . 31
5.2 Aumento no Número de Entidades Identificadas pela Ferramenta Exploratória em Diferentes Grupos Econômicos. Fonte: o próprio autor. 38

6.1 Arquitetura do Artefato 2. Fonte: o próprio autor. 42
6.2 Tabela de Comparação. Fonte: o própio autor. 46
6.3 Avaliação Média dos Grupos de Perguntas com Erro Padrão Fonte: o própio autor. 49

Lista de Tabelas

4.1 Relação das características para a identificação de Grupos Econômicos *de Fato* com base no CNPJ 14

5.1 Alocação Amostral com Grupos Ajustados segundo Neyman 37
5.2 Comparativo de Entidades Identificadas pelo SNIPER e pela Ferramenta Exploratória 37
5.3 Validação Estatística do Desempenho da Ferramenta Exploratória 39

Lista de Abreviaturas e Siglas

BERT Bidirectional Encoder Representations from Transformers.

CNAE Classificação Nacional de Atividades Econômicas.

CNJ Conselho Nacional de Justiça.

CNPJ Cadastro Nacional de Pessoas Jurídicas.

CPC Código de Processo Civil.

CRISP-DM Cross Industry Standard Process for Data Mining.

CSV Comma-Separated Values.

ETL Extract, Transform, Load.

JSON JavaScript Object Notation.

RDD Resilient Distributed Datasets.

SNIPER Sistema Nacional de Investigação Patrimonial e Recuperação de Ativos.

UDF User Defined Function.

Capítulo 1

Introdução

O Conselho Nacional de Justiça (CNJ) é o órgão de controle administrativo e financeiro do Poder Judiciário brasileiro, responsável por supervisionar o cumprimento dos princípios constitucionais de eficiência, transparência e autonomia dentro do sistema judiciário. Entre suas competências, destacam-se a fiscalização das atividades dos tribunais, o controle disciplinar e o planejamento estratégico. Com o objetivo de promover a eficiência e a celeridade processual, o CNJ atua na modernização tecnológica de todos os tribunais brasileiros, desenvolvendo e implementando sistemas integrados, como o Processo Judicial Eletrônico (PJe) e o Sistema Nacional de Investigação Patrimonial e Recuperação de Ativos (SNIPER), que visam automatizar rotinas judiciais, facilitar o acesso a informações e otimizar o trâmite de processos. Essas iniciativas buscam uniformizar os procedimentos e melhorar a prestação jurisdicional em todo o país.

Segundo o relatório *Justiça em Números 2024 do CNJ*, a fase de execução é apontada como o principal gargalo processual no Judiciário brasileiro, concentrando a maior parte dos processos em tramitação e sendo a etapa mais demorada para a efetivação das decisões judiciais. A fase de execução, conforme estabelecido no Código de Processo Civil (CPC), corresponde ao estágio do processo judicial em que se busca a satisfação de um direito reconhecido judicialmente, utilizando medidas como penhora, arresto e leilão de bens do devedor (art. 830 e ss. do CPC). Essa dificuldade é acentuada em casos envolvendo a recuperação de ativos, nos quais a localização de bens dos devedores e de Grupos Econômicos ligados a eles torna-se fundamental para a efetiva execução das sentenças.

O conceito de **Grupo Econômico**, conforme o art. 2º, § 2º, da Consolidação das Leis do Trabalho (CLT), refere-se ao conjunto de empresas que, apesar de possuir personalidade jurídica própria, encontram-se sob direção, controle ou administração comuns, respondendo solidariamente pelas obrigações decorrentes da relação de emprego.

No âmbito societário, a Lei 6.404/1976 (Lei das S.A.) complementa essa definição ao dispor, em seu art. 116, que se considera "controladora" a sociedade que detém o poder

de eleger a maioria dos membros do órgão de administração de outra ("controlada"), formalizando assim a existência de um grupo de sociedades empresariais.

Quanto à natureza dos vínculos, distinguem-se:

- **Grupos Econômicos *de Direito***: constituídos formalmente, com estrutura organizacional registrada perante a Junta Comercial ou Comissão de Valores Mobiliários (CVM) e amparados por acordos de acionistas ou instrumentos de controle societário;
- **Grupos Econômicos *de Fato***: caracterizados pela atuação coordenada e interdependente de empresas sem vínculo societário oficial, mas com intenso compartilhamento de gestão, recursos ou estratégias, de modo a evidenciar unidade econômica e finalidade comum.

A identificação e o mapeamento precisos de **Grupos Econômicos** — sejam *de direito* ou *de fato* — constituem etapa preliminar essencial para a aplicação de instrumentos jurídicos capazes de coibir fraudes e garantir a efetividade das execuções.

Uma vez localizado o grupo econômico, torna-se possível acionar o instituto da **Responsabilização Solidária** tributária, previsto no art. 124 do Código Tributário Nacional ([3]). Sob essa regra, todos os integrantes com "interesse comum" no fato gerador do tributo respondem de forma conjunta e ilimitada pelo débito, facultando ao Fisco cobrar a totalidade do crédito de qualquer um dos solidários, que, por sua vez, poderá exercer o direito de regresso contra os demais. Na prática, isso amplia o polo passivo da execução, acelera a recuperação de valores e impõe maior diligência à governança corporativa, pois eventuais falhas ou omissões de um membro afetam todo o universo econômico.

Paralelamente, o deferimento da **Desconsideração da Personalidade Jurídica**, nos termos do art. 50 do Código Civil e do art. 135 do CTN ([4, 3]), permite transpor o véu societário quando comprovados abuso da personalidade — seja por desvio de finalidade ou confusão patrimonial — para alcançar bens particulares de sócios e administradores. Esse remédio jurídico, de caráter excepcional, assegura que credores não fiquem limitados ao patrimônio da pessoa jurídica, garantindo acesso a ativos que, de outra forma, permaneceriam inacessíveis.

Do ponto de vista doutrinário, [5] observa que a personalidade jurídica deve ser preservada como regra, mas não pode servir de "manto protetivo" para práticas fraudulentas. Já [6] ressalta que a teoria maior da desconsideração exige a demonstração inequívoca de abuso, sob pena de se instaurar um ambiente de insegurança jurídica. No mesmo sentido, [7] adverte que o uso indiscriminado do instituto compromete a previsibilidade das relações empresariais, criando obstáculos ao ambiente de negócios e ao investimento privado. Por outro lado, no campo trabalhista, [8] destaca que a solidariedade entre empresas integran-

tes de grupo econômico é fundamental para resguardar créditos de natureza alimentar, cujo adimplemento goza de prioridade absoluta no ordenamento.

A jurisprudência tem consolidado esse entendimento. O Superior Tribunal de Justiça, em diversas ocasiões, reforçou a excepcionalidade da desconsideração, exigindo a demonstração de confusão patrimonial ou desvio de finalidade. Já o Tribunal Superior do Trabalho, em decisões recentes, reafirmou que a caracterização de grupo econômico para fins trabalhistas não depende de prova de fraude, mas da verificação da comunhão de interesses e da atuação conjunta das empresas. Esses precedentes evidenciam o duplo movimento: de um lado, a rigidez no campo civil-tributário; de outro, a flexibilidade ampliada nas relações de trabalho, em função da proteção da dignidade do trabalhador.

Não obstante, é preciso reconhecer os limites desses institutos. A banalização da desconsideração pode levar a um efeito contrário ao pretendido: em vez de coibir abusos, cria-se um cenário de instabilidade jurídica que desestimula o empreendedorismo legítimo. Como sublinha [5], a utilização da desconsideração deve sempre respeitar o equilíbrio entre a proteção do crédito e a preservação da autonomia patrimonial da sociedade, sob pena de se esvaziar a própria razão de ser da pessoa jurídica. Do mesmo modo, a solidariedade, embora poderosa ferramenta de arrecadação, não pode ser estendida a ponto de atingir agentes sem efetiva participação no fato gerador, sob pena de violação do princípio da legalidade tributária.

Em conjunto, a responsabilização solidária e a desconsideração da personalidade jurídica formam a espinha dorsal de qualquer investigação patrimonial capaz de penetrar estruturas complexas e protegidas por múltiplas camadas de controle societário. Seu emprego coordenado fortalece a cobrança de créditos fiscais e trabalhistas, amplia o leque de responsáveis e constitui elemento dissuasório contra a pulverização de patrimônio em artifícios de blindagem corporativa. Como observa [6], a conjugação de tais instrumentos representa verdadeiro "antídoto jurídico" contra o uso abusivo da forma societária, assegurando que a personalidade jurídica não seja deformada em instrumento de fraude.

Diante da crescente complexidade e sofisticação das estruturas empresariais, bem como do recorrente uso de arranjos societários para ocultação patrimonial, o mapeamento rigoroso de vínculos formais e informais revela-se fundamental. Essa investigação robusta não só amplia o rol de responsáveis em execuções fiscais e trabalhistas, mas também reforça a efetividade na recuperação de ativos, justificando a relevância e a urgência do presente estudo.

1.1 Descrição do Problema

A fase de execução dos processos judiciais no Brasil é caracterizada pela morosidade na localização de bens dos devedores, pois essa atividade depende fortemente do esforço e do conhecimento especializado dos peritos em localização de ativos. A *responsabilização solidária* ocorre quando outras entidades, ainda que não formalmente partes no processo, compartilham interesses econômicos, controle ou influência com o devedor, sendo passíveis de terem seus bens utilizados para a satisfação do crédito executado.

O SNIPER foi desenvolvido para auxiliar na identificação de conexões patrimoniais entre empresas e seus respectivos sócios, permitindo uma investigação mais rápida e visual das relações entre entidades jurídicas. Contudo, o SNIPER utiliza como fonte primária de dados o Cadastro Nacional de Pessoas Jurídicas (CNPJ), o que possibilita a localização imediata apenas de Grupos Econômicos *de Direito*, formalmente constituídos e registrados. A identificação de outras empresas que potencialmente comporiam um grupo econômico exige a intervenção de especialistas, que dependem de sua experiência e conhecimento para explorar informações adicionais e não estruturadas.

Empresas que compõem Grupos Econômicos *de Fato*, definidos como aqueles que não possuem vínculo formal, mas operam com estruturas de controle ou interesses comuns, não são localizadas automaticamente pelo SNIPER. Segundo Lopes (2015) [9], esses grupos apresentam características como administração compartilhada, confusão patrimonial, uso comum de ativos, entre outras práticas que indicam a subordinação ou controle conjunto entre entidades. No entanto, a falta de documentação formal dificulta sua identificação automatizada, o que aumenta o tempo e o esforço demandados na análise manual por parte dos peritos. Como consequência, muitos dos ativos pertencentes a esses grupos acabam ocultos, impedindo a aplicação de responsabilidade solidária e limitando a efetividade do processo de recuperação de ativos.

1.2 Pergunta de Pesquisa

Como proporcionar ao especialista em investigação patrimonial, a partir das informações das pessoas a serem investigadas, a identificação de outras pessoas físicas que possam compor um Grupo Econômico *de Fato*, permitindo ampliar o espectro de investigação de ativos passíveis de recuperação e incorporá-los à responsabilização solidária conforme previsto na Lei?

1.3 Hipótese

Uma ferramenta de investigação automatizada, que utilize técnicas de mineração de dados para identificar a formação de Grupos Econômicos *de Fato* a partir de conexões entre pessoas físicas, ampliaria o espectro de análise do especialista e reduziria o tempo de pesquisa, permitindo uma maior identificação de vínculos ocultos e aumentando as possibilidades de responsabilização solidária durante o processo de recuperação de ativos.

1.4 Justificativa da Hipótese

A justificativa da hipótese reside na necessidade de otimizar a investigação patrimonial diante do crescente volume e complexidade dos dados disponíveis. A introdução de uma ferramenta automatizada de mineração de dados, voltada para a identificação de conexões entre pessoas físicas e a formação de Grupos Econômicos *de Fato*, oferece a promessa de ampliar o espectro de análise e reduzir o tempo de investigação. Com isso, torna-se possível fornecer ao especialista mais elementos para uma análise eficiente e aprofundada, permitindo uma responsabilização solidária mais abrangente e contribuindo para a efetividade do processo de recuperação judicial. Dessa forma, o sistema beneficia tanto as autoridades, ao aprimorar a capacidade investigativa, quanto as partes envolvidas no processo, ao possibilitar a identificação de vínculos ocultos ou não documentados com maior rapidez e precisão.

1.5 Objetivos

O Objetivo principal do projeto é ampliar o espectro de identificação de Grupos Econômicos *de Fato* no contexto de investigações patrimoniais, facilitando a responsabilização solidária de pessoas físicas e jurídicas relacionadas ao devedor. A proposta visa automatizar o processo de análise de conexões não documentadas entre entidades, reduzindo o tempo de investigação e aumentando a eficácia na localização de vínculos que possam contribuir para a recuperação de ativos em conformidade com a legislação vigente.

Os objetivos específicos desta dissertação são apresentados a seguir:

- **Objetivo 1:** Aplicar e customizar técnicas de mineração de dados para identificar padrões relevantes em grandes conjuntos de informações financeiras e societárias, facilitando a localização de Grupos Econômicos *de Fato* e a estruturação de conexões ocultas entre entidades.
- **Objetivo 2:** Desenvolver e integrar os artefatos propostos em uma ferramenta com interface amigável, que permita a visualização e exploração eficiente de re-

lacionamentos entre pessoas físicas e jurídicas durante o processo de investigação patrimonial.

- **Objetivo 3:** Avaliar a eficácia da ferramenta desenvolvida por meio de estudos de caso e comparações com métodos tradicionais de investigação, validando sua contribuição para a identificação e recuperação de ativos em processos judiciais.
- **Objetivo 4:** Propor diretrizes para a expansão e utilização da ferramenta em novos contextos, explorando a escalabilidade e adaptabilidade do sistema a diferentes fontes de dados e cenários de investigação.

Capítulo 2

Trabalhos Relacionados e Fundamentação Teórica

Representações de dados baseadas em grafos tornaram-se uma ferramenta essencial para a extração de conhecimento a partir de conjuntos de dados complexos, ao modelar de forma natural as relações intrincadas entre entidades. A literatura jurídica define grupos econômicos como estruturas empresariais compostas por múltiplas entidades juridicamente independentes, mas interligadas por relações de controle, administração ou coordenação [9]. No Brasil, o reconhecimento de um grupo econômico pode ocorrer de duas formas: *de direito*, quando formalmente constituído por meio de contratos e registros empresariais, e *de fato*, quando empresas atuam de maneira coordenada sem um vínculo societário formal, sendo essa última categoria mais desafiadora para identificação [10].

A identificação de grupos econômicos *de fato* é um desafio crítico tanto para fins de investigação patrimonial quanto para a execução de obrigações legais e fiscais, uma vez que sua caracterização exige análise de vínculos não evidentes entre pessoas jurídicas [9, 10]. Com o avanço das técnicas de mineração de dados e modelagem de grafos, novas abordagens surgiram para apoiar essa tarefa. A análise de grafos tem se mostrado uma ferramenta eficaz na detecção de conexões ocultas entre empresas, permitindo a construção de redes de relacionamentos e inferência de vínculos indiretos. [11] destaca a aplicabilidade da análise de grafos para detectar fraudes em aquisições públicas no Brasil, demonstrando que grupos econômicos podem ser identificados por meio da sobreposição de sócios e padrões de comportamento anômalos nas licitações públicas.

Estudos como o de Brandão et al. [11] exploram o uso de algoritmos de detecção de comunidades e previsão de ligações para mapear redes de empresas interligadas. A formalização de atributos heurísticos para caracterizar *grupos econômicos de fato* tem sido estudada por [9], que definem critérios como identidade de gestores, quadro societário, denominação social, CNAE, localização de sedes e independência meramente formal. Esses

parâmetros orientam a modelagem de *features* no grafo, permitindo a aplicação de regras de negócio alinhadas a fundamentos legais.

Casos reais de fraude corporativa ilustram a importância dessas técnicas. Investigações após o colapso da Enron nos EUA [12] e o escândalo Wirecard na Alemanha [13] revelaram estruturas societárias complexas usadas para ocultar passivos e facilitar desvios de recursos. Pesquisas em detecção de anomalias em redes financeiras — exemplificadas por NetProbe [14] e por OddBall [15] — demonstram o valor de algoritmos baseados em grafos para identificar padrões atípicos em transações e relações.

Estudos também mostram que a correta identificação de grupos econômicos *de fato* desempenha papel essencial na fundamentação da desconsideração da personalidade jurídica. No contexto jurídico, a comprovação de vínculos ocultos pode fornecer evidências robustas para o redirecionamento de execuções fiscais e trabalhistas, especialmente em casos de confusão patrimonial e abuso da personalidade jurídica [10, 16].

No campo da computação, trabalhos seminais como Pregel [17] introduziram abordagens escaláveis centradas no vértice para o processamento de grafos em larga escala. PowerGraph foi posteriormente proposto para lidar com grafos caracterizados por distribuições de grau assimétricas [18]. Estudos como [19] e [20] mostram como hardware convencional pode ser utilizado para computações em grafos, enfrentando desafios de conversão e gerenciamento de conjuntos extensos. O Apache Spark [21] e seu componente GraphX [22] integram o processamento de grafos ao processamento distribuído, permitindo a transformação eficiente de bases heterogêneas em modelos de grafo.

Para operacionalizar a construção e atualização desses grafos, adotam-se pipelines de dados automatizados. Ferramentas como Apache Airflow ou agendamentos Cron disparam fluxos de Extract, Transform, Load (ETL) em Spark combinados com GraphFrames, mantendo sincronizadas as relações entre entidades [21]. Essa automação assegura que análises e consultas reflitam o estado mais atual das bases e reduz a necessidade de intervenção manual.

Mais recentemente, Graph Neural Networks (GNNs) surgem como evolução natural para a inferência de conexões ocultas em grafos. Modelos como Node2Vec [23], GraphSAGE [24] e Graph Attention Networks [25] demonstram capacidade superior de capturar propriedades estruturais e de aprender representações vetoriais de nós, facilitando a predição de vínculos não explícitos.

