# Relatório Técnico Parcial 4 do Projeto KnEDLe / NIDO

Alice Borges Carolina A. Okimoto Fabrício A. Braz
Luís P. F. Garcia Nilton C. Silva Thiago P. Faleiros
Marcelo Mandelli Vinícius R. P. Borges

**Universidade de Brasília**
**Departamento de Ciência da Computação**
http://nido.unb.br

29 de março de 2022

## 1 Introdução

O presente relatório tem como objetivo elencar os resultados produzidos no Projeto de Pesquisa *KnEDLe - Extração de Informações de Publicações Oficiais usando Inteligência Artificial*. O projeto é fruto de uma parceria entre a Universidade de Brasília (UnB), a Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF) e a Fundação de Empreendimentos Científicos e Tecnológicos (FINATEC)[1]. Este relatório trata das atividades e resultados produzidos na quarta fase do projeto (*release* 4) no período de 01/07/2021 a 31/12/2021.

O relatório está estruturado da seguinte forma: a Seção 2 apresenta as atividades do time de Viabilidade Técnica e as ferramentas utilizadas para gerência das atividades do projeto; as Seções 3 e 4 discutem as evoluções alcançadas nas ferramentas DOFMiner e DODFKge com inclusão dos modelos baseados em algoritmos inteligentes e construção do Banco de Dados (BD) baseado em grafo, respectivamente; a Seção 5 descreve o processo de anotação e validação para a criação da base de dados de atos de pessoal; a Seção 6 apresenta as atividades realizadas com atos de licitação e pré-contrato; a Seção 7 discute as atividades relacionadas a construção das expressões regulares para atos de contrato; a Seção 8 apresenta os resultados de pesquisa como artigos publicados e defesas de mestrado realizadas. A Seção 9 apresenta algumas considerações finais em relação à release em questão. Por fim, a Seção 10 encerra o relatório elencando os membros do projeto duante a *release* 4.

[1]Este projeto possui estes registros nas respectivas instituições envolvidas: FAPDF convênio 07/2019; UnB SEI:23106.058975/2019-62; Finatec 6429 - FAPDF/CIC.

# 2 Viabilidade Técnica

A *release* 4 teve como foco principal a evolução dos produtos e ferramentas em desenvolvimento durante as etapas anteriores do projeto. O objetivo era mitigar problemas anteriormente detectados nas ferramentas, avançar com soluções mais robustas e, quando necessário, propor novas soluções. Também é importante mencionar que a equipe passou por reestruturação ao final da *release* 3 com a saída e entrada de novos membros e criação de novos times. Além disso, foi realizada a compra de equipamento para o desenvolvimento científico e tecnológico do projeto. Essas atividades demandaram do time de Viabilidade Técnica um melhor acompanhamento das atividades das equipes por meio da adoção de novas tecnologias. Para isso foram utilizadas ferramentas como o ZenHub, *plugin* do GitHub, e a OKR que serão discutidas em maiores detalhes nas próximas seções.

## 2.1 ZenHub

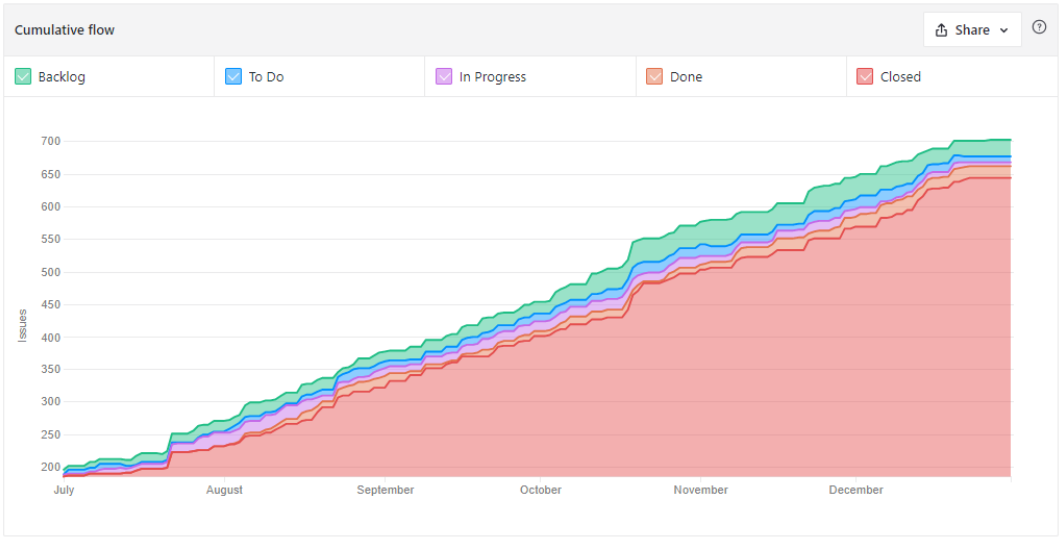
Uma das ferramentas escolhidas para acompanhar as tarefas dos times, foi o ZenHub[2]. O ZenHub é uma ferramenta de gestão de projetos poderosa, que permite a integração de utilização das *issues* do GitHub[3] como histórias ou *tasks*. Atualmente o ZenHub é utilizado por grandes companhias como a Adobe e a Microsoft. A grande vantagem de utilizar essa ferramenta é permitir a centralização das atividades e possibilitar unir a gestão, coleta de métricas e rastreio das *issues* nos *commits*. Ele também permite criar *sprints*, utilizar rótulos para organizar as atividades e *releases* para manter um acompanhamento melhor do projeto. A adoção do ZenHub visa centralizar as informações, reduzir o número de ferramentas utilizadas durante o processo de desenvolvimento e execução das atividades, a centralização das informações, além de possibilitar um rastreio das atividades da *issue*.

Coube à equipe de Viabilidade Técnica, em um primeiro momento, fornecer treinamento da ferramenta ZenHub para aqueles que não tinham conhecimento e, em um segundo momento, acompanhar semanalmente como os times estavam utilizando-a. Também coube à equipe verificar e acompanhar o avanço das atividades de cada time descritas na ferramenta.

A Figura 1 mostra o gráfico do fluxo de tarefas registrado no ZenHub durante a *release* 4 para todos os times. O eixo-$x$ apresenta os meses referentes a *release* 4 e o eixo-$y$ o número de *issues*. Durante essa *release* foram criados aproximadamente 700 *issues*. Para uma melhor organização do projeto, as *issues* foram classificadas em 5 tipos: (1) *Backlog* para tarefas que devem ser executadas eventualmente; (2) *To Do* para tarefas referentes à aquela *sprint* específica; (3) *In Progress* para tarefas que estão em andamento; (4) *Done* para tarefas finalizadas e; (5) *Closed* para tarefas que foram concluídas e comentadas nas reuniões de *sprint*. Dessa forma, todas as *issues* com as atividades devem seguir esse fluxo. Ao final da *release*, o número total de tarefas finalizadas deve se aproximar da quantidade de tarefas abertas.

A Figura 1 mostra que durante a *release* 4 o número de tarefas finzalizadas foi bem próximo do número de atividades propostas. Além disso, durante o projeto foi possível perceber que a quantidade de tarefas novas, em progresso e finalizadas se manteve em constante

[2] `https://www.zenhub.com/`
[3] `https://github.com/`




| Issues in pipeline | December 28th, 2021 | December 29th, 2021 | December 30th, 2021 | December 31st, 2021 |
|---|---|---|---|---|
| Backlog | 25 | 25 | 25 | 25 |
| To Do | 9 | 9 | 9 | 9 |
| In Progress | 7 | 7 | 7 | 7 |
| Done | 18 | 18 | 18 | 18 |
| Closed | 643 | 643 | 643 | 643 |

Figura 1: Gráfico do fluxo de tarefas.

crescimento, dessa forma não houve grandes diferenças entre número de atividades propostas e concluídas. A equipe de Viabilidade Técnica realizou o acompanhamento das tarefas semanalmente para garantir que não permanecessem muitas tarefas abertas e dar apoio ao time para o gerenciamento das *issues*.