Coletivamente, esses avanços em processamento distribuído, automação de pipelines, aprendizado profundo sobre grafos e fundamentos jurídicos formam a base teórica e metodológica para sistemas de investigação patrimonial automatizada, contribuindo para o enfrentamento da ocultação patrimonial e a efetivação das obrigações legais.

Capítulo 3

Metodologia

A metodologia Design Science Research (DSR) [1] foi utilizada para guiar o processo de desenvolvimento de artefatos nesta pesquisa, proporcionando uma estrutura sistemática para a criação, avaliação e aprimoramento de soluções inovadoras. O DSR é particularmente adequado para trabalhos que buscam resolver problemas complexos por meio do desenvolvimento de novos artefatos, como modelos, sistemas, metodologias e frameworks.

No contexto deste estudo, a aplicação do DSR permitiu a definição de um conjunto de artefatos voltados para a identificação automatizada de Grupos Econômicos, com foco em estender as funcionalidades de ferramentas existentes e superar as limitações atuais na investigação patrimonial.

3.1 Fases no Desenvolvimento dos Artefatos

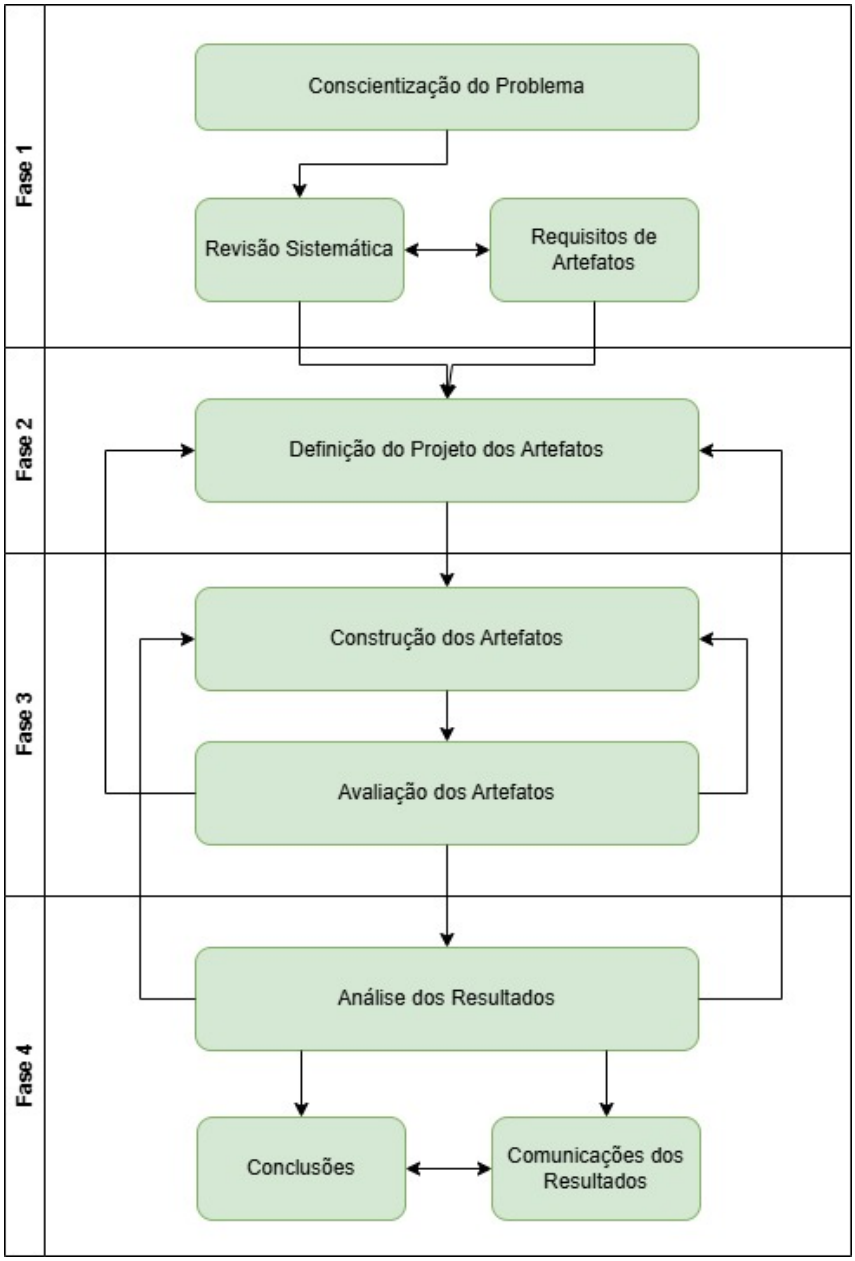


Figura 3.1: Etapas Design Science Research (DSR). Fonte: adaptado de [1].

O fluxo apresentado na Figura 3.1 é uma adaptação do modelo Design Science Research (DSR), voltada para atender às necessidades específicas desta pesquisa. Este modelo adota uma abordagem iterativa para o planejamento, construção e avaliação dos artefatos desenvolvidos.

A seguir, são descritas as quatro fases do DSR, suas respectivas etapas e o status final de execução no contexto desta investigação, considerando que todas as atividades previstas foram realizadas e documentadas.

A seguir, são descritas as quatro fases do DSR, suas respectivas etapas e o status final de execução no contexto desta investigação, considerando que todas as atividades previstas foram realizadas e documentadas.

- **Fase 1: Identificação do Problema**
  - **Conscientização do Problema (Concluído)**
    - Análise dos Desafios.
    - Identificação das Necessidades.
  - **Revisão Sistemática da Literatura (Concluído)**
    - Levantamento de Estudos Anteriores.
    - Análise de Lacunas.
  - **Requisitos de Artefatos (Concluído)**
    - Identificação dos Artefatos.
    - Especificação dos Requisitos dos Artefatos.
- **Fase 2: Projeto de Artefatos**
  - **Definição do Projeto do Artefato (Concluído)**
    - **Técnicas de Mineração Aplicadas:** Aplicação dos processos definidos no CRISP-DM para mineração de dados e resolução dos problemas apontados no artefato de automação da investigação patrimonial.
    - **Arquitetura Utilizada:** Definição da arquitetura modular para o artefato, dividindo em componentes para coleta, processamento, análise e visualização de dados.
- **Fase 3: Desenvolvimento e Avaliação do Artefato**
  - **Construção do Artefato (Concluído)**
    - Construção, testes, documentação e promoção de colaboração dos especialistas.
  - **Avaliação do Artefato (Concluído)**
    - A avaliação dos artefatos foi realizada por meio de uma pesquisa com os usuários especialistas da área, contemplando o levantamento do perfil dos avaliadores.

- **Fase 4: Conclusões e Comunicação dos Resultados**
  - **Análise dos Resultados (Concluído)**
    * Sumário dos Achados: Recapitulação dos principais resultados e destaque das descobertas mais significativas.
    * Comparação com Expectativas: Comparação entre os resultados alcançados e as expectativas iniciais.
    * Identificação de Lições Aprendidas: Registro e análise das lições aprendidas ao longo do processo de desenvolvimento e avaliação.
  - **Conclusões (Concluído)**
    * Contribuições para a Prática e Teoria: Discussão das contribuições dos artefatos para a prática da investigação patrimonial.
    * Limitações e Possíveis Melhorias: Reconhecimento e discussão das limitações, identificando oportunidades para aprimoramentos e estudos futuros.
  - **Comunicação dos Resultados (Parcialmente Concluído)**
    * Elaboração da Dissertação.
    * Apresentações e Publicações: Preparação de apresentações e submissão de trabalhos para conferências, workshops e periódicos relevantes.
    * Divulgação para Stakeholders: Compartilhamento dos resultados da pesquisa com os principais stakeholders envolvidos na investigação patrimonial.

Capítulo 4

Arquitetura Proposta

Para a resolução do problema apontado e a preparação da arquitetura proposta, foi utilizado o Cross Industry Standard Process for Data Mining (CRISP-DM) [2] como abordagem metodológica. O CRISP-DM é um modelo de processo amplamente reconhecido no campo da Mineração de Dados, dividido em seis fases sequenciais: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implementação. Esse modelo foi escolhido devido à sua estrutura flexível e ao seu foco em transformar dados brutos em conhecimento, o que permite iterar e ajustar as etapas conforme surgem novos achados durante o processo investigativo.

No contexto deste estudo, o CRISP-DM é particularmente adequado para gerenciar a complexidade envolvida na identificação de Grupos Econômicos *de Fato*, pois oferece uma abordagem sistemática que facilita a análise e a transformação de grandes volumes de dados em informações acionáveis para recuperação de ativos. Além disso, a aplicação do CRISP-DM proporcionou uma base estruturada para a escolha da arquitetura proposta, possibilitando alinhar as fases de processamento de dados, modelagem e análise com as necessidades específicas do projeto.

4.1 Entendendo o Domínio

Conforme [9], desconsideração da personalidade jurídica acontece em casos por abuso de personalidade e que em geral há presença de algumas características, tais como:

1. A identidade de gestores comuns (administradores, contadores, advogados, etc.);
2. Formação de quadro societário comum ou com pessoas vinculadas (parentes);
3. Denominação social aproximada;
4. Atuação complementar ou semelhante;

5. Compartilhamento de Bens e Infraestruturas;

6. A independência meramente formal de pessoas jurídicas.

A desconsideração da personalidade jurídica é especialmente relevante nos casos em que se identificam grupos econômicos *de fato*, nos quais a interdependência entre entidades formalmente independentes configura uma relação econômica substancial, mesmo que não documentada formalmente. Em situações como essas, a identificação de um grupo econômico *de fato* permite estender a responsabilização solidária a todas as entidades envolvidas, incluindo aquelas que, em teoria, estariam isoladas legalmente. Nesse contexto, a aplicação de critérios como a independência meramente formal, identidade de gestores, quadro societário comum e outros mencionados anteriormente, serve como base para uma abordagem investigativa mais robusta. Assim, com o suporte de especialistas — um Juiz Federal do Tribunal Regional Federal da 1ª Região e um Juiz do Trabalho do Tribunal Regional do Trabalho da 14ª Região — , foi desenvolvida uma heurística capaz de identificar indícios que caracterizam grupos econômicos *de fato*, integrando os elementos descritos de maneira sistemática para facilitar a análise e conexão entre as entidades envolvidas.

As abordagens adotadas compatibilizaram as características citadas com os dados contidos no Cadastro Nacional de Pessoas Jurídicas, permitindo efetuar comparações entre as empresas alvo e as empresas candidatas a compor o Grupo Econômico, considerando as informações disponíveis no registro formal de entidades empresariais. Para tanto montou-se o seguinte mapeamento:

	Característica Desconsideração	**Dado do CNPJ**
1	A identidade de gestores comuns	Representantes Legais em comum
2	Formação de quadro societário comum	Sócios em comum
3	Denominação social aproximada	Semelhança entre nomes fantasia ou empresarial
4	Atuação complementar ou semelhante	Analisar CNAE
5	Compartilhamento de Bens e Infraestruturas	Comparar endereços físicos, e-mail, número telefônico
6	A independência meramente formal de pessoas jurídicas	Todos os dados já citados

Tabela 4.1: Relação das características para a identificação de Grupos Econômicos *de Fato* com base no CNPJ

A heurística proposta será detalhada ao longo do trabalho, especialmente na seção dedicada aos artefatos desenvolvidos, onde será demonstrada a abordagem adotada para a automatização da investigação de grupos econômicos *de fato*. Nessa etapa, discutiremos como cada um dos critérios mencionados foi integrado ao modelo de análise, permitindo que as relações e padrões de interdependência entre entidades sejam identificados de forma

automatizada e sistemática, com base nas informações disponíveis no CNPJ e outras fontes relevantes.

4.2 Entendendo o Dado

A fonte primária dos dados utilizados neste trabalho foi o Cadastro Nacional de Pessoas Jurídicas (CNPJ), disponível no Portal de Dados Abertos do Governo Federal [26]. Essa base fornece informações estruturadas sobre entidades jurídicas registradas no Brasil, essenciais para a identificação e análise de grupos econômicos *de fato*. O trabalho inicial consistiu em estudar o dicionário de dados disponível no portal, o que permitiu compreender a estrutura dos dados e identificar os atributos relevantes para a formação da heurística proposta. Esses atributos estão descritos na Tabela 4.1 e incluem elementos-chave para caracterizar relações societárias e interdependências entre as entidades investigadas.

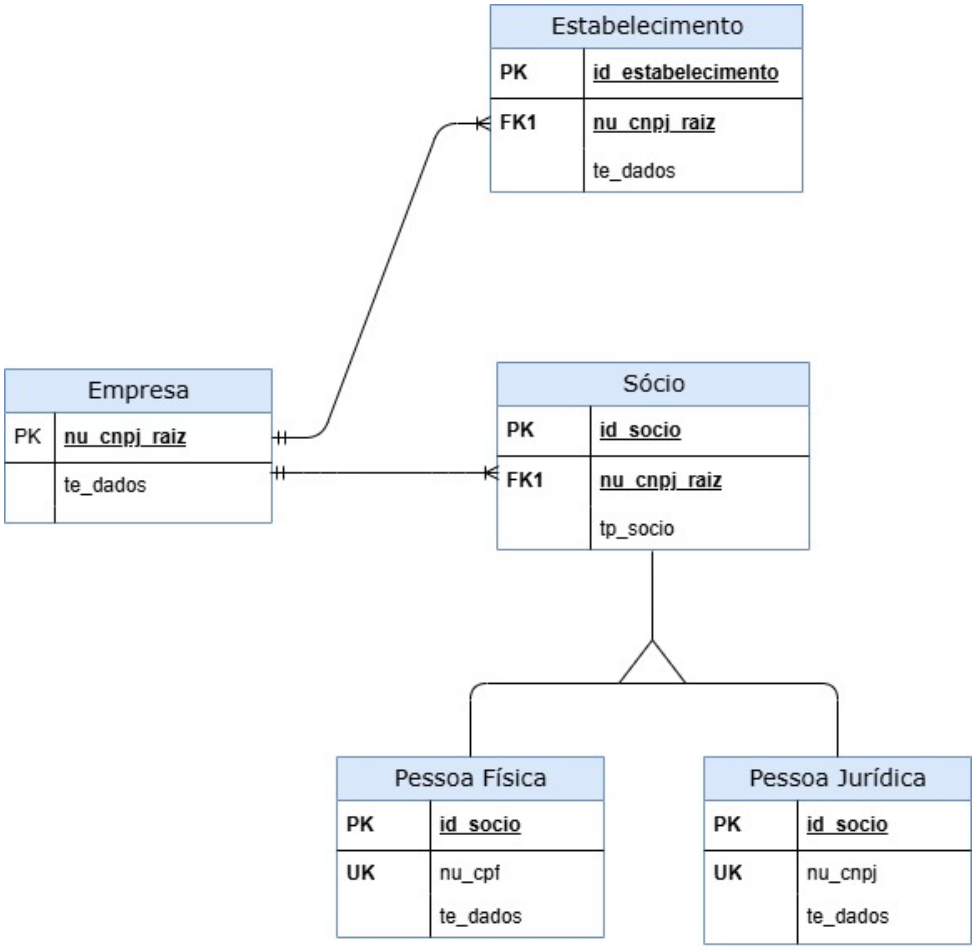


Figura 4.1: Modelo de Dados adaptado de CNPJ. Fonte: o própio autor.

As estruturas de dados identificadas como de interesse para a análise foram mapeadas no modelo da Figura 4.1, representando a estrutura de dados adaptada obtida de parte do Cadastro Nacional de Pessoas Jurídicas. Esse modelo é composto por cinco principais entidades: Empresa, Estabelecimento, Sócio, Pessoa Física e Pessoa Jurídica.

A estrutura de dados **Empresa** contém informações gerais sobre cada pessoa jurídica registrada, identificada pelo atributo nu_cnpj_raiz (CNPJ raiz). Esse campo é utilizado como chave primária para identificar unicamente cada empresa e está relacionado com outras estruturas do modelo. Esta estrutura também contém o campo te_dados, que armazena demais atributos da empresa em formato JavaScript Object Notation (JSON).

A estrutura de dados **Estabelecimento** detalha os diferentes locais de atuação de uma mesma empresa, permitindo que uma única empresa possua múltiplos estabelecimentos. Cada empresa deve ter pelo menos um estabelecimento, que representa sua Matriz, enquanto os demais estabelecimentos dessa mesma empresa são considerados filiais. Essa estrutura é identificada por uma chave primária formada pelo campo id_estabelecimento. O campo nu_cnpj_raiz atua como uma chave estrangeira que se relaciona com a estrutura Empresa, garantindo a associação entre cada estabelecimento e sua respectiva empresa. Além disso, o campo te_dados nesta estrutura está em formato JSON, armazenando informações adicionais específicas para cada estabelecimento, incluindo atributo que distinguem a matriz de suas filiais.

A estrutura de dados **Sócio** armazena os dados dos sócios de cada empresa, identificados por uma chave primária composta formada pelo campo id_socio. O campo nu_cnpj_raiz funciona como uma chave estrangeira, indicando a relação com a empresa à qual o sócio está associado. Cada empresa deve ter pelo menos um sócio registrado, garantindo a vinculação entre a entidade empresarial e seus responsáveis ou participantes. O campo tp_socio especifica o tipo de sócio, permitindo diferenciar entre sócios do tipo Pessoa Física e Pessoa Jurídica. Além disso, o campo te_dados, em formato JSON, armazena informações complementares sobre os sócios. Nesse contexto, a estrutura Sócio atua como uma abstração, unificando os dados de Pessoa Física e Pessoa Jurídica em uma mesma entidade, permitindo tratar ambos de maneira integrada no modelo de dados.

A estrutura **Pessoa Física** identifica sócios que são pessoas naturais, enquanto a estrutura **Pessoa Jurídica** identifica sócios que são outras entidades jurídicas. Ambas as estruturas são especializações da entidade Sócio e possuem uma chave primária id_socio. Além disso, armazenam nu_cpf e nu_cnpj como chave única (UK), além do campo te_dados em formato JSON, que armazena demais atributos sobre esses sócios de forma estruturada.