## 2.2 Objective and Key Results (OKR)

O acompanhamento dos riscos foram realizados mensalmente por meio do *Objective and Key Results* (OKR)[14]. Coube ao time de Viabilidade Técnica conectar os objetivos do OKR com os da Estrutura Analítica do Projeto (EAP), os objetivos chave da *release* 4 e as atividades registradas no ZenHub.

No início da *release* 4, os objetivos chaves foram definidos para cada time e adicionados à planilha de acompanhamento. O acompanhamento do OKR foi realizado a cada 15 dias, onde os times atualizavam o andamento de seus objetivos na planilha criada para garantir a evolução do projeto e identificação de lacunas a serem preenchidas. O OKR dessa *release* ficou extenso, o que não foi uma boa estratégia pois dificultou o acompanhamento pela equipe de viabilidade técnica, trazendo assim esse ponto de melhoria para a próxima *release*.

A Figura 2 revela a evolução do OKR da equipe pelas *sprints* da *release* 4. O eixo-$x$ destaca as *sprints* enquanto o eixo-$y$ destaca a porcentagem de conclusão dos objetivos da *release*. Os objetivos do time de Viabilidade Técnica, a evolução dos produtos gerados e as atividades de pesquisa são ilustrados pelas linhas verde, azul e amarela, respectivamente.

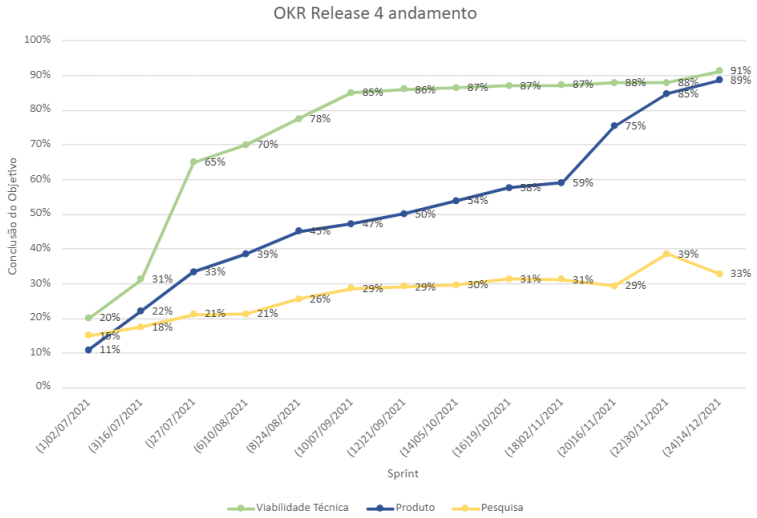



Figura 2: Evolução das atividades do OKR por *sprints*

Ao analisar a Figura 2, pode-se notar que o foco da *release* 4 foi melhorar os produtos já desenvolvidos no projeto por meio da solução de problemas e desenvolvimento de novos modelos inteligentes. Em torno de 89% dos objetivos de produtos foram alcançados, o que mostra que a *release* foi muito produtiva. No entanto, a pesquisa foi um dos objetivos que menos avançou. Apenas 33% da meta foi concluída, o que se deve ao fato de que grande parte dos alunos de Mestrado concluíram o curso ou então se afastaram do projeto. Já os objetivos de viabilidade técnica foram 91% concluídos, não chegando aos 100% pois a equipe

teve dificuldades em acompanhar o OKR por ser muito extenso e um descuido por parte dos membros do projeto em não preenchê-lo nas devidas datas.

# 3 DODFMiner

DODFMiner é o software em desenvolvimento para extração de informações de documentos em formato PDF referentes às publicações do Diário Oficial do Distrito Federal (DODF). Ele foi, primeiramente, desenvolvido como uma biblioteca em linguagem Python para apoiar outras aplicações. Atualmente, pode ser utilizado tanto como biblioteca quanto *command-line interface* (CLI). A biblioteca tem funções que apoiam a extração de informações de atos de pessoal por meio de algoritmos inteligentes.

O DODFMiner implementa três principais funções: (1) permitir o download automático dos DODFs; (2) extrair atos e; (3) extrair entidades das publicações de forma estruturada. Enquanto as tarefas de download e extração de texto dos PDFs depende de módulos baseados em bibliotecas bastante conhecias como *Beautiful Soup*[4] e PyMuPDF[5], as tarefas de extração de atos e entidades dependente de módulos baseados em algoritmos inteligentes.

Os módulos implementados no DODFMiner são apresentados na Figura 3. *Downloader* é o primeiro grande módulo da biblioteca, responsável por executar uma função de web scraper e executar o *download* dos DODFs em um intervalo de data previamente informado para o usuário. O segundo maior módulo do código é o *Extract*, responsável por extrair as informações contidas dentro dos PDFs dos DODFs. Dentro de tal módulo existem dois menores: o *Pure* é responsável por extrair a informação textual dos PDFs e o *Polished*, que contém todas as funcionalidades para extração dos atos e entidades. Atualmente, esse módulo se baseia em modelos de algoritmos de Aprendizado de Máquina (AM) como os de Reconhecimento de Entidades Nomeadas (*Named Entity Recognition* - NER) capazes de extrair as informações textuais com maior desempenho do que estratégias analíticas como expressões regulares.

A biblioteca do DODFMiner está, atualmente, na versão 1.3.11. Pode ser instalada via pip[6] ou via repositório do GitHub[7]. A documentação esta disponível no readthedocs[8]. As principais modificações realizadas desde a última *release* foram:

- Correção de *bugs* em módulos como *Downloader*, *Extract* e *Pure*. Ajuste das expressões regulares, extração de títulos e blocos de texto.
- Atualização das dependências para versões mais recentes, inclusão de *linter* para uma padronização de código mais robusta, implementação de testes unitários para melhorar a cobertura, expansão da documentação e a utilização do Jenkins[9] para *build*.
- Inclusão de modelos de AM para segmentação e NER de atos de pessoal com o objetivo de substituir a estratégia de expressões regulares.

[4] `https://www.crummy.com/software/BeautifulSoup/`
[5] `https://github.com/pymupdf/PyMuPDF`
[6] `https://pypi.org/project/dodfminer/`
[7] `https://github.com/UnB-KnEDLe/DODFMiner`
[8] `https://dodfminer.readthedocs.io/en/main/`
[9] `https://www.jenkins.io/`

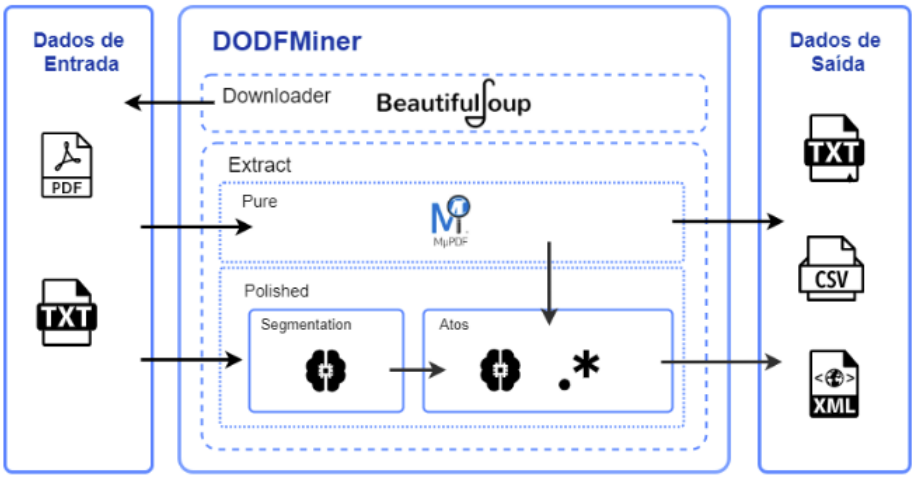



Figura 3: Ilustração da arquitetura dos componentes da ferramenta DODFMiner.

Com o objetivo de contextualizar o uso dos algoritmos de AM mais eficientes do que as estratégias analíticas, a Seção 3.1 irá discutir a abordagem utilizada para segmentação de atos enquanto a Seção 3.2 irá discutir a abordagem utilizada para NER, ambas utilizando algoritmos clássicos de AM.

## 3.1 Segmentação de atos

A segmentação de atos é a tarefa de identificar atos no texto de um DODF. Como cada ato tem limites bem definidos e pode ser considerado uma única entidade, esse problema é semelhante ao problema de NER. Portanto, a implementação, indução, validação e teste dos modelos pode ser considerada de forma similar.

Para essa tarefa, foi considerado, como um primeiro esforço, uma classe de algoritmos da área de NER, os *Conditional Random Fields* (CRFs) devido à sua simplicidade e eficácia para diversas tarefas correlatas. CRFs são métodos de modelagem estatística usados para previsão estruturada [6]. Enquanto um classificador prevê um rótulo para uma única amostra sem considerar amostras vizinhas, o CRF pode levar em consideração o contexto. Para isso, as previsões são modeladas como um modelo gráfico, que representa a presença de dependências entre as previsões.

Os modelos CRF funcionam atribuindo um rótulo a uma instância em uma sequência baseada nos atributos e os rótulos de suas instâncias vizinhas. Ele atribui um peso a cada combinação de atributo e rótulo, bem como a cada combinação de dois rótulos, e usa esses pesos para determinar a classe mais provável de uma instância. Durante o treinamento, a otimização pode ser feita por algoritmos como *Broyden–Fletcher–Goldfarb–Shanno algorithm* (L-BFGS) [15] e *Stochastic Gradient Descent* (SGD) [17]. Para esses experimentos, a biblioteca *sklearn-crfsuite*[10] foi utilizada.

Para gerar um conjunto de dados de treinamento, foram utilizados os atos da base de dados *dodf_atos_pessoal_v3*, juntamente com os DODFs dos quais foram extraídos. A metodologia inclui a utlização de técnicas de *5-fold cross validation* com *tuning* de parâmetros. A métrica

[10] `https://github.com/TeamHG-Memex/sklearn-crfsuite`

utilizada para avaliar o desempenho foi a *F1-score*. Para essa medida, valores próximos de 1 são desejados.

O resultado da segmentação utilizando CRF pode ser visto na Tabela 1. Para cada tipo de ato, é apresentado o desempenho da medida de avaliação *F1-score*. Como alguns atos nos DODFs usados para treinamento não estão presentes no conjunto de dados, os resultados podem ser considerados parciais ou em evolução.

| **Tipo de ato** | **F1-score** |
|---|---|
| Abono | 0.905 |
| Aposentadoria | 0.995 |
| Cessões | 0.834 |
| Exoneração Comissionado | 0.964 |
| Exoneração Efetivo | 0.876 |
| Nomeação Comissionado | 0.949 |
| Nomeação Efetivo | 0.345 |
| Retificação Comissionado | 0.756 |
| Retificação Efetivo | 0.802 |
| Reversões | 0.600 |
| Sem Efeito Aposentadoria | 0.371 |
| Sem Efeito Exo./Nom. | 0.956 |
| Substituição | 0.948 |

Tabela 1: F1-score da segmentação por tipo de ato.

De forma geral, podemos perceber que para aproximadamente metade dos atos, o desempenho alcançado pelo CRF foi superior a 0.9. Atos como *Cessões*, *Exoneração Efetivo*, *Retificação Comissionado* e *Retificação Efetivo* tiveram desempenho mediano. Os atos com menor desempenho foram *Nomeação Efetivo*, *Reversões* e *Sem Efeito Aposentadoria*. De forma geral, o bom ou mau desempenho do CRF pode ser justificado pela complexidade em caracterizar determinados atos e pelo baixo volume de dados rotulados em alguns casos.

Para adicionar esses modelos gerados à biblioteca DODFMiner, uma nova classe, *ActSeg*, foi criada. Esta classe inclui o código principal relacionado à segmentação de atos para técnicas de expressão regular e modelos CRF. Em trabalhos futuros, esperamos abordar técnicas estado-da-arte em NER, principalmente aquelas que fazem uso de estratégias de Aprendizado Profundo [5], mecanismos de Atenção [16] ou transferência de aprendizado [18].

## 3.2 Extração de entidades

A indução dos modelos para extração de entidades utilizou o conjunto de dados *dodf_atos_pessoal_v3*. Essa tarefa envolve a extração de informações bem definidas de atos em DODFs, como nomes, cargos e datas. A classe de modelos utilizada também foi o CRF, devido à sua simplicidade e eficácia para NER.

A maior parte do código inicial foi feita com base no trabalho anterior que já havia adicionado modelos para três tipos de atos à biblioteca DODFMiner. O código de pré-processamento e os recursos usados foram aprimorados ao longo do tempo para se adaptar

ao conjunto de dados de treinamento.

O resultado da extração de entidades utilizando CRF pode ser visto na Tabela 2. Para cada tipo de ato, é apresentado o desempenho da medida de avaliação *F1-score*. Como alguns atos nos DODFs usados para treinamento não estão presentes no conjunto de dados, os resultados podem ser considerados parciais ou em evolução.

| **Tipo de ato** | **F1-score** |
|---|---|
| Abono | 0.901 |
| Aposentadoria | 0.997 |
| Cessões | 0.911 |
| Exoneração Comissionado | 0.973 |
| Exoneração Efetivo | 0.948 |
| Nomeação Comissionado | 0.982 |
| Nomeação Efetivo | 0.796 |
| Retificação Comissionado | 0.875 |
| Retificação Efetivo | 0.912 |
| Reversões | 0.923 |
| Sem Efeito Aposentadoria | 0.971 |
| Sem Efeito Exo./Nom. | 0.979 |
| Substituição | 0.944 |

Tabela 2: F1-score do NER por tipo de ato.

De forma geral, podemos perceber que para grande parte dos atos, o desempenho alcançado pleo CRF foi superior a 0.9, exceto para *Retificação Comissionado* e *Nomeação Efetivo* que apresentam uma pequena quantidade de amostras. Isso indica que, mesmo o CRF sendo uma técnica simples, apresenta bom desempenho. Como a biblioteca DODFMiner já suportava a adição de modelos NER para extração de entidades, poucas alterações foram necessárias para incluir novos modelos.

Para as próximas *releases*, a ideia é estudar o comportamento de técnicas estado-da-arte como já mencionado na seção anterior focando, principalmente, naquelas que utilizam Aprendizado Profundo [5], mecanismos de Atenção [16] ou transferência de aprendizado [18].

# 4 DODFKge

O DODF*Knowledge* ou DODFKge é um software em desenvolvimento que tem como objetivo consolidar e gerenciar as informações oriundas da extração de informação dos DODFs via DODFMiner. Ele foi desenvolvido como: (1) um Banco de Dados (BDs) do tipo NoSQL e orientado à grafos e; (2) uma API em Python para pesquisa. Ele permite que as informações sejam recuperadas via consultas semânticas ou pesquisa simples.

Os BDs baseados em grafos são bancos NoSQL que permitem estruturar os dados em entidades e relacionamentos e permitem consultas semânticas avançadas. O grafo gerado por esses BDs relaciona os itens de dados a uma coleção de nós e arestas. Enquanto os nós representam entidades, as arestas são relacionamentos existentes entre os nós. As relações permitem que os nós sejam vinculados diretamente e recuperados com uma única operação.

Nesse sentido, o Neo4j[11] surge como uma possível solução para persistir dados uma vez que é um BD orientado à grafos e tem sido amplamente empregado na indústria e academia. O Neo4j é acessível a partir da maioria das linguagens de programação usando uma interface API REST. Além disso, ele está disponível em uma edição comunitária de código aberto licenciada pela GPLv3 e a linguagem Cypher utilizada para recuperar informações é intuitiva e simples de usar. Esses fatores contribuíram para a escolha do Neo4j.