A modelagem adota um padrão de especialização onde a chave primária da entidade **Sócio** é herdada pelas entidades **Pessoa Física** e **Pessoa Jurídica**, assegurando con-

sistência e unicidade. Essa estrutura permite distinguir de forma clara os diferentes tipos de sócios, além de viabilizar consultas especializadas sobre cada tipo de entidade.

As chaves estrangeiras foram posicionadas de forma a refletir as relações $1:N$ entre empresas e estabelecimentos, bem como entre empresas e sócios. A modelagem do campo **te_dados** em formato JSON proporciona flexibilidade para armazenar atributos complementares e facilita o processamento paralelo dos dados utilizando ferramentas como Spark e GraphFrames.

Essa estrutura relacional foi posteriormente transformada em um grafo híbrido, descrito na seção seguinte, para representar as interdependências entre entidades e permitir análises de conectividade e similaridade com maior eficiência e expressividade.

As estruturas de dados identificadas contêm um grande volume de registros, essenciais para a análise das interconexões entre entidades. A estrutura Empresa possui um total de 60.159.846 registros, representando cada pessoa jurídica cadastrada. A estrutura Estabelecimento, que detalha os locais de atuação dessas empresas, possui 63.194.841 registros. Além disso, a estrutura de dados dos Sócios (incluindo tanto Pessoa Física quanto Pessoa Jurídica) totaliza 24.869.668 registros.

Estrutura dos Campos `te_dados` (JSON)

Os campos `te_dados`, presentes nas entidades `Empresa`, `Estabelecimento` e `Sócio`, armazenam informações adicionais em formato JSON. A seguir, são destacados os atributos mais relevantes para o processo de investigação patrimonial e identificação de grupos econômicos *de fato*:

- **Empresa**
  - `cpfResponsavel` – CPF do responsável pela empresa.
  - `capitalSocial` – Valor do capital social (como proxy de capacidade financeira).
  - `nomeEmpresarial` – Nome legal registrado da empresa.
  - `porteEmpresa` – Pode indicar micro ou grande porte.
  - `naturezaJuridica` – Útil para filtrar tipos societários.
- **Estabelecimento**
  - `nomeFantasia` – Nome comercial, usado para verificação semântica de similaridade.
  - `logradouro`, `numero`, `bairro`, `municipio`, `cep`, `uf` – Campos de endereço completos, fundamentais para detectar compartilhamento de local.

  - `telefone1`, `telefone2`, `email` – Indícios de infraestrutura compartilhada.
  - `cnaeFiscal`, `cnaesSecundarias` – Classificação de atividade econômica, usada para verificar atuação complementar ou sobreposta.
  - `identificadorMatrizFilial` – Diferencia matriz de filial, útil para consolidação de dados.

- **Sócio**
  - `identificadorSocio` – Identifica se é pessoa física, jurídica ou estrangeira.
  - `entradaSociedade` – Data de entrada na empresa, útil para ver vínculos temporais.
  - `cpfRepresentanteLegal` – Em casos de sócio pessoa jurídica, pode indicar vínculo oculto com PF.
  - `qualificacaoSocio`, `qualRepresentanteLegal` – Indicadores do tipo de participação.

4.2.1 Estrutura de Grafo

O foco principal deste trabalho é estudar os relacionamentos entre as entidades Empresa (e Estabelecimento) e Sócios, levando em consideração os atributos citados na Tabela 4.1. Estruturas orientadas a grafos são eficazes na modelagem de relações que envolvem múltiplos níveis de conexão e interdependência, comuns em grupos econômicos *de fato*. Conforme argumentado por [27], "os grafos permitem uma representação mais natural de dados com relações complexas, que são intrinsecamente difíceis de mapear e consultar eficientemente em um banco de dados relacional", o que é particularmente relevante ao trabalhar com as estruturas de dados do CNPJ, onde pessoas podem atuar em múltiplas empresas e empresas podem ser sócias de outras empresas.

Neste trabalho, propõe-se a organização dos dados em uma estrutura de grafos, para persistência e execução de consultas, permitindo uma representação mais intuitiva das relações societárias. Nesse grafo, são considerados dois tipos de nós principais: Sócio Pessoa Jurídica (PJ) e Sócio Pessoa Física (PF). As conexões entre eles são representadas como arestas, que simbolizam uma união societária entre os nós. Para simplificar a modelagem, o nó Sócio PJ é uma junção (join) entre as tabelas Empresa e Estabelecimento, permitindo um único ponto de representação para entidades jurídicas com múltiplos estabelecimentos. Essa abordagem resulta em um grafo híbrido, onde os nós representam tanto entidades únicas (pessoas físicas) quanto agregações de dados (pessoas jurídicas com estabelecimentos), facilitando a análise e a visualização das interdependências dentro de grupos econômicos *de fato*.

Para representar matematicamente o grafo descrito, consideramos um grafo híbrido

$$G = (V, E) \tag{4.1}$$

Onde:

- V é o conjunto de **nós** do grafo, composto por dois tipos de nós:

- **Sócio Pessoa Física (PF)**: Representa pessoas físicas associadas a uma empresa.
- **Sócio Pessoa Jurídica (PJ)**: Representa uma junção entre a tabela **Empresa** e a tabela **Estabelecimento**, simplificando a representação de entidades jurídicas com múltiplos estabelecimentos em um único nó.

Podemos definir o conjunto de nós V como:

$$V = V_{PF} \cup V_{PJ} \tag{4.2}$$

Onde:

- $V_{PF} = \{v \in V : v \text{ é um sócio pessoa física}\}$
- $V_{PJ} = \{v \in V : v \text{ é uma pessoa jurídica, com um ou mais estabelecimentos}\}$

As arestas $E \subseteq V \times V$ representam as **uniões societárias** entre os nós, indicando as relações de associação entre **Sócio PJ** e **Sócio PF**. Cada aresta $e = (u, v) \in E$ conecta um nó u a um nó v e define uma relação societária.

Definimos então o grafo G como:

$$G = (V, E) = (V_{PF} \cup V_{PJ}, E) \tag{4.3}$$

Onde:

- V representa os dois tipos de nós (**PF** e **PJ**).
- E representa as **uniões societárias** entre esses nós.

Complexidade Prática na Estruturação do Grafo

A transposição dos dados do Cadastro Nacional da Pessoa Jurídica (CNPJ) e das sociedades empresariais para um modelo em grafo envolve desafios que vão além da mera representação de relações. A primeira dimensão a ser considerada é a escala: milhões de nós e dezenas de milhões de arestas tornam operações rotineiras, como consultas e junções, computacionalmente intensivas. Essa magnitude impõe restrições práticas tanto

na alocação de memória quanto no tempo de execução, exigindo técnicas de indexação, particionamento e filtragem regional, como a materialização de visões específicas, para viabilizar análises.

Outro aspecto relevante é a heterogeneidade dos nós. O grafo contém tanto pessoas jurídicas quanto pessoas físicas, categorias que apresentam naturezas distintas e demandam representações diferenciadas de atributos. Essa diversidade reflete-se na definição das features utilizadas no treinamento de modelos de aprendizado de máquina, uma vez que as pessoas jurídicas possuem informações ricas (CNAE, endereço, telefone, embeddings de nomes), enquanto as pessoas físicas exigem simplificações e codificações mais restritivas. O desafio, portanto, consiste em equilibrar granularidade informacional e viabilidade computacional, evitando vetores de features demasiadamente esparsos ou dimensionalmente onerosos.

Além disso, a multiplicidade de relações cria padrões densos e complexos de conectividade. Um mesmo sócio pode figurar em diversas empresas, formando estruturas de alta centralidade que aumentam a complexidade do grafo e impactam algoritmos de busca, clusterização e aprendizado. Esse fenômeno contribui para a formação de comunidades densas e sobrepostas, cujas fronteiras são difíceis de delimitar sem técnicas avançadas de modelagem. Assim, a construção do grafo não é apenas uma etapa de transformação de dados, mas uma decisão arquitetural crítica que condiciona a eficiência dos métodos analíticos e preditivos aplicados posteriormente.

Todos esses elementos foram cuidadosamente considerados no processo de preparação do dado, de modo a assegurar que a estrutura em grafo pudesse refletir, de forma fiel e eficiente, as relações entre empresas e sociedades. A atenção a aspectos como escala, heterogeneidade dos nós e densidade das conexões foi fundamental para orientar as escolhas de modelagem e processamento, estabelecendo as bases sobre as quais as análises e experimentos subsequentes puderam ser conduzidos com robustez metodológica.

4.3 Preparação do Dado

Os dados brutos foram baixados do Portal de Transparência [26] em formato *Comma-Separated Values (CSV)*. O formato CSV é um tipo de arquivo de texto simples onde os dados são organizados em linhas e colunas, separados por vírgulas (ou outro delimitador, como ponto e vírgula). Cada linha do arquivo representa um registro, e as colunas contêm os diferentes atributos de cada registro. Esse formato é amplamente utilizado para exportar e compartilhar grandes volumes de dados devido à sua simplicidade e compatibilidade com diversos softwares e ferramentas de análise, incluindo planilhas eletrônicas e sistemas de banco de dados.

No caso do Cadastro Nacional de Pessoa Jurídica (CNPJ), foram baixados três arquivos principais: Empresas, Sócios e Estabelecimentos. Esses arquivos contêm as informações essenciais para a análise das relações societárias, incluindo dados sobre as entidades jurídicas (Empresas), as pessoas físicas e jurídicas associadas como sócios (Sócios) e os locais de atuação das empresas (Estabelecimentos). Esses dados em formato CSV serviram como ponto de partida para a fase de preparação e transformação, sendo convertidos e organizados para facilitar a modelagem e análise no formato de grafo.

4.3.1 Resumo dos Arquivos Utilizados

Os arquivos utilizados foram extraídos da base pública do CNPJ disponibilizada pela Receita Federal em maio de 2025. A Figura 4.2 apresenta um resumo estimado com os principais indicadores de volume e qualidade dos dados.

Entidade	Arquivos ZIP	Tamanho Total	Registros Estimados	Observações sobre Qualidade dos Dados
Empresas	10	~1,17 GB	~60 milhões	Campos como `capitalSocial` com zeros à esquerda; possíveis duplicações por CNPJ raiz.
Estabelecimentos	10	~4,15 GB	~63 milhões	Muitos registros com e-mails e telefones ausentes; endereços incompletos; usados para análise de locais.
Sócios	10	~599 MB	~25 milhões	Dados nulos em representante legal; `identificadorSocio` e `qualificação` exigem decodificação auxiliar.

Figura 4.2: Resumo dos arquivos utilizados na pesquisa.Fonte: o própio autor.

4.3.2 ETL

Para a realização dos processos de ETL(Figura 4.3) e das consultas para a avaliação do resultado preliminar, foi utilizada uma máquina de alto desempenho como laboratório. O termo Extract, Transform, Load (ETL) refere-se a um conjunto de processos essenciais para a manipulação e preparação de dados. Na fase de extração (Extract), os dados são obtidos da fonte original; na fase de transformação (Transform), eles são limpos e organizados conforme as necessidades do modelo; e na fase de carregamento (Load), são inseridos no ambiente de destino, no caso, uma estrutura de grafo.

A máquina utilizada como laboratório é um servidor Dell PowerEdge R820, com as seguintes especificações:

- **Memória**: 512 GB.

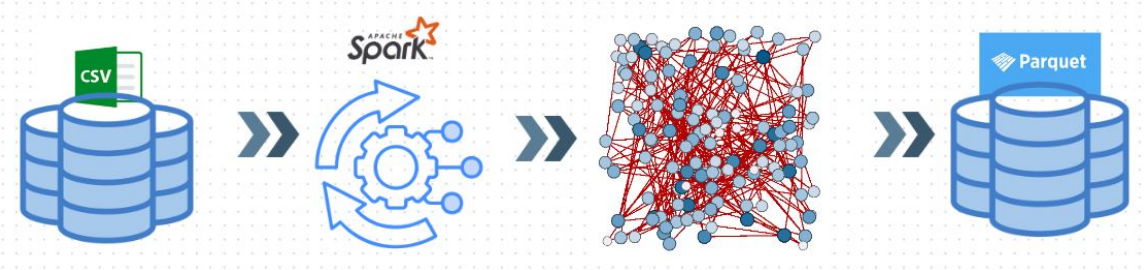


Figura 4.3: Processo de ETL. Fonte: o própio autor.

- **Processador**: Intel(R) Xeon(R) CPU E5-4650 @ 2.70GHz, configurado com 4 processadores, cada um com 8 cores físicos e 16 lógicos, totalizando 64 cores.

O elevado número de cores torna essa máquina especialmente adequada para o uso de recursos multiprocessados, otimizando o tempo de execução das tarefas de ETL e das consultas complexas necessárias para a análise das relações societárias. Esse recurso de paralelismo permite distribuir os processos de manipulação de dados, aumentando significativamente a eficiência ao processar grandes volumes de informações do CNPJ.

A escolha do **Apache Spark** e de sua ferramenta **GraphX**, para a realização das atividades de Extração e Transformação do dado da pesquisa, foi motivada pela necessidade de lidar com grandes volumes de dados e executar computações paralelas em grafos de forma eficiente. O Spark é uma estrutura de processamento de dados em larga escala que oferece recursos robustos para escalabilidade e tolerância a falhas, essenciais ao trabalhar com dados do CNPJ. Segundo [21], o Spark introduz o conceito de **Resilient Distributed Datasets (RDD)** como sua abstração principal, permitindo o cache de dados em memória em várias máquinas e facilitando múltiplas operações paralelas semelhantes ao **MapReduce**, além de proporcionar uma infraestrutura resiliente a falhas [21].

O **GraphX** [28], especificamente, é uma API do Spark projetada para computação em grafos, incluindo algoritmos como **PageRank** e **filtragem colaborativa**. Em um nível elevado, o GraphX estende a abstração RDD do Spark, introduzindo o **Resilient Distributed Property Graph**: um multigrafo direcionado com propriedades associadas a cada vértice e aresta, ideal para a análise de relações complexas entre entidades como empresas e sócios.

O Spark é baseado em uma estrutura de processamento distribuído chamada **MapReduce**, que foi inicialmente desenvolvida pelo Google para simplificar o cálculo em grandes conjuntos de dados distribuídos. No modelo MapReduce, as tarefas são divididas em duas fases principais:

Map: O conjunto de dados é dividido em várias partes menores, e cada parte é processada independentemente para gerar pares de chave-valor. **Reduce**: Os pares de chave-

valor são então combinados para produzir o resultado final, permitindo a paralelização e a distribuição de tarefas entre várias máquinas. Dean e Ghemawat (2008) descrevem o MapReduce como um framework que aborda a complexidade do processamento distribuído, paralelização e tolerância a falhas, facilitando o processamento de grandes volumes de dados distribuídos por centenas ou milhares de máquinas [29]. Essa abordagem permite que o Spark e o GraphX manipulem grandes conjuntos de dados com eficiência, dividindo as tarefas de análise de grafos em processos paralelos distribuídos.

A aplicação do Spark e do GraphX, portanto, justifica-se pela escalabilidade e desempenho no processamento de grafos complexos, facilitando a análise de relações societárias em grandes bases de dados como a do CNPJ, onde a necessidade de paralelismo e distribuição de tarefas é essencial.

A fase de transformação será realizada com foco na utilização de pesquisas no **GraphX** para realizar consultas avançadas em grafos. As consultas serão conduzidas utilizando **GraphFrames** com o recurso de **motifs**. Os motifs permitem a identificação de padrões de conexão específicos dentro do grafo, como caminhos e ciclos, tornando possível detectar subestruturas que indicam relações entre empresas e sócios. Isso facilita a descoberta de grupos econômicos *de fato* por meio de padrões previamente definidos de conexão entre as entidades.

Para a persistência dos dados transformados, será utilizado o formato **Parquet**, que armazena os dados em colunas, proporcionando um armazenamento mais eficiente e leitura otimizada para consultas futuras. O **Parquet** é especialmente vantajoso para grandes volumes de dados, pois permite compressão eficiente e acesso rápido a colunas específicas, essencial para a análise iterativa e exploratória dos dados no grafo.

4.4 Integração metodológica: da análise dos dados à modelagem da solução

O processo de construção da solução proposta foi orientado por um fluxo metodológico adaptado do modelo de referência CRISP-DM (Cross Industry Standard Process for Data Mining), amplamente utilizado em projetos de mineração de dados. Esse modelo foi ajustado para refletir as fases específicas do trabalho de identificação de Grupos Econômicos *de Fato* a partir da base de dados do CNPJ.

A Figura 4.4 apresenta essa adaptação, evidenciando seis fases principais: (1) entendimento do negócio, (2) entendimento dos dados, (3) preparação dos dados, (4) modelagem, (5) avaliação e (6) implantação. As três primeiras fases — já descritas neste capítulo — abordam a fundamentação do problema jurídico, a análise exploratória da base CNPJ e

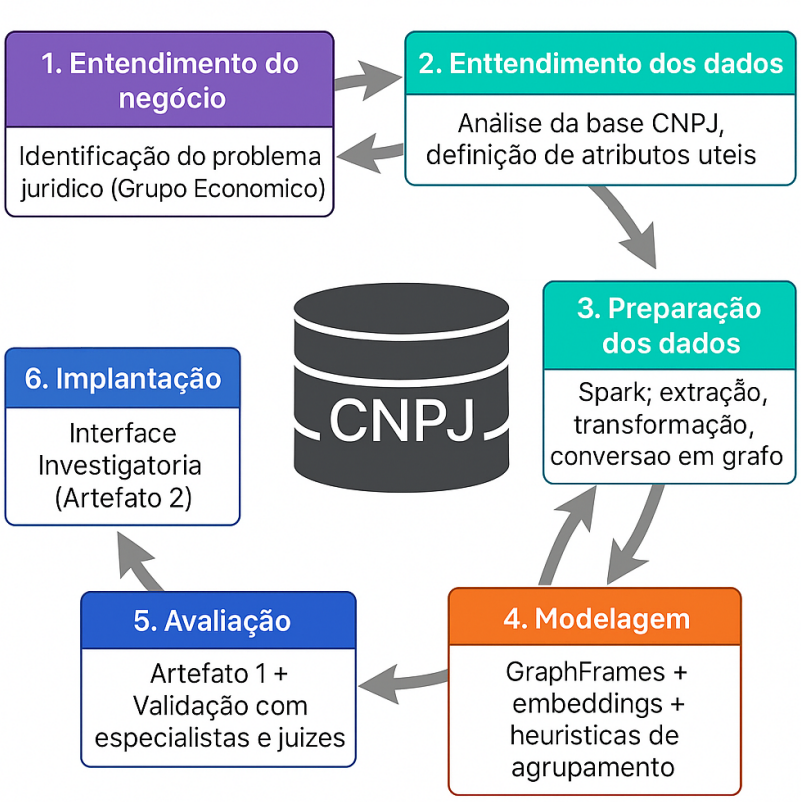


Figura 4.4: Modelo de Referência CRISP-DM. Fonte: adaptado de [2].

a transformação dos dados em uma estrutura grafo relacional com informações relevantes para o processo investigativo.