Para que o BD pudesse ser populado com as informações dos DODFs, foram utilizados os modelos inteligentes de NER implementados no DODFMiner. Portanto, milhares de publicações do ano de 2010 à 2020 foram coletadas, processadas e extraída as informações. Para a geração desse banco, foi definido um esquema baseado nos tipos de atos, entidades e relações nos atos de pessoal. A Figura 4 mostra o esquema definido para o BD. Nele é possível ver as relações que existem entre as entidades *DODF*, *Documento*, *Edital_Normativo*, *Ordem_de_Serviço*, *Pessoa*, *Matrícula*, *Orgão* e *Cargo*. As relações podem ser diversas como *Publicado_no*, *Aprovado_no*, *Aposenta_de* e entre outras.

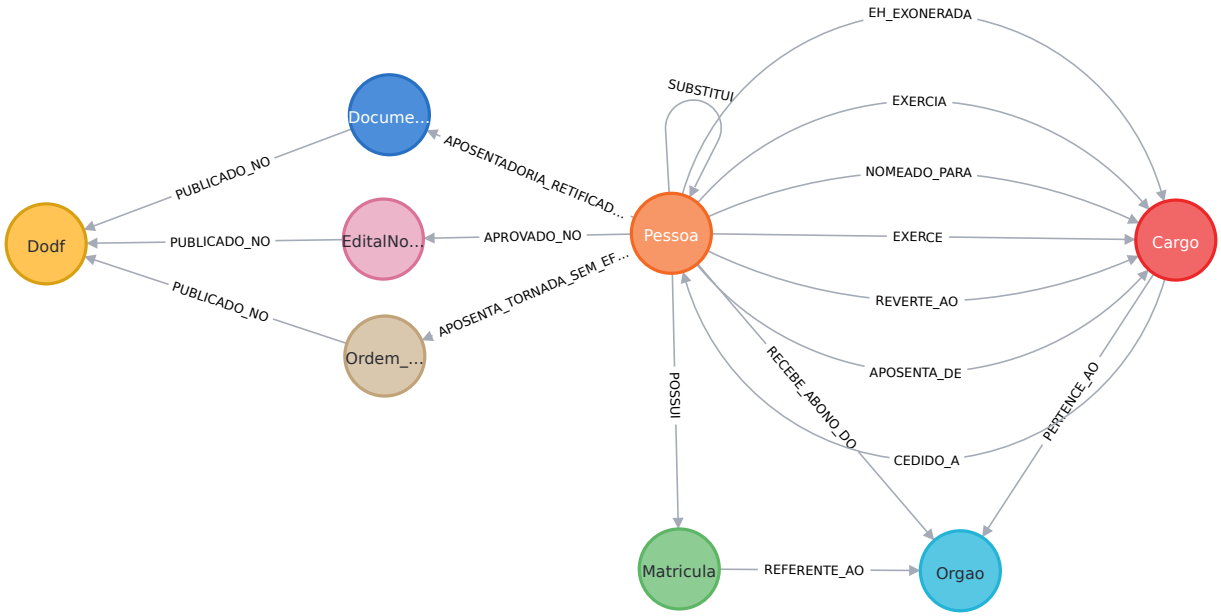



Figura 4: Esquema do BD baseado em grafos para atos de pessoal.

Dessa forma, consultas semânticas podem ser construídas para recuperar as informações relacionadas aos atos de pessoal. Seguem alguns exemplos de consultas: (1) recuperar todas as pessoas que aposentaram de um cargo do Orgão "Secretaria de Estado de Saúde"; (2) recuperar as pessoas nomeadas para um determinado cargo de chefia em algum orgão e; (3) recuperar todas as relações de exerce/exercia um determinado cargo. Seguem as respectivas consultas:

- `MATCH (n:Pessoa) - [:APOSENTA_DE] -> (c:Cargo) - [:PERTENCE_AO] -> (o:Orgao{nome:"SECRETARIA DE ESTADO DE SAUDE"}) RETURN n,c,o`

[11] `https://neo4j.com/`

- `MATCH (n:Pessoa) - [:NOMEADO_PARA] -> (c) - [:PERTENCE_AO] -> (o)`
  `WHERE c.nome STARTS WITH "Chefe"`
  `RETURN n,c,o`
- `MATCH (c:Cargo) <- [:EXERCE] - (n:Pessoa) - [:EXERCIA] -> (c:Cargo)`
  `RETURN c,n`

Para realizar essas e outras consultas, o usuário pode, com as credenciais corretas, acessar o link da ferramenta DODFKge[12]. Essa ferramenta é baseada na interface *Neo4j Browser* para consulta e visualização. Atualmente o BD está utilizando os modelos NER disponíveis na última versão estável do DODFMiner. Por conta da constante evolução dos modelos, esquema e consultas podem sofrer alterações conceituais contantes.

Além do acesso ao BD via *Neo4j Browser*, foi construída uma API expondo alguns *endpoints* para execução de consultas. Nesse caso, foram projetados filtros específicos por tipo ato. Esses filtros permitem combinar os campos de busca para consulta. O acesso a essa ferramenta é feito por meio do DASH[13] na aba "Pesquisa". A principal vantagem dessa ferramenta é a não necessidade de conhecimento técnico específico na linguagem de consulta. A principal limitação é que os filtros não têm capacidade de realizar consultas complexas.

# 5 Anotação de Textos

O processo de anotação visa a criação de *corpus* (conjunto de textos) para os atos de pessoal do Diário Oficial do Distrito Federal. O *corpus* é importante para o treinamento de modelos de processamento de linguagem natural e de aprendizado para reconhecimento de entidades nomeadas e segmentação de textos a partir das publicações associadas a atos como nomeação, exoneração, cessão entre outros. Vale ressaltar que o processo de anotação foi idealizado para obter um *corpus* padrão ouro (*Gold Standard*), significando que o processo de anotação seguiu um rigoroso processo de marcação de palavras que denotam as entidades nos documentos, além da revisão dessas marcações por diferentes anotadores. Ademais, as anotações (rótulos e segmentos de textos anotados) devem apresentar um alto nível de concordância entre os anotadores para ser considerado um padrão ouro.

A Tabela 3 sumariza como ficou dividida a quantidade total de 100 documentos alocadas para anotação e criação do DODF Corpus. O Batch KnEDLe foi parcialmente anotado pelos membros do projeto KnEDLe devido ao processo de anotação estar consumindo tempo necessário de pesquisa e desenvolvimento para cumprimento dos objetivos do projeto. Os Batches 1, 2, 3 e de Validação foram parcialmente anotados pelos voluntários externos ao projeto durante a *Release* 3, não sendo concluídos por causa de problemas técnicos na ferramenta NidoTat e com o servidor que a hospeda. A equipe de anotação se encarregou de concluir o processo de anotação dos documentos e revisão das anotações realizadas durante a *Release* 4. Informações detalhadas sobre como o processo de anotação foi conduzido podem ser encontrados nos relatórios das *Releases* 2 e 3.

De maneira geral, durante a *Release* 4, a equipe focou em corrigir os erros de anotação no *corpus*, além de disponibilizá-lo para todos os membros do projeto KnEDLe em formatos

[12] `http://164.41.76.30/browser/`
[13] `http://164.41.76.30/dash/`

Tabela 3: Descrição dos batches de DODFs concebidos para o processo de anotação.

| Nome do Batch | Quantidade de documentos |
|---|---|
| Batch KnEDLe | 33 |
| Batch 1 | 21 |
| Batch 2 | 20 |
| Batch 3 | 20 |
| Batch de Validação | 6 |

*csv* e *xml*. Inicialmente, a equipe de anotação do KnEDLe, composta pelos membros Lívia, Matheus e Tatiana, iniciou as atividades de análise e revisão de todas as anotações realizadas pelos voluntários externos ao projeto, como também as anotações ainda incompletas realizadas na *Release* 2 pelos membros do KnEDLe. Tais revisões e eventuais correções foram fundamentais para maximizar a quantidade de anotações, pois alguns voluntários abandonaram o processo de anotação, como também problemas técnicos afetaram a ferramenta de anotação NidoTat por curtos períodos de tempos. Outro fator importante é que os anotadores voluntários tiveram que ser treinados para ganharem entendimento da estrutura e do padrão dos atos, uma vez que eles não eram especialistas em publicações do Diário Oficial do Distrito Federal.