As próximas fases, a partir do Capítulo seguinte, tratarão da modelagem computacional da solução, das avaliações conduzidas com especialistas do domínio e da implantação da interface final voltada ao uso prático da ferramenta investigatória.

Capítulo 5

Artefato 1 - Ferramenta Exploratória

A *Ferramenta Exploratória* é o primeiro artefato gerado neste trabalho e é responsável por realizar a **Preparação do Dado**, compreendendo as três fases do processo de **Extract, Transform, Load (ETL)**. Esse processo começa com a **extração** dos dados brutos do CNPJ, armazenados em arquivos CSV, que servem como fonte primária das informações iniciais. Em seguida, a ferramenta executa a **transformação** dos dados, adaptando-os para a estrutura em grafo proposta (Equação 4.3). Nesta fase, são aplicadas regras de modelagem que possibilitam a organização das entidades e das relações societárias em um grafo híbrido, facilitando a análise de grupos econômicos *de fato*. Por fim, na etapa de **carga**, os dados processados são armazenados em formato **Parquet**.

A ferramenta implementa a heurística descrita na Seção 4.1 para identificar indícios de grupos econômicos *de fato*, comparando os dados de uma empresa selecionada como alvo com clusters de empresas identificados no conjunto de dados reduzido do CNPJ. Essa heurística analisa os tipos de dados listados na Tabela 4.1 e permite verificar similaridades estruturais e relacionais que sugerem vínculos societários significativos.

Ao final da aplicação da heurística, a ferramenta organiza os resultados encontrados e apresenta os dados em uma forma visual, que facilita a identificação e análise das possíveis conexões entre as entidades investigadas. Essa visualização destaca agrupamentos e relações identificadas, tornando o processo de avaliação dos indícios mais ágil e acessível para análises exploratórias.

Este capítulo descreve em detalhe a implementação da *Ferramenta Exploratória*.

5.1 Modelagem do Dado

5.1.1 Extração

Após a criação de uma **Spark Session**, a ferramenta inicia a primeira etapa do processo de ETL, que é a **extração** dos dados brutos. Para isso, utiliza-se o **PySpark**, uma interface do Apache Spark para o Python. O PySpark serve para processar grandes volumes de dados de forma distribuída e paralela, facilitando a execução de operações complexas de ETL (como leitura, transformação e carga) em datasets de grandes dimensões. Essa interface permite que dados armazenados em arquivos CSV sejam lidos e manipulados como estruturas distribuídas chamadas **DataFrames**, que funcionam de maneira semelhante a tabelas de banco de dados, mas com capacidade de processamento paralelo.

Durante a leitura dos arquivos CSV contendo dados das empresas, sócios e estabelecimentos, são aplicados **esquemas específicos** para interpretar corretamente as colunas, incluindo a definição de tipos de dados adequados para cada campo. Essa aplicação de esquemas assegura que os dados sejam carregados de maneira consistente, respeitando a tipagem necessária para operações posteriores. Além disso, durante a leitura, são configuradas opções de **delimitador** e **escape** para lidar com a complexidade dos arquivos CSV e assegurar a integridade dos dados extraídos, evitando problemas comuns como erros de formatação e dados incompletos.

Em preparação para a criação dos nós do grafo, conforme definido na fórmula 4.3, a ferramenta realiza uma operação de join entre as tabelas Empresas e Estabelecimento, consolidando-as em uma única tabela desnormalizada. Essa operação de junção é executada em um DataFrame Spark, garantindo uma estrutura de dados consistente e simplificada para o processamento subsequente. A tabela desnormalizada resultante permite que as entidades jurídicas e seus respectivos estabelecimentos sejam representados como nós únicos no grafo, facilitando a modelagem e análise das relações entre entidades.

5.1.2 Transformação

A fase de **transformação** envolve uma série de etapas para preparar os dados para serem organizados em um grafo. Esta fase utiliza funcionalidades do **PySpark**, juntamente com as bibliotecas **GraphX** e **GraphFrames**, para processar e estruturar os dados.

O **GraphX** é uma API do Apache Spark especificamente projetada para computação em grafos e análise de redes em larga escala. Ela permite a representação de dados como grafos, onde as entidades (nós) e suas relações (arestas) podem ser processadas de maneira distribuída e paralela. O **GraphX** foi desenvolvido para realizar operações em grafos de forma eficiente, permitindo a aplicação de algoritmos complexos, como PageRank e

outras análises de conectividade, fundamentais para identificar padrões e estruturas em redes complexas.

O **GraphFrames** é uma extensão do **GraphX** que fornece uma interface mais flexível para manipulação de grafos usando DataFrames do Spark. Diferente do **GraphX**, que opera diretamente sobre **Resilient Distributed Datasets (RDD)**, o GraphFrames utiliza DataFrames, facilitando a integração com outras operações do Spark e permitindo consultas mais avançadas usando a sintaxe SQL e expressões de consulta sobre grafos. O **GraphFrames** também permite a execução de operações avançadas de busca de padrões no grafo, chamadas **motifs**, que facilitam a identificação de subestruturas complexas dentro da rede, como caminhos e ciclos específicos, essenciais para a análise de grupos econômicos.

Na transformação dos dados de empresas, é realizada a desestruturação do campo JSON `te_dados` utilizando a estrutura de **DataFrames**. A operação é executada por meio de uma **explosão de colunas** (usando a função `selectExpr`) em combinação com um **esquema pré-definido** que especifica os tipos de dados esperados para cada subcampo do JSON. Primeiro, o campo JSON contendo informações agregadas sobre cada empresa é lido como uma string dentro do DataFrame. Em seguida, aplica-se o esquema sobre essa string para expandi-la em colunas individuais, representando atributos específicos como **porte**, **capital social** e **natureza jurídica**. A operação permite que cada atributo do JSON seja mapeado para uma coluna independente no DataFrame, onde cada coluna adquire um tipo de dado apropriado, como `StringType`, `IntegerType` ou `DoubleType`, conforme a definição do esquema. O resultado final é um DataFrame onde as informações antes contidas no campo JSON foram desmembradas em colunas separadas, facilitando o acesso direto a cada atributo e permitindo que esses dados sejam manipulados individualmente em operações subsequentes.

Para garantir que cada empresa tenha um ponto de representação único no grafo, foi aplicado um filtro para manter apenas o estabelecimento principal (matriz) de cada empresa, eliminando duplicidades resultantes de registros de filiais. Esse processo foi realizado em um DataFrame contendo os dados de estabelecimentos, onde foi aplicada uma condição específica para selecionar apenas os registros onde o campo indicador de tipo de estabelecimento é igual a "matriz". A operação foi implementada utilizando uma função de filtragem (`filter`), que percorre o DataFrame e retém somente as linhas que correspondem ao critério estabelecido. O campo que identifica o tipo de estabelecimento (matriz ou filial) foi utilizado como critério para a filtragem, garantindo que, para cada empresa, apenas o registro da matriz fosse mantido. O resultado final é um DataFrame simplificado, contendo um único ponto de representação por empresa.

Para os dados de sociedade (Sócios), a estrutura JSON no campo `te_dados` é expandida, extraindo-se atributos relevantes como país, data de entrada na sociedade e tipo de sociedade (pessoa física ou jurídica), realizada da mesma maneira que em empresa. Isso permite que cada atributo no campo `te_dados` seja acessado individualmente, facilitando a manipulação e análise dos dados no contexto das relações de sociedade.

O campo de identificação da sociedade (`id`) foi ajustado para seguir o mesmo padrão utilizado na tabela de Empresas, mantendo apenas as 8 primeiras posições (raiz do CNPJ) para representar as entidades jurídicas de forma consistente. Além disso, os campos que representam o identificador de sociedade e de empresa são renomeados para `src` (empresa) e `dst` (sociedade), adaptando a estrutura para a construção das arestas do grafo.

Neste ponto do processo, o DataFrame Sociedade, que estabelece a relação entre um sócio e uma empresa, já foi finalizado. Esse DataFrame servirá como base para a criação do grafo e será utilizado para representar as arestas do grafo, ou seja, as conexões entre entidades (Pessoas Físicas e Pessoas Jurídicas) que indicam vínculos societários. No entanto, para que o grafo seja completo, é necessário construir também o DataFrame de vértices (nós), que representará todas as entidades envolvidas.

O DataFrame de vértices é construído como uma combinação entre duas fontes: a lista de empresas (obtida a partir do DataFrame de Empresas) e todas as pessoas físicas que participam de uma sociedade (extraídas do DataFrame de Sociedade). A combinação dessas entidades, realizada com uma operação de união (`union`), permite que o grafo contenha tanto as empresas quanto os indivíduos associados a elas.

5.1.3 Gerando Embeddings de Nome Empresarial

Um dos critérios adotados pela Heurística Investigativa é a localização de nomes empresariais semelhantes (Item 3 da Tabela 4.1). Durante a fase de **Preparação e Modelagem dos Dados**, utilizamos a biblioteca *SentenceTransformers*[1] para processar os embeddings de todos os nomes fantasia e, na ausência destes, do nome empresarial, armazenando-os em arquivos Parquet. Os embeddings armazenados são utilizados para a avaliação de similaridade, com a biblioteca *Cosine-Similarity*[2] durante o processo de investigação das empresas-alvo.

A escolha do *SentenceTransformers* para gerar embeddings de sentenças baseia-se em sua capacidade de capturar o significado semântico dos nomes empresariais de maneira mais robusta e precisa. Em vez de se limitar a uma análise superficial de palavras ou caracteres, o SentenceTransformers permite representar cada nome empresarial como um

[1]SBERT.net. Sentence Transformers. 2024. URL:https://sbert.net/ (acessado em 03/06/2024)

[2]scikit-learn. Machine Learning in Python. 2024. URL:https://shorturl.at/XiQBU (acessado em 03/06/2024)

vetor denso em um espaço multidimensional, conhecido como embedding de sentença. Essa representação vetorial permite que empresas com nomes semanticamente semelhantes (mesmo com diferenças de palavras ou ordem) fiquem próximas no espaço vetorial, facilitando a identificação de similaridade semântica entre nomes que, de outra forma, poderiam ser interpretados como diferentes.

O uso de embeddings de sentenças é particularmente vantajoso para este estudo, pois possibilita uma análise mais aprofundada e detalhada das relações entre nomes empresariais. O método tradicional de comparação de strings, como o cálculo de distância de edição, pode identificar similaridades na forma, mas falha em capturar o contexto e o significado. Já os embeddings de sentenças, ao serem processados por modelos de aprendizado profundo, representam nuances semânticas, proporcionando uma base mais eficaz para identificar vínculos entre empresas com nomes similares.

```
from sentence_transformers import SentenceTransformer
from pyspark.sql.functions import udf, col
from pyspark.sql.types import ArrayType, FloatType

# Funcao para criar embeddings usando SentenceTransformer
def generate_embeddings_with_model(text, model):
    embeddings = model.encode(text)
    return embeddings.tolist()

# Criando o modelo SentenceTransformer
model = SentenceTransformer('sentence-transformers/paraphrase-
   multilingual-MiniLM-L12-v2')

# Criando a UDF com o modelo como parametro
generate_embeddings_udf = udf(lambda text:
   generate_embeddings_with_model(text, model), ArrayType(
   FloatType()))

# Gera embeddings
nomes_empresas = nomes_empresas.withColumn('embeddings',
   generate_embeddings_udf(col('nomeFantasia')))

# Persiste o DataFrame para consulta posterior
nomes_empresas.write.mode('overwrite').parquet(parquet_dir_name)
```

Listing 5.1: Função para gerar e persistir embeddings de nomes empresariais

O O código Python apresentado em *Listing 5.1* tem como objetivo criar embeddings para os nomes empresariais (nome fantasia) de uma base de dados, utilizando a biblioteca **SentenceTransformers**.

Na linha 5 é definida a função `generate_embeddings_with_model` que aplica o enconde do modelo `SentenceTransformer`, definido na linha 11, no texto recebido, retornando uma lista de embeddings para armazenamento.

O Modelo de Transformação de Sentença `paraphrase-multilingual-MiniLM-L12-v2` [30] foi selecionado devido a sua capacidade de gerar embeddings de alta qualidade que capturam o significado semântico de frases e textos curtos, como os nomes empresariais, em diversos idiomas. Esse modelo é particularmente útil para aplicações que envolvem busca e comparação de similaridade semântica, pois mapeia o texto para um espaço vetorial denso, onde frases com significados similares tendem a estar próximas entre si.

Na Linha 14 Uma *User Defined Function (UDF)* é criada para que a função de geração de embeddings (`generate_embeddings_udf`) possa ser aplicada diretamente nas colunas de um DataFrame Spark. Esta UDF recebe como entrada o texto de cada nome fantasia e retorna o embedding correspondente como uma lista de floats (ArrayType(FloatType())). Na linha 17 a UDF `generate_embeddings_udf` é aplicada à coluna nomeFantasia do DataFrame `nomes_empresas`. A função withColumn cria uma nova coluna chamada embeddings onde cada registro contém o embedding correspondente ao nome da empresa Por fim, na linha 20, `nomes_empresas` é salvo em formato Parquet para realização consultas.

5.2 Formulação Proposta

O produto final do processo de Extração, Transformação e Carga (ETL) são três DataFrames persistidos em formato Parquet, cada um preparado para desempenhar uma função específica na análise de grupos econômicos *de fato*. O primeiro DataFrame contém a relação dos sócios (pessoas físicas e jurídicas), que representa os nós do grafo. O segundo DataFrame representa as arestas (ligações entre os nós), detalhando a relação societária entre as pessoas. O terceiro DataFrame armazena os conjuntos de embeddings dos nomes empresariais, que permitem realizar buscas por similaridade de nomes entre empresas.

Esses dados foram organizados e processados especificamente para a construção do grafo proposto na Equação 4.3 e para suportar a execução da heurística investigativa. A estrutura em grafo permite uma visualização detalhada das relações e facilita a aplicação de algoritmos que identificam padrões de interdependência entre entidades.

Além disso, este artefato possui um módulo de pesquisa que utiliza os recursos do Spark e do GraphFrames para realizar consultas e análises complexas, destinadas à validação da formulação proposta. Esse módulo de pesquisa permite a aplicação da heurística

de similaridade e conexões societárias, facilitando a identificação e análise dos grupos econômicos *de fato* com alto desempenho e escalabilidade.

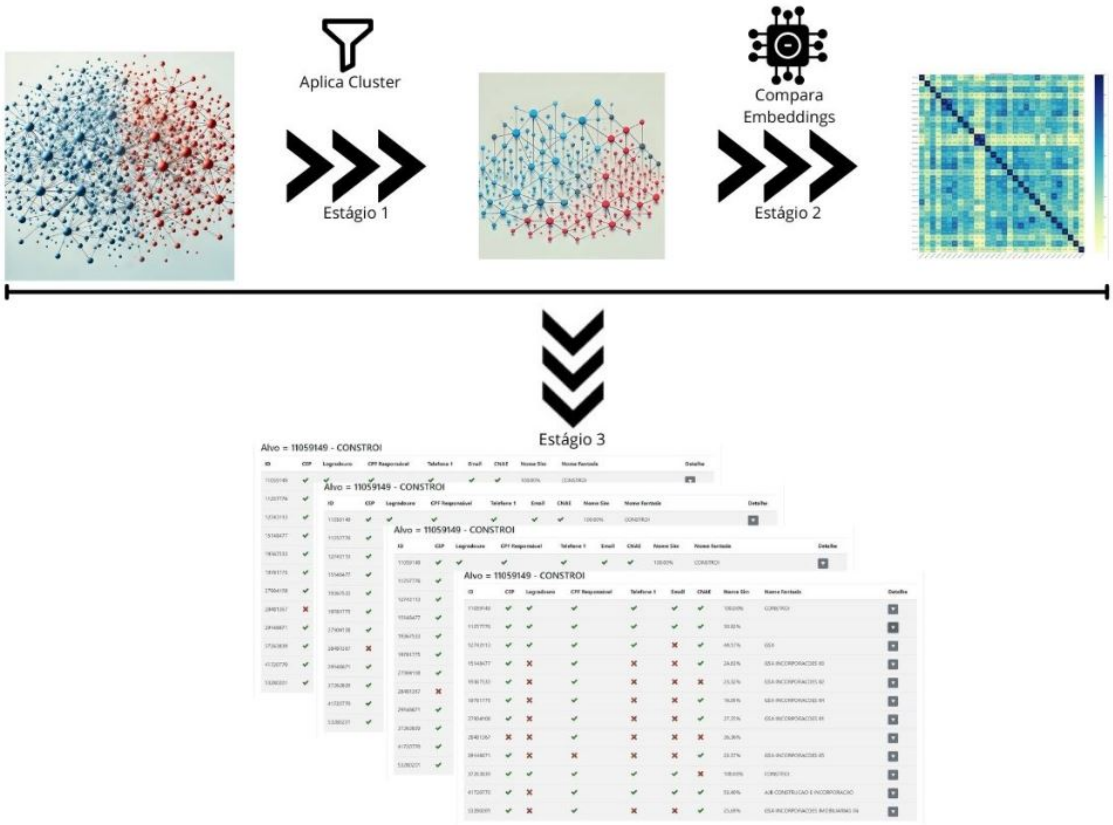


Figura 5.1: Fluxo Operacional da Heurística Exploratória. Fonte: o próprio autor.