De acordo com a equipe de anotação, o processo de revisão das anotações de todos os documentos demandou muito esforço devido aos seguintes fatores:

- Atos e entidades não anotados: correspondem aos segmentos de textos que deveriam ter sido marcados e rotulados pelos anotadores (consequentemente pelos revisores) durante o processo de anotação, mas que, por algum motivo, não foram anotados. Essas situações estão relacionadas com o abandono de voluntários do processo de anotação que deixaram os documentos que lhe foram atribuídos sem anotações.

- Rótulos incorretos: significa que um segmento de texto foi incorretamente rotulado por um anotador e que continuou incorreto mesmo após o processo de revisão. A grande maioria dessas ocorrências estiveram relacionadas com entidades difíceis de serem reconhecidas, como por exemplo "órgão" e "hierarquia de lotação". Assim, esses erros foram ocasionados pelo fato dos anotadores voluntários não serem especialistas em DODF, não conhecendo todos os órgãos do Governo do Distrito Federal, como também seus sub-órgãos, sub-secretarias, entre outros.

- Padrões e estruturas heterogêneas dos atos: era comum os anotadores voluntários encontrarem textos associados aos atos apresentando estruturas e padrões diferentes daqueles especificados no tutorial de anotação - que foi escrito com base no documento de requisitos fornecido pelo Tribunal de Contas do Distrito Federal.

A Tabela 4 descreve a quantidade de instâncias de cada tipo de ato na versão 3 (v3) do *corpus* considerando seus 99 documentos. Vale ressaltar que a ausência de um documento no *corpus* se deveu a problema técnico no NidoTat, em que não foi possível recuperar as anotações realizadas. A equipe de anotação está analisando estratégias para resolver o problema. Observa-se um desbalanceamento entre esses atos, em que é possível verificar que

os atos de "Nomeação de cargo efetivo", "Reversão" e "Ato Tornado sem Efeito em Aposentadoria" apresentam poucas instâncias, o que pode afetar o aprendizado dos modelos de reconhecimento de entidades nomeadas e segmentação de textos.

Tabela 4: Total de instâncias por ato para a versão 3 (v3) do *corpus* contendo 99 documentos.

| Tipo de ato | Quantidade de instâncias |
|---|---|
| Substituição | 2423 |
| Exoneração de Cargo Comissionado | 2380 |
| Nomeação de Cargo Comissionado | 2314 |
| Retificação de Cargo Efetivo | 1241 |
| Exoneração de Cargo Efetivo | 303 |
| Cessão | 267 |
| Ato Tornado Sem Efeito (Exo. e Nom.) | 251 |
| Retificação Comissionado | 201 |
| Abono de Permanência | 137 |
| Nomeação de Cargo Efetivo | 79 |
| Reversão | 62 |
| Ato Tornado sem Efeito (Aposentadoria) | 21 |
| Total | 9679 |

No final da *Release* 4, a equipe de anotação passou a realizar correções em marcações que eram mais difíceis de serem identificadas em uma anotação manual e que estão relacionadas com a maneira que o NidoTat opera. É o caso das relações que são importantes para conectar diversas entidades a um determinado tipo de ato. Por exemplo, em um ato de retificação, deve-se conectar o texto do ato completo de retificação e as entidades, como nome do servidor, órgão, hierarquia de lotação, informação errada e informação corrigida). Entretanto, sabendo-se que existem dois tipos de atos de retificação ("Retificação de Cargo Comissionado" e "Retificação de Cargo Efetivo"), foram identificados em experimentos de reconhecimento de entidades nomeadas que o ato anotado estava relacionado com "Retificação de Cargo Comissionado", mas a relação criada identificava "Retificação de Cargo Efetivo". Esses erros estão sendo corrigidos pela equipe de anotação por meio de uma ferramenta desenvolvida pela própria equipe utilizando o aplicativo streamlit [14].

# 6 Licitações e Pré-contratos

Dando continuidade à frente de licitações e pré-contratos desde a *Release* 3, a equipe investiu seus esforços nas seguintes frentes: Evolução da interface com usuário da *Timeline*, Identificação de blocos de atos com *pytesseract*, Implementação de um Autômato Determinístico Finito (DFA) para melhorar a extração de atos, Extração de entidades, Supervisão fraca para maior generalização das classes de atos extraídas com regex, Pesquisa sobre a implementação de banco de dados nosql.

[14] `https://streamlit.io/`

## 6.1 Evolução da interface com usuário da Timeline

A primeira modificação foi feita no *header* da aplicação, deixando-o mais intuitivo para o usuário de acordo com a primeira heurística de Nielsen, onde esse ressalta sobre a importância da visibilidade do status do sistema, deve-se informar ao usuário sobre qual ambiente ele estava, em qual ele está e para quais outros ambientes ele poderá se dirigir a partir de sua localização [11] (Figura 5 e Figura 6).


Figura 5: Novo header desenvolvido para facilitar a experiência de usuário.


Figura 6: Novo header desenvolvido para facilitar a experiencia de usuario.

Outra mudança importante foi em relação ao *flip card*, novas entidades foram adicionadas ao tipo de ato referente a extrato de contrato. São elas: contrato, partes, objeto, valor,

lei orçamentaria, união orçamentária, programa de trabalho, natureza da despesa, nota da empresa, data de assinatura, signatários e vigência.

A equipe populou o banco de dados da Timeline com quase 90 mil atos referentes a licitações e extratos de contratos. Vale salientar que primeiro é processado o arquivo responsável por popular a tabela certame, essa contém o número do processo, e depois o arquivo contendo o conteúdo dos atos e suas entidades. A equipe também realizou experimentos para mapear as entidades de um processo licitatório utilizando o *framework spaCy* (Figura 7).


Figura 7: Exemplo de resultado com spaCy.

## 6.2 Identificação de blocos de atos com *pytesseract*

Com o uso do *pytesseract*, que é uma biblioteca do Python para reconhecimento ótico de caracteres (OCR), nós conseguimos analisar os DODFs e identificar os blocos de texto que existem no arquivo pdf de cada DODF desde 2000 [10] (conforme pode ser visto na Figura 8).

Foi criado um script que gera um dataframe com o texto completo da seção 3 de cada um dos DODFs do período de 2000 a 2021 (como visto nas Figuras 9 e 10), separados por página, sendo que esses textos possuem tags que marcam os blocos de textos do DODFs, facilitando o processo de localizar o início e fim de um ato.

## 6.3 Implementação de um Autômato Determinístico Finito (DFA) para melhorar a extração de atos

Foi implementado um DFA que analisa cada linha do dataframe evidenciado na Figura 10. No estado 1 do DFA, é verificado se a linha representa algum dos títulos dos atos de licitação. No caso de falso, o DFA permanece no estado 1 e analisa a próxima linha, até analisar a última linha do texto. No caso de verdadeiro, o DFA passa para o estado 2, o que representa que um ato de licitação foi encontrado, e nesse caso, as linhas posteriores são adicionadas ao texto do ato. Caso a linha seja alguma das heurísticas que representam o fim do ato, o DFA volta para o estado 1.

Extraímos 38235 atos usando o DFA, os quais pode ser de um dos seguintes tipos:

1. AVISO DE ABERTURA DE LICITAÇÃO
2. AVISO DE ADJUDICAÇÃO E HOMOLOGAÇÃO
3. AVISO DE HOMOLOGAÇÃO E CONVOCAÇÃO


Figura 8: Divisão de blocos de textos gerados pelo *pytesseract*.

4. AVISO DE CONVOCAÇÃO
5. AVISO DE DECLARAÇÃO DE VENCEDOR
6. AVISO DE RESULTADO
7. AVISO DE RESULTADO DE JULGAMENTO
8. AVISO DE JULGAMENTO
9. AVISO DE LICITAÇÃO
10. AVISO DE JULGAMENTO DE HABILITAÇÃO
11. AVISO DE RESULTADO DE RECURSO E JULGAMENTO
12. AVISO DE SUSPENSÃO DE LICITAÇÃO
13. AVISO DE ADIAMENTO DE LICITAÇÃO
14. AVISO DE ALTERAÇÃO

| | file_name | number | day | month | year | page | text |
|---|---|---|---|---|---|---|---|
| 0 | DODF 021 31-01-2000 | 21 | 31 | 1 | 2000 | 24 | \n\nxxbob\nPÁGINA 24\nxxeob\n\nxxbob\nDIÁRIO O... |
| 1 | DODF 021 31-01-2000 | 21 | 31 | 1 | 2000 | 23 | \n\nxxbob\nNº 21 SEGUNDA-FEIRA, 31 JAN 2000\nx... |
| 2 | DODF 021 31-01-2000 | 21 | 31 | 1 | 2000 | 22 | \n\nxxbob\nPÁGINA 22 DIÁRIO FEDERAL SEGUNDA-FE... |
| 3 | DODF 021 31-01-2000 | 21 | 31 | 1 | 2000 | 21 | \n\nxxbob\nNº 21 SEGUNDA-FEIRA, 31 JAN 2000\nx... |
| 4 | DODF 021 31-01-2000 | 21 | 31 | 1 | 2000 | 20 | \n\nxxbob\nPÁGINA 20\nxxeob\n\nxxbob\nrensl e ... |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 155926 | DODF 186 01-10-2021 | 186 | 1 | 10 | 2021 | 5 | \n\nxxbob\nPÁGINA 5\nxxeob\n\nxxbob\nDiário Of... |
| 155927 | DODF 186 01-10-2021 | 186 | 1 | 10 | 2021 | 4 | \n\nxxbob\nPÁGINA 4\nxxeob\n\nxxbob\nDiário Of... |
| 155928 | DODF 186 01-10-2021 | 186 | 1 | 10 | 2021 | 3 | \n\nxxbob\nPÁGINA 3\nxxeob\n\nxxbob\nDiário Of... |
| 155929 | DODF 186 01-10-2021 | 186 | 1 | 10 | 2021 | 2 | \n\nxxbob\nPÁGINA 2 Diário Oficial do Distrito... |
| 155930 | DODF 186 01-10-2021 | 186 | 1 | 10 | 2021 | 1 | \n\nxxbob\nGOVERNO DO DISTRITO FEDERAL\nxxeob\... |

155931 rows × 7 columns

Figura 9: Dataframe gerado com a utilização do OCR.

| | Unnamed: 0 | Dodfs_list |
|---|---|---|
| 0 | 0 | DODF 001 03-01-2000\n\nxxbob\nPÁGINA 4\nxxeob\... |
| 1 | 1 | DODF 002 04-01-2000\n\nxxbob\nAGENTE FALA, A G... |
| 2 | 2 | DODF 003 05-01-2000\n\nxxbob\nGoverno\nTRIO\nF... |
| 3 | 3 | DODF 004 06-01-2000\n\nxxbob\nDO GOvERRO\nDIST... |
| 4 | 4 | DODF 005 07-01-2000\n\nxxbob\nDIÁRIO OFICIAL\n... |
| ... | ... | ... |
| 5315 | 5315 | DODF 195 18-10-2021\n\nxxbob\nGOVERNO DO DISTR... |
| 5316 | 5316 | DODF 196 19-10-2021\n\nxxbob\nGOVERNO DO DISTR... |
| 5317 | 5317 | DODF 197 20-10-2021\n\nxxbob\nPÁGINA 13\nxxeob... |
| 5318 | 5318 | DODF 198 21-10-2021\n\nxxbob\nGOVERNO DO DISTR... |
| 5319 | 5319 | DODF 199 22-10-2021\n\nxxbob\nGOVERNO DO DISTR... |

5320 rows × 2 columns

Figura 10: Dataframe com o texto completo da seção 3 de cada DODF.

15. AVISO DE REABERTURA
16. AVISO DE NOVA DATA DE ABERTURA

## 6.4 Extração de entidades

Durante a release 3 foi utilizado regex para muitas entidades dos atos de licitação. Como não teve muita variação na estrutura dos atos na release 4, mesmo que extraídos de outra maneira, muitos dos regex puderam ser reutilizados e outros foram apenas modificados. Nesse sentido, até o momento, conseguimos extrair as seguintes entidades:

1. **Número do processo**: que é uma forma de relacionar os atos de licitação;
2. **Data do ato**: que é como conseguimos criar a linha do tempo, sendo que é a data em que o ato foi escrito, ou a data do DODF em que o ato foi publicado;
3. **Id do ato**: que é uma forma de reconhecer e agrupar facilmente as fases do processo de licitação que foram extraídos;
4. **Texto do ato**: que mostra todo o texto do ato;

5. **Empresa**: que mostra quais empresas estão fazendo parte do processo licitatório ou que ganharam determinados itens do processo;

6. **CNPJ**: Facilita o rastreio de atos de uma determinada empresa.

7. **Valor da licitação**: juntamente com os itens licitados: que é a informação de qual foram os itens que determinada empresa venceu e qual foi o valor desses itens;

8. **Itens da licitação fracassados**: que representa quais itens não foram licitados por nenhuma empresa

## 6.5 Supervisão fraca para maior generalização das classes de atos extraídas com regex

Utilizamos as regex que já havíamos criado para iniciar o processo de rotular as entidades no intuito de criar um modelo preditivo e a marcação IOB que pode ser utilizada também para treinar modelos, como por exemplo, o algoritmo CRF, que é um modelo padrão para prever a sequência de rótulos mais provável que corresponde a uma sequência de entradas [2] (Figura 11).

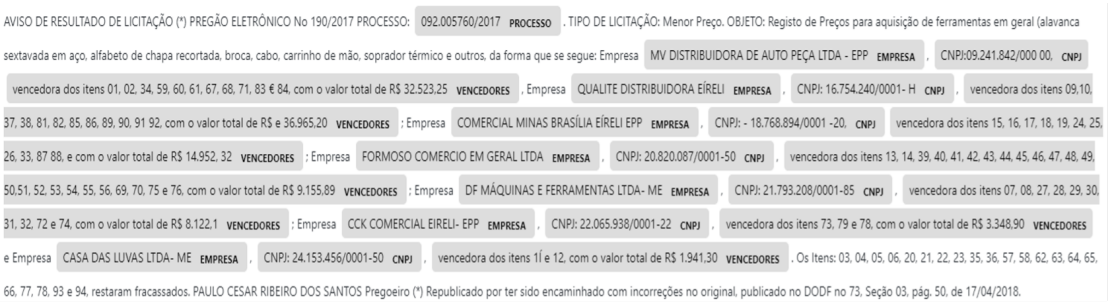



Figura 11: Texto rotulado utilizando o skweak.

## 6.6 Pesquisa sobre a implementação de banco de dados *NoSQL*

Uma investigação sobre o uso alternativo de um banco de dados *NoSQL* foi realizada também. Iniciamos uma pesquisa sobre o uso do *neo4j*, um banco de dados orientado a grafos como uma alternativa ao nosso atual banco de dados *PostgreSQL*. No entanto, outras prioridades assumiram nossa agenda e foi necessário pausar o estudo sobre o *neo4j*. Não chegamos a ter resultados práticos para o projeto.

Como próximos passos, a equipe planejou utilizar o volume de dados até aqui rotulados para iniciar uma fase de treinamento de modelos de IA com o intuito de trazer maior generalização às funcionalidades de identificação de atos e de entidades nomeadas. Neste momento estamos numa força tarefa de anotação de atos e entidades prioritários (levantados em reunião com cliente) junto com a equipe de contratos.