A formulação proposta busca definir um fluxo operacional para proceder a investigação de identificação de Grupos Econômicos *de Fato* (Figura 5.1), utilizando como base as regra da heurística estabelecida a partir das características identificadas em [9]. Para guiar a aplicação da heurística, foram mapeadas as características a serem observadas, associadas aos dados disponíveis no CNPJ, conforme indicado na Tabela 4.1. Esses critérios são utilizados para a criação de um fluxo operacional que realiza, de maneira automatizada, a localização de empresas suspeitas de formação de Grupos Econômicos *de Fato* com empresas-alvo informadas. O fluxo operacional foi desenvolvido em três estágios.

5.2.1 Primeiro Estágio

No primeiro estágio da heurística, os DataFrames `nodes` (nós do grafo), `edges` (arestas do grafo) e `embeddings` são carregados com os dados já tratados e organizados (Linhas de 1 a 2 no *Listing* 5.2). O Grafo é montado com os `nodes` e os `edges` (linha 5 no *Listing* 5.2).

Nesse momento todos as empresas e seus quadros societários estão contidos na estrutura definida.

```
1  nodes = spark.read.parquet(parquet_nodes)
2  edges = spark.read.parquet(parquet_edges)
3  embeddings = spark.read.parquet(parquet_embeddings)
4
5  g = GraphFrame(nodes, edges)
6
7  id = cnpj_alvo
8
9  pFind  = "(a)-[e]->(b)"
10 socios_target = g.find(pFind) \
11         .filter(f"(a.id = '{id}') OR (b.id = '{id}')") \
12         .select(col("a.id").alias("src"), col("a.te_dados_es.
              nomeFantasia").alias("src_nome"), col("a.te_dados_es.
              cep").alias("src_cep"),
13         col("b.id").alias("dst"), col("b.te_dados_es.nomeFantasia
              ").alias("dst_nome"), col("b.te_dados_es.cep").alias("
              dst_cep"),
14         col("b.id_tipoPessoa").alias("id_tipoPessoa")) \
15         .distinct()
```

Listing 5.2: Consulta Usando Motif

Uma empresa é previamente definida como empresa-alvo, sendo o principal objeto de investigação. Neste ponto, todas as empresas que integram o quadro societário da empresa-alvo principal são também incluídas como empresas-alvo adicionais para análise.

Para identificar os sócios da empresa-alvo, foi realizada uma consulta específica no grafo utilizando o recurso de motif da biblioteca GraphFrames (Linhas de 9 a 15 no *Listing* 5.2), que permite buscar padrões entre os nós e as arestas do grafo. No código, o padrão de busca é definido pela variável `pFind`, que especifica `(a)-[e]->(b)`, indicando uma relação direcionada onde um nó `a` está conectado a um nó `b` através de uma aresta `e`.

A consulta `g.find(pFind)` aplica esse padrão ao grafo `g`, e, em seguida, é filtrada para incluir apenas as conexões onde `a.id` ou `b.id` correspondem ao identificador da empresa-alvo (`id`), com a condição `filter(f"(a.id = 'id') OR (b.id = 'id')")`. Essa filtragem permite que o grafo retorne as conexões diretas entre a empresa-alvo e outras entidades, capturando assim os sócios e demais empresas relacionadas.

Após a filtragem, os resultados são refinados com a seleção de colunas específicas. Os atributos `id`, `nomeFantasia` e `cep` são extraídos tanto para o nó de origem `(a)` quanto para o nó de destino `(b)`

Com base nessa seleção inicial, é criado um segmento (ou cluster) dentro do universo de empresas registradas no CNPJ, aplicando-se os critérios de similaridade previamente definidos para identificação de grupos econômicos *de fato*. Entre os atributos analisados, alguns são especialmente adequados para aplicação de filtros (*cep, e-mail principal, telefone, cpf do respeonsável)* que auxiliam na criação de um cluster segmentado, gerando um grafo com um universo de empresas reduzido e mais focado para a investigação.

5.2.2 Segundo Estágio

No segundo estágio, é realizada uma análise comparativa da similaridade semântica entre os nomes das empresas-alvo e os nomes das empresas contidas no cluster formado no primeiro estágio. Este estágio envolve a comparação entre os embeddings dos nomes empresariais das empresas-alvo e das empresas que compartilham características comuns, como o mesmo CEP, ampliando a busca para entidades localizadas próximas umas das outras, por exemplo.

Esse processo de análise semântica utiliza os embeddings previamente gerados e armazenados, permitindo avaliar a proximidade entre os nomes das empresas com base na similaridade de significado, e não apenas na correspondência exata de caracteres. Esse critério ajuda a identificar empresas que, embora tenham variações nos nomes, podem fazer parte de um mesmo grupo econômico *de fato*, facilitando a identificação de vínculos ocultos ou indiretos.

Neste momento, é montada uma matriz de similaridade, que consiste em uma estrutura de dados utilizada para comparar o grau de semelhança entre pares de elementos. No caso específico, a matriz cruza o nome das empresas-alvo com o nome das empresas pertencentes ao cluster, utilizando um método de comparação de similaridade, com o `cosine-similarity`.

Após a construção da matriz de similaridade, é aplicado um filtro para selecionar apenas as empresas cujo escore de similaridade seja superior a 90%, ou a um outro percentual pré-estabelecido conforme os parâmetros de interesse do investigador. Somente as empresas que atendem a esse critério passam para o próximo estágio, onde os demais critérios heurísticos definidos na 4.1 serão aplicados para refinar ainda mais o processo de identificação de potenciais Grupos Econômicos *de Fato*.

5.2.3 Terceiro Estágio

No terceiro estágio, os resultados da aplicação dos filtros de similaridade e dos critérios heurísticos são apresentados em tabelas (Figura 6.2) organizadas conforme os atributos da Tabela 4.1. Cada tabela compara as características de interesse entre a empresa-alvo e todas as empresas filtradas, oferecendo ao investigador uma visão clara das relações entre as entidades analisadas.

Esse formato facilita a identificação de empresas com indícios de participação em grupos econômicos com a empresa-alvo. A disposição visual das tabelas permite que o analista compare rapidamente as entidades, evidenciando as similaridades e diferenças em cada atributo relevante.

As comparações são realizadas de forma cruzada, em uma análise `n × m`, onde `n` é o número de empresas eleitas como alvo e `m` representa as empresas identificadas após a aplicação da Heurística Exploratória.

Adicionalmente, os resultados de similaridade de nomes são particularmente importantes, pois ajudam a identificar vínculos entre empresas com nomes semanticamente relacionados, mesmo quando esses não são idênticos. Esse critério destaca relações entre empresas registradas sob nomes diferentes, mas que compartilham um contexto ou finalidade comercial, sugerindo uma possível ligação econômica que pode não ser aparente apenas pela estrutura societária. Esse formato visual, fundamentado nas características listadas, facilita a localização de empresas semelhantes, permitindo ao analista identificar potenciais vínculos de maneira mais eficiente e precisa.

5.3 Experimentos e Resultados

Compreender as limitações das metodologias atuais de recuperação de ativos é essencial para o aprimoramento das ferramentas investigativas. O sistema SNIPER, principal plataforma institucional voltada à identificação de vínculos patrimoniais, concentra-se predominantemente na detecção de grupos econômicos legalmente constituídos. No entanto, essa abordagem não contempla, de forma automatizada, a localização de outras empresas que possam integrar grupos econômicos *de fato*, exigindo do especialista um esforço adicional para expandir manualmente a investigação e identificar conexões ocultas entre entidades.

Com o objetivo de superar essa limitação, este trabalho propõe o uso de uma Ferramenta Exploratória que emprega mecanismos heurísticos para ampliar o escopo da identificação de vínculos societários. Para validar a eficácia dessa abordagem, foi realizada uma avaliação comparativa entre os resultados obtidos com o SNIPER e os gerados pela Ferramenta Exploratória, a partir de uma amostra previamente extraída segundo

a metodologia descrita na Seção 5.3.2. Essa avaliação permite demonstrar a capacidade da ferramenta proposta em identificar relações econômicas não detectadas pelas soluções tradicionais.

5.3.1 Definição da Linha de Base

A principal dificuldade nos processos de recuperação de ativos está na identificação de grupos econômicos, etapa essencial para garantir a efetividade da fase de execução judicial. O sistema SNIPER, em sua concepção original, fundamenta-se na consulta de vínculos empresariais formalmente registrados — ou seja, aqueles que configuram Grupos Econômicos *de Direito*. Dessa forma, sua atuação está restrita à identificação de relações declaradas oficialmente, não alcançando empresas que possam estar funcionalmente vinculadas, mas que não possuem registro formal dessa associação.

Em contraposição, a Ferramenta Exploratória utiliza uma abordagem baseada em heurísticas para expandir a detecção de interconexões empresariais, com o objetivo de identificar outras empresas que possam, de forma razoável, integrar um grupo econômico mais amplo. A metodologia proposta busca revelar conexões informais, mas relevantes, que normalmente demandariam investigação manual por parte dos especialistas.

Para avaliar a eficácia comparativa dos dois sistemas, foi utilizada uma amostra extraída conforme a metodologia apresentada na Seção 5.3.2. Essa amostra foi processada separadamente no SNIPER e na Ferramenta Exploratória, permitindo uma análise comparativa dos resultados obtidos por cada abordagem.

O principal aspecto considerado nesta avaliação de linha de base foi o **escopo de detecção**: enquanto o SNIPER identifica apenas as relações formalmente registradas, a Ferramenta Exploratória busca ampliar o grupo econômico identificado por meio da detecção de vínculos indiretos. A mensuração dessa expansão permite avaliar o potencial de automação no processo de identificação de grupos econômicos *de fato*.

Outro ponto de interesse nesta avaliação é a análise da suscetibilidade dos diferentes estratos de grupos econômicos à identificação de empresas adicionais. A influência da ocorrência de novas entidades detectadas, conforme a quantidade de empresas originalmente pertencentes a cada grupo econômico, deve ser examinada para que se possa identificar padrões relevantes na distribuição desses resultados.

Os achados decorrentes dessa análise comparativa são detalhados nas seções seguintes, nas quais se avalia a efetividade da abordagem heurística no processo de ampliação da identificação de grupos econômicos.

5.3.2 Seleção da Amostra e Alocação de Neyman

O conjunto de dados utilizado neste estudo foi extraído do Cadastro Nacional da Pessoa Jurídica (CNPJ), considerando todas as empresas registradas com sede no Distrito Federal. Esse conjunto abrange uma ampla variedade de estruturas societárias, incluindo desde microempresas com apenas um sócio até grandes corporações com múltiplos sócios. Dada a elevada concentração de empresas com apenas um sócio, foi necessário aplicar um processo de filtragem para garantir uma distribuição amostral mais equilibrada.

Tratamento de Outliers Uma parcela significativa da base era composta por empresas com apenas um sócio pessoa jurídica. Para mitigar a super-representação desse grupo sem comprometer a integridade estatística, foi adotada uma abordagem de amostragem truncada (*trimmed sampling*). Nesse processo, manteve-se uma fração representativa das empresas com apenas um sócio, garantindo sua presença na amostra final, ao mesmo tempo em que se removeram os valores extremos na outra extremidade da distribuição (acima do percentil 99), reduzindo o impacto de empresas com número excepcionalmente alto de sócios.

Estratificação e Alocação de Neyman Para construir uma amostra estatisticamente robusta, foi aplicada a técnica de *Alocação de Neyman* [31], que realiza a alocação ótima de tamanhos amostrais entre estratos com base em sua variabilidade. A estratificação foi definida da seguinte forma:

- **Baixo**: empresas com 1 a 2 sócios;
- **Médio**: empresas com 3 a 5 sócios;
- **Alto**: empresas com 6 ou mais sócios.

A fórmula utilizada para o cálculo da alocação ótima é dada por:

$$n_h = \frac{N_h S_h}{\sum_h N_h S_h} \cdot n_{total}, \tag{5.1}$$

em que:

- n_h é o tamanho da amostra para o estrato h;
- N_h é o tamanho da população do estrato h;
- S_h é o desvio-padrão do estrato h;
- n_{total} é o tamanho total da amostra ($n_{total} = 300$).

A Tabela 5.1 apresenta a alocação obtida, assegurando que os grupos com maior variabilidade recebam uma fração proporcionalmente maior da amostra, o que contribui para o aumento do poder estatístico das análises subsequentes.

Tabela 5.1: Alocação Amostral com Grupos Ajustados segundo Neyman

Grupo	**População (N_h)**	**Desvio-Padrão (S_h)**	**Tamanho da Amostra (n_h)**
Alto	340	3,6579	141
Baixo	1782	0,4981	101
Médio	701	0,7296	58

Essa amostra estratificada garante uma representação adequada dos diferentes perfis de complexidade societária e assegura maior poder explicativo para as análises que seguem.

5.3.3 Resultados e Análise

Os resultados obtidos com a aplicação da Ferramenta Exploratória foram comparados com aqueles produzidos pelo SNIPER, a fim de avaliar a eficácia de cada abordagem na identificação de grupos econômicos. As métricas centrais analisadas incluem a quantidade total de entidades identificadas e sua distribuição entre diferentes estratos econômicos.

Comparação das Entidades Identificadas A Tabela 5.2 apresenta o total de entidades identificadas por cada sistema. A Ferramenta Exploratória demonstrou ser capaz de expandir os grupos econômicos ao identificar conexões adicionais que não haviam sido detectadas previamente pelo SNIPER.

Tabela 5.2: Comparativo de Entidades Identificadas pelo SNIPER e pela Ferramenta Exploratória

Sistema	**Total de Entidades Identificadas**
SNIPER	2101
Ferramenta Exploratória	2353

Impacto nos Grupos Estratificados A quantidade de empresas adicionais identificadas variou conforme o estrato econômico analisado. A Figura 5.2 apresenta o impacto da Ferramenta Exploratória sobre os diferentes grupos, demonstrando um aumento consistente na identificação de entidades em todos os níveis de complexidade.

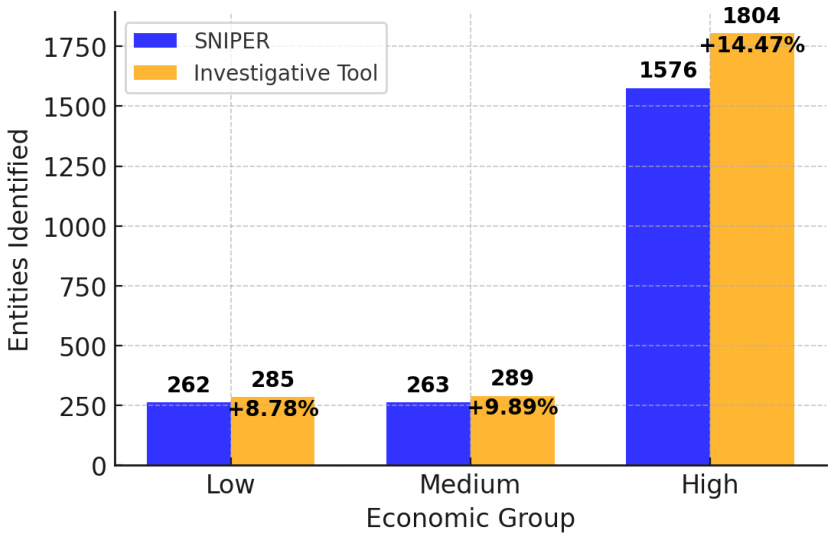


Figura 5.2: Aumento no Número de Entidades Identificadas pela Ferramenta Exploratória em Diferentes Grupos Econômicos. Fonte: o próprio autor.

5.3.4 Validação Estatística dos Resultados

Para assegurar a confiabilidade das diferenças observadas entre o SNIPER e a Ferramenta Exploratória, foram conduzidos testes estatísticos com o intuito de verificar se o aumento na quantidade de entidades identificadas possui significância estatística.

Foi aplicado um **teste t pareado** para comparar o número de entidades identificadas por cada método. O resultado apresentou um **valor de** p de $5{,}96 \times 10^{-8}$, valor significativamente inferior ao limiar convencional de 0,05, indicando que a diferença observada é estatisticamente significativa. O teste t pareado é amplamente utilizado para testar hipóteses relacionadas à média de dois grupos relacionados.

Para confirmar esses achados, também foi realizado um **teste não paramétrico de Wilcoxon para amostras pareadas**, que produziu um **valor de** p de $1{,}76 \times 10^{-17}$. Esse teste é especialmente apropriado quando não se pode assumir a normalidade da distribuição. Seus resultados reforçam a conclusão de que a Ferramenta Exploratória, de forma consistente, identifica mais entidades do que o SNIPER.

Além disso, foi calculado um **intervalo de confiança de 95%** para a diferença média entre os dois métodos, resultando em um intervalo de (0,60, 1,25). Isso significa que, em média, a Ferramenta Exploratória identifica entre **0,60 e 1,25 entidades a mais por investigação** do que o SNIPER, com 95% de confiança. Intervalos de confiança oferecem uma faixa plausível de valores para a diferença real entre as médias, permitindo uma inferência estatística mais robusta [32].

A Tabela 5.3 resume os resultados obtidos nos testes estatísticos aplicados.

Tabela 5.3: Validação Estatística do Desempenho da Ferramenta Exploratória

Teste	**Valor de p**	**Intervalo de Confiança (95%)**
Teste t pareado	$5{,}96 \times 10^{-8}$	(0,60, 1,25)
Teste de Wilcoxon	$1{,}76 \times 10^{-17}$	–

Esses testes estatísticos confirmam que o aumento na identificação de entidades promovido pela Ferramenta Exploratória não decorre de variação aleatória, mas sim de uma melhoria real e significativa em relação ao SNIPER. Os baixos valores de p obtidos em ambos os testes oferecem forte evidência para rejeição da hipótese nula de que não há diferença entre os dois métodos.

5.4 Conclusão Parcial

O principal aspecto considerado nesta avaliação de linha de base foi o **escopo de detecção**: enquanto o SNIPER identifica apenas as relações formalmente registradas, a Ferramenta Exploratória busca ampliar o grupo econômico identificado por meio da detecção de vínculos indiretos. A mensuração dessa expansão permite avaliar o potencial de automação no processo de identificação de grupos econômicos *de fato*. A Ferramenta Exploratória desenvolvida neste trabalho tem como principal objetivo automatizar a identificação de indícios de Grupos Econômicos *de Fato* a partir de dados estruturados do Cadastro Nacional da Pessoa Jurídica (CNPJ), oferecendo uma ampliação significativa das funcionalidades atualmente disponíveis no sistema SNIPER. A adoção do framework CRISP-DM permitiu estruturar o processo de mineração de dados em fases bem definidas, assegurando sistematização, reprodutibilidade e escalabilidade.

Ao aplicar técnicas de processamento paralelo com Apache Spark e modelagem em grafos com GraphFrames, foi possível transformar grandes volumes de dados relacionais em uma estrutura orientada a grafos, mais apropriada para a análise de interdependências complexas. A incorporação de modelos de similaridade semântica com embeddings de nomes empresariais, utilizando Sentence-BERT, introduziu uma abordagem inovadora na identificação de vínculos não evidentes entre entidades jurídicas.

O desenvolvimento do artefato foi concluído, abrangendo o processo de extração e transformação dos dados, a aplicação das heurísticas de pesquisa propostas para localizar as conexões relevantes, e a formatação da saída dos dados para uma visualização clara.

A Ferramenta foi submetida a um processo experimental, no qual se comparou sua capacidade de detecção com a do SNIPER. Os resultados demonstraram que a nova abordagem é capaz de identificar um número significativamente maior de entidades potencialmente integrantes de grupos econômicos *de fato*. A validação estatística reforçou

esses achados, com testes como o t pareado e o teste de Wilcoxon indicando diferenças significativas entre os dois métodos. A média de expansão na identificação de entidades oscilou entre 0,60 e 1,25 empresas adicionais por investigação, com 95% de confiança.

Com base nos resultados obtidos, conclui-se que a Artefato é eficaz e representa um avanço relevante na investigação automatizada de Grupos Econômicos *de Fato*, contribuindo para o fortalecimento da capacidade estatal de enfrentamento à ocultação patrimonial.

Capítulo 6

Artefato 2 - Interface Investigatória

Devido ao caráter sigiloso que rege os inquéritos e as investigações patrimoniais conduzidas por órgãos públicos, não há disponibilidade de um banco de dados consolidado que contenha os resultados prévios das análises que levaram à desconsideração da personalidade jurídica em casos concretos. Em razão dessa limitação, torna-se inviável a comparação direta entre os resultados gerados pelo Artefato 1 e decisões ou investigações pretéritas já concluídas.

Para contornar essa restrição e validar empiricamente a proposta desenvolvida, foi implementado o Artefato 2 — a Interface Investigatória — com o objetivo de fornecer uma camada de acesso simplificada à Ferramenta Exploratória (Artefato 1) e permitir sua utilização prática por especialistas do domínio. A interface permite que o usuário forneça os CNPJ das entidades-alvo, os quais são processados automaticamente e comparados com os dados persistidos, apresentando tabelas estruturadas que destacam similaridades e vínculos relevantes entre as empresas investigadas.

Além de facilitar a utilização da heurística investigativa, a Interface Investigatória foi desenvolvida como instrumento para coleta de avaliações qualitativas e quantitativas junto a um grupo de especialistas, composto por juízes e servidores públicos federais com experiência em investigação patrimonial. Durante um período de 14 dias, foram realizadas 160 investigações com o uso da ferramenta. Ao final de cada interação, os especialistas avaliaram a ferramenta em quatro dimensões: *Precisão e Eficácia*, *Usabilidade e Eficiência*, *Impacto na Investigação Patrimonial* e *Comparação com outras ferramentas*. Também foi aplicada a métrica *Net Promoter Score (NPS)* para aferir a probabilidade de recomendação da ferramenta a outros profissionais.

Essas avaliações permitiram obter uma visão abrangente sobre a utilidade prática, a eficácia percebida e a aceitabilidade da ferramenta no contexto real de uso, oferecendo subsídios valiosos para o aperfeiçoamento do sistema e para o desenvolvimento de futuras versões com base em evidências empíricas e feedback qualificado.

6.1 Metodologia de Implementação Técnica

O desenvolvimento da Interface Investigatória adotou uma arquitetura orientada a serviços, composta por três camadas principais: persistência de dados, serviços de backend e interface de acesso ao usuário. Essa estrutura visa permitir a interação eficiente com os dados previamente processados pelo Artefato 1, oferecendo aos especialistas uma ferramenta prática e acessível para análise investigatória.

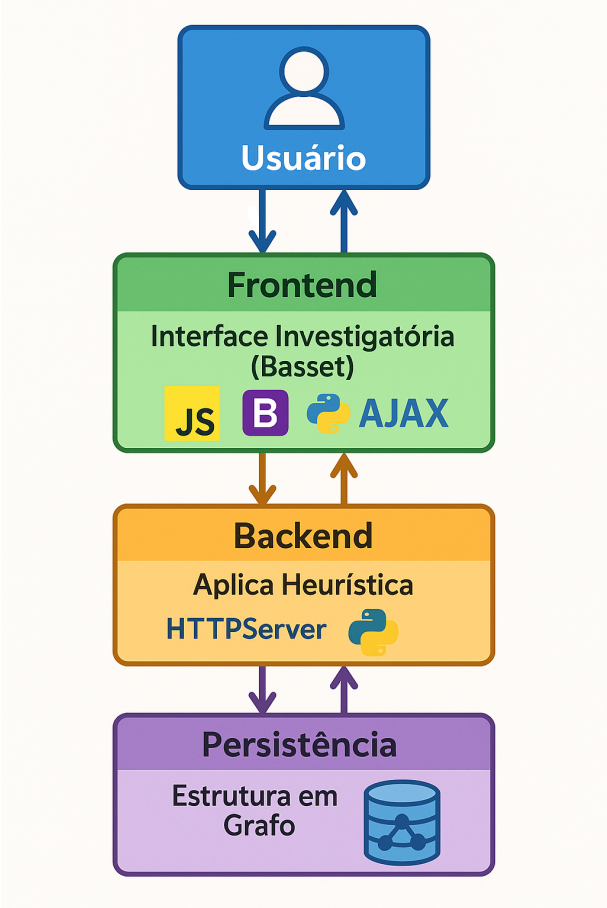


Figura 6.1: Arquitetura do Artefato 2. Fonte: o próprio autor.

A Figura 6.1 apresenta a organização modular da solução. Na base, encontra-se a camada de persistência, responsável por armazenar os dados do grafo empresarial em um banco PostgreSQL. Esses dados derivam dos DataFrames gerados no Artefato 1 e foram estruturados para otimizar as consultas frequentes realizadas durante as investigações. A camada intermediária corresponde ao backend, implementado em Python com suporte a autenticação por tokens, controle de sessão e processamento das requisições de consulta e comparação de CNPJs. Por fim, a camada de frontend disponibiliza uma interface web intuitiva, desenvolvida em HTML, JavaScript e Bootstrap, permitindo que os especialis-

tas realizem consultas inserindo CNPJs ou CPFs e visualizem os resultados em tabelas dinâmicas e comparativas.

Cada camada é descrita em mais detalhes nas subseções a seguir, abordando sua função e tecnologias utilizadas, mas sem aprofundamento técnico exaustivo. Essa abordagem modular permitiu uma implementação flexível e escalável, além de facilitar a manutenção do sistema e futuras integrações com novas funcionalidades, como a coleta de avaliações dos usuários e a geração de indicadores de uso.

6.1.1 Camada de Banco de Dados

A camada de banco de dados da Interface Investigatória foi concebida com o objetivo de prover consultas rápidas e eficientes sobre os dados processados no Artefato 1. Para isso, foi adotado o Sistema Gerenciador de Banco de Dados PostgreSQL, em razão de sua robustez, ampla compatibilidade com ferramentas externas e elevada performance em operações analíticas.

A modelagem dos dados buscou aproveitar integralmente o pipeline de ETL do Artefato 1, estruturando as informações extraídas dos arquivos Parquet em três principais tabelas relacionais:

- **nodes**: armazena os nós do grafo, representando pessoas físicas e jurídicas, com aproximadamente 75 milhões de registros. Cada entrada contém dois campos do tipo `JSONB`, que concentram os dados cadastrais da empresa (`te_dados_em`) e do estabelecimento (`te_dados_es`).
- **edges**: representa as conexões societárias entre os nós, totalizando cerca de 25 milhões de registros. As conexões são registradas com informações adicionais no campo `te_dados_sc`, também em formato `JSONB`.
- **embeddings**: contém vetores de embeddings de nomes empresariais e nomes fantasia, com cerca de 59 milhões de registros. Cada vetor é armazenado como array de `double precision`, associado ao identificador da empresa.

A modelagem adotou um formato desnormalizado, utilizando colunas do tipo `JSONB` para permitir acesso direto e flexível aos atributos relevantes, sem a necessidade de múltiplos joins. Essa decisão foi motivada pelo perfil analítico e de leitura intensiva da aplicação, em que os dados são carregados periodicamente de fontes externas e não sofrem atualizações interativas.

Com o intuito de otimizar a performance, não foram estabelecidas chaves estrangeiras entre as tabelas. As regras tradicionais de integridade referencial foram dispensadas,

pois a origem dos dados já assegura consistência estrutural suficiente para o propósito investigativo da ferramenta.

Adicionalmente, foram criados índices específicos para acelerar as consultas mais frequentes:

- Índices sobre os campos `id` e `cep` da tabela `nodes`;
- Índices sobre os campos `src` e `dst` da tabela `edges`;
- Índice funcional para filtros por Unidade da Federação, utilizando a extração do campo `uf` de dentro do `JSONB` `te_dados_es` com a função `LOWER()`.

A opção pelo PostgreSQL em detrimento do Apache Spark para esta fase do projeto se justifica pelo padrão de acesso mais interativo exigido pela interface. Consultas SQL tradicionais são mais apropriadas para as interações realizadas por especialistas durante a navegação, enquanto o Spark permanece como base para o processamento em larga escala no pipeline de ingestão e transformação.

6.1.2 Camada de Serviços

A camada de serviços é responsável por receber as requisições do frontend, coordenar o processamento dos dados de entrada e retornar os resultados consolidados. Seu papel principal é executar o fluxo operacional da Ferramenta Exploratória utilizando os dados previamente persistidos no banco PostgreSQL.

O backend foi implementado em Python, utilizando a biblioteca `http.server` como base para a definição das rotas e controle do fluxo. O serviço central expõe três rotas principais:

- `/login`: recebe um par `usuário/senha`, valida contra as credenciais definidas em arquivo de configuração e retorna um `token` de autenticação com validade de 6 horas;
- `/search`: recebe uma lista de CNPJs em formato JSON, valida o token e aciona a lógica de comparação de entidades, retornando um dicionário JSON contendo os resultados para cada CNPJ alvo;
- `/status`: rota auxiliar usada para verificação do servidor, permitindo confirmar o funcionamento da aplicação.

A autenticação é gerenciada pelo módulo `auth.py`, com controle de expiração automático, garantindo que sessões inativas sejam descartadas. A lógica principal de análise está concentrada no módulo `query.py`, onde ocorre a execução sequencial de etapas que correspondem diretamente à heurística investigativa proposta no Capítulo 5.2:

1. **Primeiro Estágio – Coleta das Conexões Diretas**: a partir do CNPJ alvo fornecido, o sistema identifica todas as conexões diretas (arestas) e os nós vizinhos correspondentes, utilizando filtros por `src` ou `dst` para mapear o quadro societário e suas ramificações imediatas;
2. **Segundo Estágio – Análise de Similaridade Semântica**: os nomes das empresas localizadas na etapa anterior são comparados com o nome da empresa-alvo usando vetores de embeddings armazenados na tabela `embeddings`. A métrica de similaridade utilizada é o *cosine similarity*, e apenas entidades com similaridade acima de um limiar configurável (por exemplo, 0,9) são mantidas para análise;
3. **Terceiro Estágio – Consolidação dos Resultados**: as entidades selecionadas são organizadas em estruturas tabulares, comparando atributos como CNAE, telefone, e-mail, CEP e razão social. Essa organização visa facilitar a interpretação dos dados e a identificação visual de potenciais grupos econômicos *de fato*.

Os dados retornados pelo backend incluem, para cada empresa alvo, uma lista de entidades suspeitas emparelhadas, junto com suas similaridades nos campos previamente definidos na Tabela 4.1. O resultado é entregue em formato JSON estruturado, pronto para ser renderizado na interface web.

Essa arquitetura garante baixa latência nas respostas, suporte a múltiplos usuários simultâneos e flexibilidade para ajustes futuros na heurística ou nas métricas de similaridade.

6.1.3 Interface Web e Visualização dos Resultados

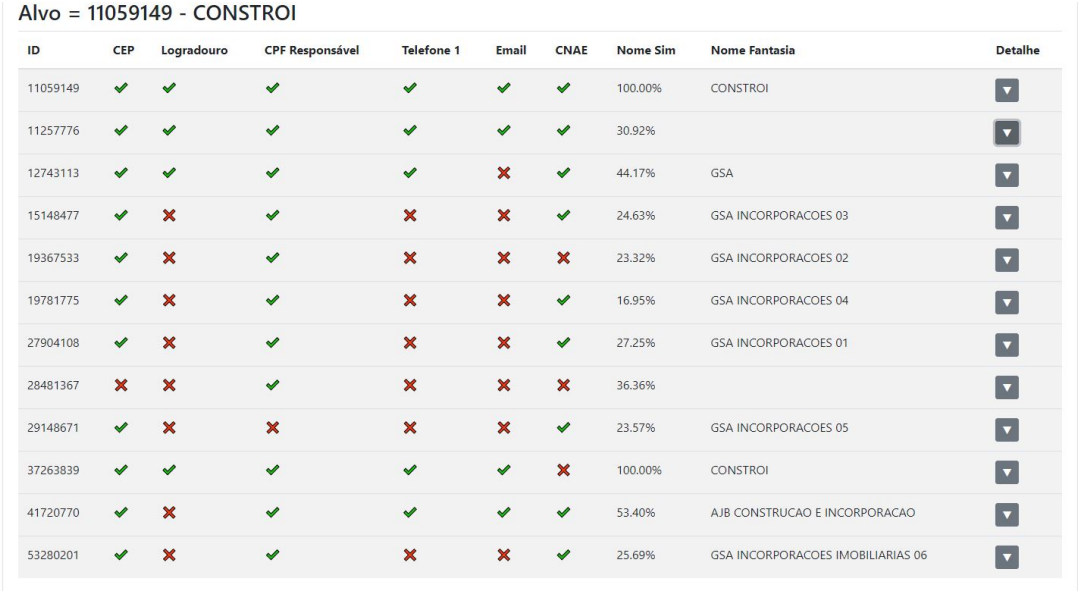
A camada de acesso da Interface Investigatória é composta por uma aplicação web leve, desenvolvida em HTML, JavaScript e Bootstrap, com o objetivo de proporcionar uma interação simples e eficiente para os especialistas responsáveis pela análise. Essa interface é o principal ponto de entrada do sistema para os usuários finais e foi projetada para operar de forma responsiva, sem necessidade de instalação local ou dependências externas.

A interface permite que o usuário autentique-se com um par de credenciais e, após o login, insira um ou mais CNPJs como entidades-alvo para investigação. O formulário de entrada é validado em tempo real no navegador, garantindo que os dados enviados estejam corretamente formatados. Em seguida, a requisição é enviada para o backend via chamada assíncrona (AJAX), sendo processada e respondida em formato JSON.

O conteúdo retornado é interpretado pelo script `common.js`, que monta dinamicamente as tabelas de comparação. Para cada CNPJ alvo, é exibida uma tabela detalhada que relaciona os atributos da empresa investigada com os das entidades suspeitas, identificadas com base na heurística descrita na seção anterior. Os atributos apresentados

incluem similaridade semântica de nome, CNAE, telefone, e-mail, CEP, nome fantasia e composição societária.

A Figura 6.2 ilustra a estrutura da tabela gerada para cada investigação. Cada linha representa uma empresa suspeita, acompanhada de um conjunto de indicadores que auxiliam na inferência de vínculos informais, característica essencial na identificação de grupos econômicos *de fato*.



Alvo = 11059149 - CONSTROI

ID	CEP	Logradouro	CPF Responsável	Telefone 1	Email	CNAE	Nome Sim	Nome Fantasia	Detalhe
11059149	✔	✔	✔	✔	✔	✔	100.00%	CONSTROI	
11257776	✔	✔	✔	✔	✔	✔	30.92%		
12743113	✔	✔	✔	✔	✘	✔	44.17%	GSA	
15148477	✔	✘	✔	✘	✘	✔	24.63%	GSA INCORPORACOES 03	
19367533	✔	✘	✔	✘	✘	✘	23.32%	GSA INCORPORACOES 02	
19781775	✔	✘	✔	✘	✘	✔	16.95%	GSA INCORPORACOES 04	
27904108	✔	✘	✔	✘	✘	✔	27.25%	GSA INCORPORACOES 01	
28481367	✘	✘	✔	✘	✘	✘	36.36%		
29148671	✔	✘	✘	✘	✘	✔	23.57%	GSA INCORPORACOES 05	
37263839	✔	✔	✔	✔	✔	✘	100.00%	CONSTROI	
41720770	✔	✘	✔	✔	✔	✔	53.40%	AJB CONSTRUCAO E INCORPORACAO	
53280201	✔	✘	✔	✘	✘	✔	25.69%	GSA INCORPORACOES IMOBILIARIAS 06	

Figura 6.2: Tabela de Comparação. Fonte: o própio autor.