# 7 Contratos

Os objetivos principais da equipe foram primeiramente comprender a estrura dos contratos publicados no DODF, para então buscar a extração, a visualizaçao e a classificação desses contratos. A equipe passou por diversas metodologias e atividades, sendo que, inicialmente, focou-se na familiarização e treinamento de novos membros, posteriormente foram aplicados esses conhecimentos para a extração e classificação de contratos usando técnicas variadas, desde regex ao uso de ferramentas de aprendizado de máquina.

## 7.1 Extração por Expressões Regulares

O extrator de contratos foi feito utilizando buscas de padrões através do regex. Para a busca dos padrões, foi feita a leitura de diários oficiais de diferentes anos (entre 2000 e 2021), identificando-se os títulos que definem extratos de contrato (como "EXTRATOS CONTRATUAIS", "EXTRATOS DE CONTRATOS", "EXTRATO DE CONTRATO" etc.), bem como de outros blocos textuais, que delimitam o fim dos blocos de extratos de contrato.

De forma simplificada, o extrator executa os passos:

1. Identificação de linha inicial de cada bloco textual em um documento do DODF;
2. Identificação de cada título de extrato de contrato no documento;
3. Extração do trecho que contém cada extrato de contrato, que se inicia com um título e se termina no início do próximo bloco textual.

Na identificação dos títulos dos extratos de contrato, foi utilizada uma expressão regular que buscava:

"...EXTRAT... D... CONTRAT..."ou "...EXTRAT... CONTRAT.."(expressões simplificadas, com partes extraídas substituídas por reticências).

Essas duas expressões permitem captar diferentes títulos de extratos de contrato, desde títulos curtos como "EXTRATO DE CONTRATO Nº 2014/296" a títulos maiores, como "EXTRATO DO CONTRATO PARA PRESTAÇÃO DE SERVIÇOS Nº 04/2015-SSP, NOS TERMOS DO PADRÃO Nº 04/2002, instituído pelo Decreto/DF nº 23.287/2002".

Já para a identificação de um bloco textual genérico, foi utilizada expressão que busca linhas apenas com caracteres maiúsculos que estão centralizados na página e aparecem depois de espaçamento. Aqui foi possível se beneficiar da forma como os textos do documento foram extraídos, captando também centralização de linhas.

A validação do extrator foi feita a partir da seleção aleatória de DODFs em cada ano entre 2000 e 2021. De cada ano foram extraídos 6 documentos aleatórios. Desses 6, no entanto, foram analisados apenas 2, escolhidos a partir de ordem crescente. Caso algum desses 2 não contivesse nenhum extrato de contrato, passou-se para o próximo documento, até cada ano ter 2 documentos analisados.

Assim, foram analisados 44 documentos, com um total de 447 extratos de contrato.

Com os dados obtidos pelo extrator, foi possível observar que houve 390 extrações desses documentos com o conteúdo integral, sem nenhum tipo de falha. No entanto, outros 24 extratos foram extraídos com perdas pouco significativas, de até 1 linha.

Com isso, o processo de validação do extrator apontou uma acurácia média de cerca de 87,25% para textos sem perda nenhuma e de no máximo de 92,62% para textos sem nenhuma perda ou com perdas pouco significativas. Na Figura 12 estão listados os valores de acurácia obtido em extrações de Diários Oficiais dentro de um mesmo ano.

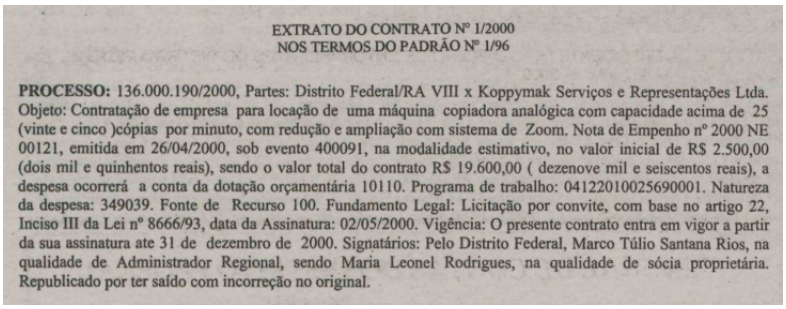
Por exemplo, a Figura 13 mostra um extrato de contrato publicado no DODF de 5 de junho de 2000. E a Figura 14 apresenta como o extrato de contrato da Figura 13 foi extraído.

| year | extracted_texts | total_texts | accuracy |
|---|---|---|---|
| 2000 | 15 | 18 | 83.33 |
| 2001 | 10 | 12 | 83.33 |
| 2002 | 9 | 10 | 90.00 |
| 2003 | 29 | 32 | 90.62 |
| 2004 | 24 | 30 | 80.00 |
| 2005 | 15 | 17 | 88.24 |
| 2006 | 13 | 13 | 100.00 |
| 2007 | 9 | 11 | 81.82 |
| 2008 | 33 | 37 | 89.19 |
| 2009 | 15 | 21 | 71.43 |
| 2010 | 11 | 16 | 68.75 |
| 2011 | 23 | 25 | 92.00 |
| 2012 | 16 | 18 | 88.89 |
| 2013 | 25 | 26 | 96.15 |
| 2014 | 38 | 39 | 97.44 |
| 2015 | 7 | 7 | 100.00 |
| 2016 | 14 | 16 | 87.50 |
| 2017 | 26 | 27 | 96.30 |
| 2018 | 14 | 18 | 77.78 |
| 2019 | 14 | 18 | 77.78 |
| 2020 | 13 | 17 | 76.47 |
| 2021 | 16 | 18 | 88.89 |

Figura 12: Tabela com a acurácia do extrator para cada ano nas amostras. A coluna "total texts"mostra o número total de extratos nas amostras, enquanto "extracted texts"mostra o número de extrações bem-sucedidas.

A aplicação de regex foi consequência do estudo das informações a serem extraídas dos atos de contratos. Essas informações são compilações das demandas das Secretarias de Controle Externo.

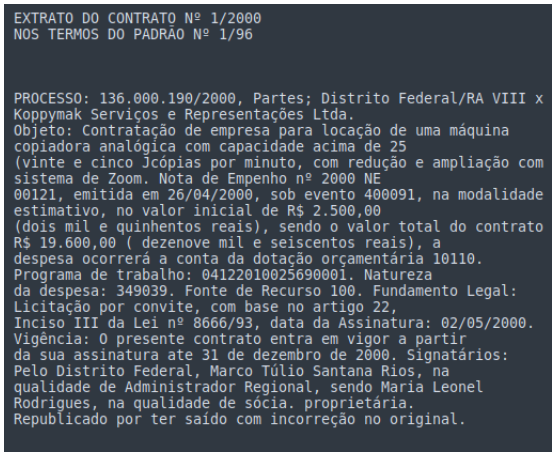
Com a descrição detalhada dos atos foram construídas expressões regulares para extração das informações. Porém, sabe-se que a utilização de expressão regular não é abrangente ao ponto de cobrir todas as regras possíveis, o que torna falha a extrações de vários atos (que ocasionalmente não tiveram os padrões descritos). Por esse motivo, utilizou-se técnicas de aprendizado de máquinas. Em especial, técnica de classificação supervisionada.



EXTRATO DO CONTRATO Nº 1/2000
NOS TERMOS DO PADRÃO Nº 1/96

**PROCESSO:** 136.000.190/2000, Partes: Distrito Federal/RA VIII x Koppymak Serviços e Representações Ltda. Objeto: Contratação de empresa para locação de uma máquina copiadora analógica com capacidade acima de 25 (vinte e cinco )cópias por minuto, com redução e ampliação com sistema de Zoom. Nota de Empenho nº 2000 NE 00121, emitida em 26/04/2000, sob evento 400091, na modalidade estimativo, no valor inicial de R$ 2.500,00 (dois mil e quinhentos reais), sendo o valor total do contrato R$ 19.600,00 ( dezenove mil e seiscentos reais), a despesa ocorrerá a conta da dotação orçamentária 10110. Programa de trabalho: 04122010025690001. Natureza da despesa: 349039. Fonte de Recurso 100. Fundamento Legal: Licitação por convite, com base no artigo 22, Inciso III da Lei nº 8666/93, data da Assinatura: 02/05/2000. Vigência: O presente contrato entra em vigor a partir da sua assinatura ate 31 de dezembro de 2000. Signatários: Pelo Distrito Federal, Marco Túlio Santana Rios, na qualidade de Administrador Regional, sendo Maria Leonel Rodrigues, na qualidade de sócia proprietária. Republicado por ter saído com incorreção no original.