A interface também inclui um menu de ajuda com orientações para o preenchimento dos campos e interpretação dos resultados, além de uma opção para reiniciar a investigação ou encerrar a sessão de forma segura. Essa camada foi pensada para maximizar a autonomia dos especialistas, mesmo sem familiaridade prévia com ferramentas técnicas ou sistemas de consulta a grafos.

Ao manter a lógica de apresentação separada da lógica de negócio e do armazenamento de dados, a camada de acesso favorece a manutenção evolutiva do sistema e a personalização futura da interface, como por exemplo para diferentes perfis de usuários ou contextos institucionais.

6.2 Avaliação do Artefato 2

A avaliação da Interface Investigatória foi realizada com o objetivo de validar tanto a eficácia da heurística de localização de Grupos Econômicos *de Fato* quanto a usabilidade e aplicabilidade da interface desenvolvida. Foram conduzidas duas frentes de análise

complementares: a avaliação técnica da heurística baseada na matriz de confusão, e a avaliação subjetiva da ferramenta pelos especialistas, por meio de um formulário estruturado (Anexo I).

6.2.1 Avaliação da Heurística

Para avaliar a capacidade da heurística em identificar conexões patrimoniais ocultas, foi formado um grupo de 12 especialistas em investigação patrimonial, incluindo juízes e servidores públicos federais. Esses profissionais realizaram 160 investigações reais utilizando a interface ao longo de 14 dias. Cada investigação consistiu na análise de um CNPJ-alvo, com base na aplicação completa dos três estágios da heurística descrita no Capítulo 5.

Ao final de cada consulta, os especialistas foram convidados a classificar os resultados segundo os quatro componentes clássicos da matriz de confusão:

- **Verdadeiro Positivo (TP)** – Empresas não formalmente vinculadas, mas corretamente identificadas como pertencentes ao grupo econômico *de fato*;
- **Falso Positivo (FP)** – Empresas sugeridas pela ferramenta que, ao final, não integravam o grupo econômico analisado;
- **Verdadeiro Negativo (TN)** – Casos em que não havia vínculos e a ferramenta corretamente não apresentou sugestões;
- **Falso Negativo (FN)** – Empresas que pertenciam ao grupo econômico *de fato*, mas não foram localizadas pela heurística.

Essas classificações possibilitaram o cálculo de métricas objetivas de desempenho da heurística, como precisão, revocação (recall) e F1-score, que serão discutidas na próxima seção.

6.2.2 Avaliação da Interface e do Artefato

A segunda vertente da avaliação consistiu na aplicação de um formulário padronizado para mensurar a percepção dos especialistas quanto à qualidade geral da ferramenta. O instrumento de coleta foi baseado em uma escala Likert de 1 a 5 pontos, e abordou os seguintes blocos temáticos:

- **Precisão e Eficácia** – Avaliação da acurácia na identificação de vínculos patrimoniais ocultos;
- **Usabilidade e Eficiência** – Facilidade de uso da interface, disposição visual dos dados e tempo de resposta;

- **Impacto na Investigação Patrimonial** – Potencial da ferramenta em ampliar a análise investigativa e dar suporte à desconsideração da personalidade jurídica;
- **Comparação com Outras Ferramentas** – Grau de superioridade percebida em relação a soluções já utilizadas pelos especialistas;
- **Confiabilidade e Redução de Esforço Manual** – Avaliação de como a ferramenta contribui para maior confiança nos dados e reduz a necessidade de verificações manuais.

Além das perguntas fechadas, o formulário incluiu questões abertas para captura de sugestões de melhoria e relato de pontos positivos percebidos. Também foi incluída a métrica *Net Promoter Score* (NPS), em que os participantes avaliaram, de 0 a 10, o quanto recomendariam a ferramenta a outros colegas da área [33]

A aplicação do formulário foi anônima, individual e realizada imediatamente após a conclusão das investigações. Os dados foram armazenados em ambiente seguro e processados com técnicas estatísticas descritivas para avaliação da consistência e dispersão das respostas.

A versão completa do instrumento utilizado encontra-se no Anexo I, sob o título *Formulário de Avaliação da Interface Investigatória.*

6.3 Resultados

Esta seção apresenta os principais resultados obtidos a partir das avaliações realizadas pelos especialistas quanto à heurística de inferência e à ferramenta como um todo. As métricas utilizadas foram tanto objetivas (baseadas na matriz de confusão) quanto subjetivas (baseadas na percepção dos usuários por meio de questionários estruturados).

6.3.1 Desempenho da Heurística

A análise dos dados classificados pelos especialistas permitiu calcular três métricas clássicas de desempenho:

- **Precisão (Precision):** 66,67%. Indica a proporção de entidades corretamente classificadas como pertencentes ao Grupo Econômico *de Fato* em relação ao total de sugestões feitas. Embora indique uma boa capacidade de acerto, revela que há espaço para reduzir falsos positivos.
- **Revocação (Recall):** 85,71%. Representa a proporção de entidades corretamente recuperadas entre todas as que efetivamente pertenciam ao grupo, indicando que a heurística tem elevada capacidade de recuperação.

- **F1-Score:** 75%. Equilíbrio entre precisão e revocação, servindo como medida composta da performance geral da heurística.

Os resultados mostram que a heurística foi eficaz em encontrar vínculos ocultos relevantes, especialmente ao apresentar alto recall. No entanto, a precisão moderada sugere que ajustes podem ser implementados para reduzir o número de falsos positivos, minimizando o esforço de triagem por parte dos especialistas. Ainda que o processo conte com validação humana ao final, a redução de ruído investigativo contribui para maior foco nos vínculos mais promissores.

6.3.2 Percepção sobre a Ferramenta

A percepção dos especialistas sobre a ferramenta foi mensurada por meio de 12 questões categorizadas em quatro blocos temáticos, utilizando escala Likert de 1 a 5. As médias por categoria, acompanhadas de seus respectivos erros padrão, são apresentadas na Figura 6.3.

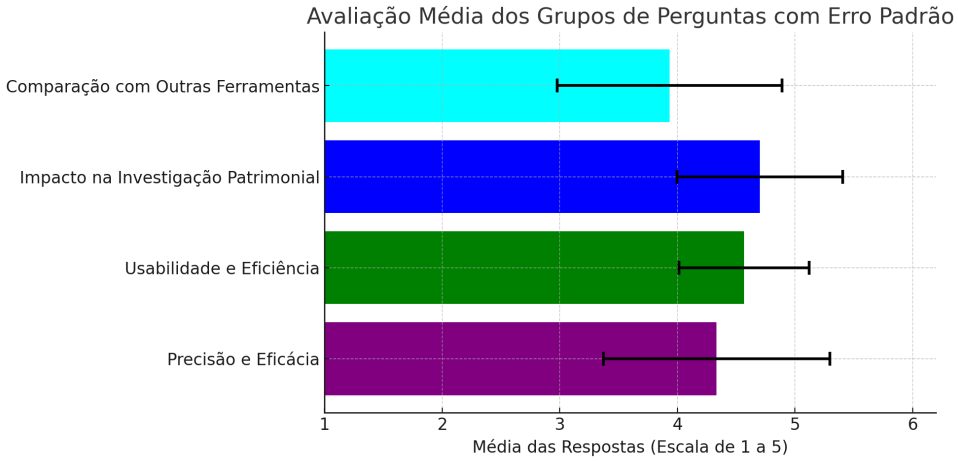


Figura 6.3: Avaliação Média dos Grupos de Perguntas com Erro Padrão Fonte: o própio autor.

- **Impacto na Investigação Patrimonial:** obteve a média mais alta (**4,70**), indicando que os especialistas reconheceram valor agregado substancial da ferramenta para as atividades de investigação.
- **Usabilidade e Eficiência:** apresentou média de **4,57**, com baixa variabilidade nas respostas, o que evidencia uma experiência de uso consistente e satisfatória.
- **Precisão e Eficácia:** registrou média de **4,33**, também positiva, mas com variação maior, sugerindo percepções distintas entre os especialistas sobre a capacidade da ferramenta em localizar vínculos precisos.

- **Comparação com Outras Ferramentas:** teve a menor média (**3,93**) e maior erro padrão, apontando que, embora bem avaliada, ainda não se consolidou como uma solução superior a outras já conhecidas no meio investigativo.

6.3.3 Net Promoter Score (NPS)

Além da avaliação categorizada, os especialistas responderam à métrica de lealdade conhecida como *Net Promoter Score* (NPS), baseada na metodologia proposta por Reichheld [33]. Nessa avaliação, os usuários informam o quanto recomendariam a ferramenta a colegas, em uma escala de 0 a 10.

O valor obtido de **80** representa um índice elevado de recomendação, indicando que a maioria dos especialistas demonstrou alto grau de satisfação e confiança na ferramenta. Esse resultado reforça a consistência das demais avaliações e sugere que a solução tem potencial para adoção institucional mais ampla em cenários de investigação patrimonial.

6.4 Conclusão Parcial

A avaliação da Interface Investigatória (Artefato II) evidenciou sua relevância como ferramenta de apoio às atividades de investigação patrimonial, especialmente no que se refere à operacionalização da heurística de localização de grupos econômicos *de fato* em ambiente interativo e acessível para especialistas.

O desempenho técnico da heurística, com destaque para o valor de revocação de 85,71%, demonstra a elevada capacidade da ferramenta em recuperar empresas relevantes envolvidas em estruturas econômicas ocultas. A precisão de 66,67% revela, contudo, a presença de falsos positivos, indicando a necessidade de ajustes nos critérios de inferência para aprimorar a assertividade. O equilíbrio entre essas métricas, refletido pelo F1-score de 75%, sugere que a heurística apresenta desempenho consistente e utilizável em cenários reais.

Do ponto de vista da experiência dos usuários, a ferramenta foi bem recebida pelos especialistas. A categoria *Impacto na Investigação Patrimonial* atingiu média de 4,70 na escala Likert, indicando uma percepção altamente positiva quanto à contribuição da ferramenta para ampliar e qualificar a análise de vínculos patrimoniais. A interface também demonstrou boa aceitação em termos de usabilidade e eficiência, com média de 4,57, sugerindo que a camada de acesso foi eficaz em traduzir os resultados complexos da heurística em visualizações compreensíveis e úteis para os usuários.

O *Net Promoter Score* (NPS) obtido foi de 80, valor elevado que denota um forte grau de satisfação e lealdade por parte dos especialistas. Esse resultado confirma que a

ferramenta não apenas atende às necessidades práticas dos usuários, como também possui potencial para adoção mais ampla por outros profissionais da área.

A ausência de bases de dados históricas com registros consolidados de decisões relacionadas à desconsideração da personalidade jurídica impediu uma comparação objetiva com decisões anteriores. Ainda assim, a avaliação baseada em uso real e julgamento especializado forneceu evidências empíricas relevantes sobre a utilidade da ferramenta no contexto da recuperação de ativos.

Esses resultados indicam que o Artefato II cumpre seu papel como componente operacional da proposta investigativa, traduzindo os dados processados pelo Artefato I em recursos acionáveis e validados por especialistas.

Capítulo 7

Considerações Finais

Devido ao caráter sigiloso dos inquéritos e dos trabalhos realizados pelos especialistas em recuperação de ativos, não foi possível realizar uma comparação direta entre os resultados obtidos pelos Artefatos 1 e 2 e os casos históricos, mesmo nos processos já transitados em julgado. A fase investigativa e as informações relacionadas à obtenção de provas permanecem restritas, o que inviabiliza a utilização de exemplos práticos concretos para validação quantitativa com base em casos reais.

Nesse contexto, a avaliação da eficácia dos artefatos foi conduzida a partir da percepção e do julgamento técnico de juízes e especialistas que contribuíram para a formulação das heurísticas utilizadas no desenvolvimento das ferramentas. Após analisarem o trabalho realizado e os resultados produzidos, esses especialistas relataram que as ferramentas atendem aos objetivos esperados e ampliam a capacidade de identificação de vínculos patrimoniais não documentados.

A análise qualitativa foi realizada por meio de um formulário de avaliação baseado na escala Likert, com número par de alternativas, de modo a eliminar a opção de neutralidade e incentivar os participantes a assumir uma posição clara quanto à eficácia dos artefatos investigatórios. Os especialistas avaliaram aspectos-chave como precisão e eficiência na identificação de grupos econômicos *de fato*, permitindo identificar pontos fortes e oportunidades de melhoria.

A aplicação do formulário contemplou um grupo de especialistas em investigação patrimonial, incluindo juízes e técnicos que atuam diretamente na recuperação de ativos. Os resultados foram analisados com abordagens estatísticas adequadas, considerando medidas como médias e desvios-padrão quando aplicável [34] [35] [36].

Essa avaliação qualitativa permitiu obter uma visão abrangente sobre a utilidade e a eficácia dos artefatos no contexto das investigações patrimoniais, confirmando sua relevância prática e fornecendo subsídios para aperfeiçoamentos futuros.

7.1 Análise de Resultados

A avaliação apresentada na Seção 5.3.3 indica que o Artefato 1 cumpriu o objetivo de ampliar o espectro de identificação de entidades potencialmente vinculadas a grupos econômicos *de fato*, quando comparado à linha de base operacional estabelecida. O contraste entre a saída do *pipeline* exploratório e o resultado obtido diretamente no SNIPER é sintetizado na Tabela 5.2, enquanto a Figura 5.2 ilustra o aumento no número de entidades identificadas em diferentes grupos, evidenciando ganhos consistentes de cobertura ao longo dos cenários analisados. Observa-se que a combinação de *features* derivadas do CNPJ (quadro societário, CNAE, endereços e *contatos*) com a similaridade semântica de nomes (embeddings baseados em Bidirectional Encoder Representations from Transformers (BERT)) contribuiu para revelar conexões não mapeadas previamente pela abordagem exclusivamente documental.

Ainda na Seção 5.3.3, a análise discute o comportamento da heurística frente à amostra delineada, destacando que os indícios de vínculo emergem tanto por sobreposição direta (sócios/representantes em comum) quanto por sinais combinados (proximidade de denominação social, atuação correlata e compartilhamento de infraestrutura). Esse arranjo favorece a recuperação de entidades adjacentes ao núcleo do grupo, sem depender apenas de relações formais explícitas.

A Seção 5.3.4 reporta a validação estatística dos achados, demonstrando que as diferenças observadas entre a linha de base e o Artefato 1 são suportadas pelos procedimentos de teste adotados. Em termos práticos, a validação afasta a hipótese de que os ganhos detectados seriam fruto de variação aleatória da amostra, fortalecendo a confiança na robustez do *pipeline* proposto. Complementarmente, a Tabela 5.3 apresenta o resumo dessa validação, alinhando os resultados às hipóteses avaliadas.

Em conjunto, a Tabela 5.2, a Figura 5.2 e as análises das Seções 5.3.3 e 5.3.4 convergem para o mesmo diagnóstico: o Artefato 1 amplia a cobertura de entidades relacionadas e traz à tona vínculos relevantes para a investigação patrimonial, mantendo um equilíbrio adequado entre abrangência e pertinência dos resultados.

Síntese do Artefato 1. O Artefato 1 entregou um *pipeline* exploratório eficaz para localizar indícios de grupos econômicos *de fato* a partir de dados públicos estruturados, integrando análise de grafos e similaridade semântica. Os resultados reforçam seu papel como etapa de triagem e expansão do escopo investigativo, reduzindo o esforço manual inicial e fornecendo insumos qualificados para a etapa seguinte de interação com especialistas e priorização de alvos.

A avaliação do Artefato 2, descrita na Seção 6.2, demonstra que a heurística incorporada na *Interface Investigatória* manteve desempenho consistente frente aos cenários

propostos, preservando a capacidade de revelar vínculos patrimoniais relevantes sem sobrecarga de resultados irrelevantes. Conforme detalhado na Seção 6.3.1, o balanceamento entre precisão e abrangência foi alcançado a partir da combinação das *features* estruturadas com os indícios semânticos identificados na etapa exploratória, resultando em listas de entidades priorizadas para análise do especialista.

Na Seção 6.2.2, a avaliação da interface e da ferramenta revelou que a usabilidade e a organização das informações foram aspectos determinantes para que os especialistas conseguissem absorver rapidamente os resultados e tomar decisões de prosseguimento na investigação. O Gráfico 6.3 sintetiza a percepção dos avaliadores em relação a critérios como clareza das visualizações, utilidade das conexões apresentadas e adequação da navegação no contexto investigativo.

Complementarmente, a Seção 6.3.2 evidencia que a percepção sobre a ferramenta foi amplamente positiva. Os especialistas destacaram a agilidade na compreensão do contexto dos grupos, a facilidade de identificação de padrões relevantes e a redução do tempo necessário para percorrer múltiplas fontes de dados. A integração entre o motor heurístico e a camada de visualização foi apontada como diferencial para a efetividade da ferramenta no dia a dia de trabalho.

Síntese do Artefato 2. O Artefato 2 consolidou os avanços obtidos no *pipeline* exploratório do Artefato 1, transformando-os em uma solução interativa e operacional para suporte à investigação patrimonial. Ao aliar desempenho heurístico estável com uma interface orientada ao fluxo cognitivo do especialista, o artefato final entrega valor direto ao processo investigativo, permitindo que a análise se concentre nas conexões mais promissoras e reduzindo o tempo de reação frente a indícios de grupos econômicos *de fato*.

7.2 Conclusões

O trabalho apresentou uma abordagem inovadora para a localização de grupos econômicos *de fato*, integrando processamento paralelo em *Apache Spark*, análise de grafos com *GraphX* e *GraphFrames*, e modelos baseados em Bidirectional Encoder Representations from Transformers (BERT) para similaridade de sentenças. O desenvolvimento foi conduzido pelo método *Design Science Research*, assegurando que a solução resultasse de um ciclo iterativo de concepção, construção, avaliação e refinamento, sempre alinhada às necessidades e restrições do domínio investigado.

A solução final se materializou em dois artefatos complementares: um *pipeline* exploratório (Artefato 1), capaz de ampliar a cobertura de entidades potencialmente vinculadas a grupos econômicos *de fato*, e uma interface investigatória (Artefato 2), que operaciona-

liza a exploração desses vínculos em um ambiente interativo voltado ao especialista. A combinação de *features* estruturadas derivadas do CNPJ e de indícios semânticos obtidos por embeddings resultou em uma abordagem híbrida que se mostrou robusta frente aos cenários analisados.

Os resultados demonstraram, com suporte de validação estatística e avaliação qualitativa por especialistas, que a proposta não apenas aumenta a abrangência da detecção de vínculos, como também mantém a pertinência dos achados, evitando sobrecarga de informações irrelevantes. Essa combinação contribui para reduzir o esforço manual, agilizar a tomada de decisão e ampliar a capacidade de identificação de conexões não formalmente documentadas, endereçando diretamente um dos principais gargalos da investigação patrimonial.

Embora concebida para integração com o SNIPER e o contexto de investigações conduzidas no âmbito do CNJ, a arquitetura e os procedimentos desenvolvidos possuem potencial de adaptação a outros ambientes que demandem análise de redes complexas de entidades, sejam elas econômicas, sociais ou criminais. Dessa forma, o trabalho oferece não apenas um produto técnico aplicável, mas também um modelo conceitual replicável, capaz de orientar o desenvolvimento de soluções semelhantes em diferentes domínios investigativos.

7.3 Contribuições

As principais contribuições deste trabalho podem ser sintetizadas em quatro dimensões: técnica, metodológica, prática e científica.

- **Técnica:** Implementação de um *pipeline* exploratório (Artefato 1) e de uma interface investigatória (Artefato 2) capazes de integrar múltiplas fontes de dados estruturados, combinar *features* do CNPJ com embeddings semânticos, e produzir visualizações interativas de grafos. Essa abordagem híbrida possibilitou revelar vínculos econômicos não explicitamente documentados, com validação estatística (Subseção 5.3.4) e avaliação positiva por especialistas (Subseção 6.2.2).
- **Metodológica:** Aplicação estruturada do *Design Science Research* ao problema de localização de grupos econômicos *de fato*, contemplando desde a elicitação de requisitos junto a especialistas até a formulação e teste de heurísticas específicas para o domínio. O processo metodológico incluiu ciclos iterativos de desenvolvimento e avaliação, permitindo alinhar continuamente os artefatos às necessidades práticas do contexto investigativo, conforme detalhado no Capítulo 3.

- **Prática:** Disponibilização de uma solução aplicável diretamente ao fluxo de trabalho do Conselho Nacional de Justiça (CNJ) e do SNIPER, capaz de ampliar a abrangência das investigações patrimoniais e reduzir o tempo necessário para identificar conexões relevantes. O uso da ferramenta mostrou potencial de acelerar a fase de execução dos processos judiciais e otimizar a alocação de recursos na recuperação de ativos, conforme evidenciado na Seção 6.2.
- **Científica:** Contribuição para o avanço do estado da arte na análise de redes econômicas complexas, ao combinar técnicas de análise de grafos, processamento paralelo e modelos de similaridade semântica em um contexto investigativo real. Os resultados e procedimentos documentados oferecem base para replicação e adaptação em outras áreas de pesquisa envolvendo detecção de relações não explícitas (Seção 7.1).

7.4 Trabalhos Futuros

No primeiro estágio da heurística exploratória (Subseção 5.2.1), a formação inicial de *clusters* é realizada por meio de filtros aplicados a atributos específicos, tais como *CEP*, *e-mail principal*, telefone e CPF do responsável. Essa estratégia permite reduzir o universo inicial de empresas, gerando subgrafos mais focados e potencialmente mais relevantes para a investigação.

Apesar de sua utilidade, a dependência de um ou dois critérios de filtragem apresenta limitações. Se o filtro for excessivamente restritivo, há risco de ocorrências de **falsos negativos**, em que entidades com vínculos relevantes não são incluídas no *cluster*. Por outro lado, filtros demasiadamente amplos aumentam a ocorrência de **falsos positivos**, agregando entidades sem relação efetiva e prejudicando a precisão das análises subsequentes. Essa tensão entre abrangência e precisão é um desafio central na etapa inicial de segmentação.

Uma alternativa promissora para superar essas limitações é a adoção de técnicas de *aprendizado de links* (*link prediction*) em grafos, especialmente quando combinadas com *Graph Neural Networks* (GNNs) [37, 38]. O aprendizado de links busca prever a existência de arestas que não estão explicitamente presentes no grafo, mas que são prováveis com base em padrões estruturais e atributos dos nós. Isso significa identificar conexões latentes entre entidades, muitas vezes invisíveis para abordagens exclusivamente determinísticas.

No contexto da heurística exploratória, o uso de aprendizado de links permitiria a formação de *clusters* iniciais mais robustos, integrando:

- **Padrões topológicos:** aproveitando a estrutura do grafo para detectar relações indiretas ou múltiplas vias de conexão entre entidades.

- **Similaridade de atributos:** combinando características explícitas (endereços, CNAEs, contatos) com representações latentes aprendidas por embeddings.
- **Sinais fracos:** capturando indícios sutis de vínculo, que isoladamente não seriam suficientes para estabelecer uma conexão, mas que, combinados, sugerem forte probabilidade de relação.

Modelos baseados em GNNs, como o *Graph Autoencoder* (GAE) [39] e o *GraphSAGE* [40], aprendem representações vetoriais (*embeddings*) para cada nó, propagando informações ao longo de suas conexões e permitindo que o algoritmo identifique relações prováveis mesmo em cenários de dados incompletos. Essa capacidade é especialmente relevante para investigações patrimoniais, nas quais parte das conexões pode estar ausente ou fragmentada nos registros públicos.

Como evolução deste trabalho, propõe-se investigar a aplicação de aprendizado de links com GNNs para aprimorar o processo de formação de *clusters* no primeiro estágio da heurística exploratória. A pesquisa deverá incluir:

1. Definição de métricas de avaliação que equilibrem *recall* e *precision* no contexto investigativo.
2. Comparação da abordagem atual baseada em filtros determinísticos com modelos de aprendizado de links.
3. Análise do impacto sobre o desempenho das etapas subsequentes da heurística.

Espera-se que essa abordagem permita reduzir perdas de conexões relevantes e, ao mesmo tempo, mitigar a inclusão de vínculos espúrios, tornando a etapa de triagem mais eficiente e efetiva para a detecção de grupos econômicos *de fato*.

Referências

[1] Dresch, Aline, Daniel Pacheco Lacerda e José Antonio Valle Antunes Junior: *Design science research: método de pesquisa para avanço da ciência e tecnologia*. Bookman Editora, 2020. x, 9, 10

[2] Shearer, Colin: *The crisp-dm model: the new blueprint for data mining*. Journal of data warehousing, 5(4):13–22, 2000. x, 13, 24

[3] Brasil: *Código Tributário Nacional*. Senado Federal, 1966. 2

[4] Brasil: *Código Civil Brasileiro*. Senado Federal, 2002. 2

[5] Coelho, Fábio Ulhoa: *Curso de Direito Comercial: Direito de Empresa*. Saraiva, São Paulo, 19ª edição, 2020. 2, 3

[6] Requião, Rubens: *Curso de Direito Comercial*. Saraiva, São Paulo, 27ª edição, 2005. 2, 3

[7] Carvalhosa, Modesto: *Comentários à Lei de Sociedades Anônimas*. Saraiva, São Paulo, 5ª edição, 2017. 2

[8] Delgado, Maurício Godinho: *Curso de Direito do Trabalho*. LTr, São Paulo, 19ª edição, 2021. 2

[9] Lopes, Christian Sahb Batista e Rafhael Frattari: *Solidariedade tributária e grupos econômicos*. Revista de Direito Tributário e Financeiro, 1(1):581–603, 2015. 4, 7, 13, 31

[10] Santos, Sara Gabriela do Nascimento, Vivian Cristine dos Santos Elias e Luiz Reinaldo Capelleti: *Desconsideração da personalidade jurídica na execução fiscal*. Revista Ibero-Americana de Humanidades, Ciências e Educação, 10(5):5422–5431, maio 2024. `https://periodicorease.pro.br/rease/article/view/14246`. 7, 8

[11] Carvalho, Álvaro Augusto Bastos de e Raimir Holanda Filho: *Detecting fraud in public acquisition of brazilian government with an analytical approach*. Procedia Computer Science, 238:248–254, 2024. 7

[12] Healy, P. e K. Palepu: *The fall of enron*. Journal of Economic Perspectives, 17(2):3–26, 2003. 8

[13] Köhler, K.: *Wirecard: A case of auditing failure*. Journal of Business Ethics, 2021. 8

[14] Pandit, Sumeet, Duen Horng Chau, Shobha Wang e Christos Faloutsos: *Netprobe: a fast and scalable system for fraud detection in online auction networks.* Em *Proceedings of the 16th International Conference on World Wide Web*, páginas 201–210. ACM, 2007. 8

[15] Akoglu, Leman, Michael McGlohon e Christos Faloutsos: *Oddball: Spotting anomalies in weighted graphs.* Em *2010 IEEE International Conference on Data Mining (ICDM)*, páginas 410–419. IEEE, 2010. 8

[16] Brandão, Michele A, Arthur PG Reis, Bárbara MA Mendes, Clara A Bacha De Almeida, Gabriel P Oliveira, Henrique Hott, Larissa D Gomide, Lucas L Costa, Mariana O Silva, Anisio Lacerda *et al.*: *Plus: A semi-automated pipeline for fraud detection in public bids.* Digital Government: Research and Practice, 5(1):1–16, 2024. 8

[17] Malewicz, G., M. H. Austern, A. Bik, J. Dehnert, I. Horn, N. Leiser e G. Czajkowski: *Pregel: A system for large-scale graph processing.* Em *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, páginas 135–146, 2010. 8

[18] Gonzalez, J. E., Y. Low, H. Gu, D. Bickson e C. Guestrin: *Powergraph: Distributed graph-parallel computation on natural graphs.* Em *Proceedings of the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI '12)*, páginas 17–30, 2012. 8

[19] Kyrola, A., G. Blelloch e C. Guestrin: *Graphchi: Large-scale graph computation on just a pc.* Em *Proceedings of the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI '12)*, páginas 31–46, 2012. 8

[20] Roy, A., I. Mihailovic e S. Zdonik: *X-stream: Edge-centric graph processing using streaming partitions.* Proceedings of the VLDB Endowment, 6(10):969–980, 2013. 8

[21] Zaharia, Matei, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker e Ion Stoica: *Spark: Cluster computing with working sets.* Em *2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 10)*, 2010. 8, 22

[22] al., M. Zaharia et: *Graphx: A resilient distributed graph system on spark.* Em *First International Workshop on Graph Data Management Experiences and Systems*, 2012. 8

[23] Grover, Aditya e Jure Leskovec: *node2vec: Scalable feature learning for networks.* Em *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, páginas 855–864. ACM, 2016. 8

[24] Hamilton, William L., Rex Ying e Jure Leskovec: *Inductive representation learning on large graphs.* Em *Advances in Neural Information Processing Systems*, volume 30, páginas 1024–1034. Curran Associates, Inc., 2017. 8

[25] Veličković, Petar, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio e Yoshua Bengio: *Graph attention networks.* Em *International Conference on Learning Representations*, 2018. 8

[26] BRASIL. RFB: *Cadastro nacional da pessoa jurídica - cnpj*, 2023. https://dados.gov.br/dados/conjuntos-dados/cadastro-nacional-da-pessoa-juridica---cnpj, URL:https://dados.gov.br/dados/conjuntos-dados/cadastro-nacional-da-pessoa-juridica—cnpj Accessed: 2024-10-30. 15, 20

[27] Angles, Renzo e Claudio Gutierrez: *Survey of graph database models*. ACM Comput. Surv., 40(1), fevereiro 2008, ISSN 0360-0300. https://doi.org/10.1145/1322432.1322433. 18

[28] Inoubli, Wissem, Sabeur Aridhi, Haithem Mezni, Mondher Maddouri e Engelbert Mephu Nguifo: *An experimental survey on big data frameworks*, 2018. ISSN 0167-739X. 22

[29] Dean, Jeffrey e Sanjay Ghemawat: *Mapreduce: Simplified data processing on large clusters*. Commun. ACM, 51(1):107–113, janeiro 2008, ISSN 0001-0782. https://doi-org.ez54.periodicos.capes.gov.br/10.1145/1327452.1327492. 23

[30] Reimers, Nils e Iryna Gurevych: *Sentence-bert: Sentence embeddings using siamese bert-networks*. Em *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, novembro 2019. http://arxiv.org/abs/1908.10084. 30

[31] Neyman, J.: *On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection*. Journal of the Royal Statistical Society, 97(4):558–625, 1934. 36

[32] Montgomery, D. C.: *Applied Statistics and Probability for Engineers*. Wiley, 7ª edição, 2019. 38

[33] Reichheld, Frederick F: *The one number you need to grow*. Harvard business review, 81(12):46–55, 2003. 48, 50

[34] Boone Jr, Harry N e Deborah A Boone: *Analyzing likert data*. The Journal of extension, 50(2):48, 2012. 52

[35] Harpe, Spencer E: *How to analyze likert and other rating scale data*. Currents in pharmacy teaching and learning, 7(6):836–850, 2015. 52

[36] Allen, I Elaine e Christopher A Seaman: *Likert scales and data analyses*. Quality progress, 40(7):64–65, 2007. 52

[37] Wu, Zonghan, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang e Philip S. Yu: *A comprehensive survey on graph neural networks*. IEEE Transactions on Neural Networks and Learning Systems, 32(1):4–24, 2020. 56

[38] Zhang, Muhan e Yixin Chen: *A survey on link prediction in complex networks*. ACM Computing Surveys (CSUR), 52(5):1–35, 2020. 56

[39] Kipf, Thomas N. e Max Welling: *Variational graph auto-encoders*. Em *Proceedings of the NIPS Workshop on Bayesian Deep Learning*, 2016. https://arxiv.org/abs/1611.07308. 57

[40] Hamilton, Will, Rex Ying e Jure Leskovec: *Inductive representation learning on large graphs*. Advances in Neural Information Processing Systems, 30:1024–1034, 2017. https://arxiv.org/abs/1706.02216. 57

Anexo I

Formulário de Avaliação da Interface Investigatória

Questionário de Avaliação da Ferramenta Investiga CNPJ - Basset

Para cada uma das afirmações a seguir, por favor, marque o grau em que você concorda ou discorda. Utilize a escala de 1 a 5, onde:

- **1** = Discordo totalmente
- **2** = Discordo parcialmente
- **3** = Neutro
- **4** = Concordo parcialmente
- **5** = Concordo totalmente

* Indica uma pergunta obrigatória

Precisão e Eficácia

1. 1. A **Ferramenta Investiga CNPJ** aumenta a precisão na identificação de Grupos Econômicos *

Marcar apenas uma oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo Totalmente

2. 2. A **Ferramenta Investiga CNPJ** reduz o esforço necessário para a localização de empresas e grupos econômicos em comparação com métodos tradicionais. *

Marcar apenas uma oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo Totalmente

3. 3. A ferramenta conduz a localização de conexões patrimoniais ocultas que não seriam facilmente detectadas manualmente. *

Marcar apenas uma oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo Totalmente

Usabilidade e Eficiência

4. 4. A interface da **Ferramenta Investiga CNPJ** é intuitiva e fácil de usar. *

Marcar apenas uma oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo Totalmente

5. 5. O tempo necessário para obter um resultado na ferramenta é adequado às necessidades da investigação. *

Marcar apenas uma oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo Totalmente

6. 6. A disposição visual dos resultados facilita a identificação de empresas associadas a um Grupo Econômico *

Marcar apenas uma oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo Totalmente

Impacto na Investigação Patrimonial

7. 7. O uso da **Ferramenta Investiga CNPJ** contribui de alguma forma para aumentar a celeridade da fase de execução processual. *

Marcar apenas uma oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo Totalmente

8. 8. A ferramenta amplia o espectro de análise patrimonial ao sugerir conexões não explícitas nos registros oficiais. *

Marcar apenas uma oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo Totalmente

9. 9. A **Ferramenta Investiga CNPJ** contribui para a ampliação do espectro de identificação de Grupos Econômicos de Fato, fortalecendo a fundamentação para desconsideração da personalidade jurídica. *

Marcar apenas uma oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo Totalmente

Comparação com Outras Ferramentas

10. 10. Em comparação com outras ferramentas investigativas disponíveis, a **Ferramenta Investiga CNPJ** oferece informações mais relevantes. *

Marcar apenas uma oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo Totalmente

11. 11. A **Ferramenta Investiga CNPJ** melhora a confiabilidade dos dados utilizados na investigação em relação a outras abordagens disponíveis. *

Marcar apenas uma oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo Totalmente

12. 12. O uso da ferramenta reduz a necessidade de verificações manuais exaustivas para confirmar relações entre empresas e sócios. *

Marcar apenas uma oval.

	1	2	3	4	5	
Disc	◯	◯	◯	◯	◯	Concordo Totalmente

Avaliação Geral e Sugestões

Esta seção busca capturar a percepção geral e sugestões para aprimorar a Ferramenta Investiga CNPJ. O feedback coletado será essencial para otimizar sua usabilidade, precisão e melhoria do produto

13. 13. Quais são os principais pontos positivos da ferramenta?

14. 14. O que poderia ser melhorado na **Ferramenta Investiga CNPJ**?

15. 15. Em uma escala de 1 a 10, o quanto você recomendaria a **Ferramenta Investiga CNPJ** para outros especialistas em investigações patrimoniais? Quando 1 indica "não recomendaria" e 10 "recomendaria totalmente" *

Marcar apenas uma oval.

1	2	3	4	5	6	7	8	9	10
◯	◯	◯	◯	◯	◯	◯	◯	◯	◯