Figura 13: Exemplo de extrato de contrato publicado no DODF de 5 de junho de 2000.



EXTRATO DO CONTRATO Nº 1/2000
NOS TERMOS DO PADRÃO Nº 1/96

PROCESSO: 136.000.190/2000, Partes; Distrito Federal/RA VIII x Koppymak Serviços e Representações Ltda.
Objeto: Contratação de empresa para locação de uma máquina copiadora analógica com capacidade acima de 25
(vinte e cinco )cópias por minuto, com redução e ampliação com sistema de Zoom. Nota de Empenho nº 2000 NE
00121, emitida em 26/04/2000, sob evento 400091, na modalidade estimativo, no valor inicial de R$ 2.500,00
(dois mil e quinhentos reais), sendo o valor total do contrato R$ 19.600,00 ( dezenove mil e seiscentos reais), a
despesa ocorrerá a conta da dotação orçamentária 10110.
Programa de trabalho: 04122010025690001. Natureza
da despesa: 349039. Fonte de Recurso 100. Fundamento Legal:
Licitação por convite, com base no artigo 22,
Inciso III da Lei nº 8666/93, data da Assinatura: 02/05/2000.
Vigência: O presente contrato entra em vigor a partir
da sua assinatura ate 31 de dezembro de 2000. Signatários:
Pelo Distrito Federal, Marco Túlio Santana Rios, na
qualidade de Administrador Regional, sendo Maria Leonel
Rodrigues, na qualidade de sócia. proprietária.
Republicado por ter saído com incorreção no original.

Figura 14: Exemplo de extrato de contrato extraído, localizado no DODF de 5 de junho de 2000.

# 8 Pesquisa

Esta seção serão descritas as principais pesquisas resultantes da *Relese 4* deste projeto. Inicialmente, é apresentado a lista de publicações envolvendo os alunos de mestrado. Em seguinda, serão apresentados os resumos de três dissertações de mestrado já defendidas por alunos bolsistas do projeto. As dissertações estão em inglês, e podem ser encontradas na base de dissertações e teses da UnB. Em seguida, serão apresentados os resumos da pesquisa de dois alunos de mestrado que ainda estão no projeto. As pesquisas desses alunos correspondem a área de Aprendizado com supervisão fraca e segmentação de textos.

As publicações obtidas por esses alunos estão listadas a seguir:

- **Deep Active-Self Learning Applied to Named Entity Recognition** [12]
- **Victor: a dataset for Brazilian legal documents classification** [8]
- **Inferring the source official texts: can SVM beat ULMFiT?** [9]
- **Topic Modelling Brazilian Supreme Court Lawsuits** [7]
- **Data Augmented 3D Semantic Scene Completion With 2D Segmentation Priors** [3]

## 8.1 Deep Active Learning Approaches to the task of Named Entity Recognition[13]

**Dissertação aprovada e defendida pelo aluno José Reinaldo da Cunha S. A. V. da Silva Neto[15]. Resumo:** *Redes neurais profundas é o atual estado-da-arte para uma variedade de tarefas variadas no campo de processamento de linguagem natural e visão computacional, mas eles exigem grandes quantidades de dados rotulados para serem treinados e alcançarem bons resultados. Algoritmos de aprendizado ativo profundo são projetados para reduzir a quantidade de dados rotulados para treinar esses modelos. Neste trabalho de pesquisa foram identificadas lacunas em trabalhos recentes na literatura em algoritmos de aprendizado ativo profundo aplicados na tarefa de detecção de entidades nomeadas, e também são propostos potenciais soluções para tais lacunas. Em particular, trabalhos recentes da literatura dependem do conjunto de teste para aplicar parada antecipada dos modelos de treinamento durante o processo de aprendizado ativo. A separação dos dados rotulados a fim de criar um conjunto de validação é inconveniente no cenário com poucos recursos. Então, neste trabalho foi proposto uma estratégia de atualização automática em cada época de treinamento (DUTE – Dynamic Update of Training Epochs) que atua como uma técnica de parada antecipada não supervisionada. Resultados experimentais sugerem que a estratégia proposta, a DUTE, é capaz de manter o desempenho do modelo treinado, quando comparada com técnicas tradicionais de parada antecipada e também não exige conjunto de validação. Também foi investigado a técnica de auto-rotulação (self-learning) como uma opção viável para mais redução do custo de rotulação no cenário de aprendizado ativo. Além disso, experimentos também investigaram estragégias de rotulação a nível de tokens e sentenças. Foi observado que, apesar de grande esforço, auto-aprendizado a nível de sentenças não traz melhorias significativas comparado com trabalhos anteriores. Entretanto, auto-rotulação a nível de tokens apresentou resultados promissores em modelos treinados com algoritmos estado-da-arte e exigiram menos dados rotulados. Mais especificadamente, experimentos aplicados na base CoNLL2003 mostraram que a técnica propos de auto-rotulação a nível de token alcançou resultados similares ao estado-da-arte com apenas* 24% *menos dados rotulados.*

## 8.2 Domain-specific datasets for document classification and named entity recognition[1]

**Dissertação aprovada e defendida pelo aluno Pedro Henrique Luz de Araujo.[16] Resumo:** *Uma enorme quantidades de dados é produzida diariamente e uma parte significativa está no formato de texto, com diversos domínios (posts de míndias sociais, livros, notícias, relatórios oficiais, processos legais). Essa rica fonte de informação pode produzir conhecimento útil. O desafio é que textos em linguagens naturais são não estruturados: um processamento é requerido para obter entendimento e conhecimento estruturado dos dados. Apesar do processamento em linguagem natural alcançar grande progresso na última década, modelos atuais requerem uma grande quantidade de exemplos anotados e tendem a não generalizar além de dados do domínio. Técnicas recentes de transferência de conhecimento podem mitigar essas necessidades, mas conjuntos de dados específicos do domínio ainda são*

[15] `https://repositorio.unb.br/handle/10482/42729`
[16] `https://repositorio.unb.br/handle/10482/42415`

*necessários para ajustar modelos pré-treinados e para avaliação. Neste trabalho, foi proposto três conjunto de dados específicos de domínio com dados anotados para duas tarefas de processamento de linguagem natural: classificação de documentos e reconhecimento de entidades nomeadas. Para estabelecer um benchmark para trabalhos futuros em domínio legal e administrativo, esses dados são treinados, avaliados e comparados com diferentes modelos. Primeiramente, foi proposto um conjunto de dados para a tarefa de reconhecimento de entidades nomeadas em documentos legais com entidades de domínios específicos e treinado um modelo biLSTM-CRF nos dados. Em seguida, foi proposto um conjunto de dados do Supremo Tribunal Federal Brasileiro anotado para duas tarefas de classificação; foram treinados modelos neurais profundos e modelos tradicionais com e sem modelos de sequência; tópicos foram avaliados por meio da técnica Latent Dirichlet Allocation. Finalmente, foi proposto um conjunto de dados do Diário Oficial com dados rotulados e não rotulados e vários modelos foram comparados e com técnicas estado-da-arte com métodos de transferência de conhecimento.*

## 8.3 The emergence of an Information Bottleneck Theory of Deep Learning[4]

**Dissertação aprovada e defendida pelo aluno Frederico Guth[17]. Resumo:** *Na última década, testemunhamos uma infinidade surpreendente de técnicas bem sucedidas aplicando Aprendizado Profundo. Apesar desses muitos sucessos, pode-se estar novamente escalando um pico de expectativas infladas. No passado, a falsa solução era "adicionar poder computacional aos problemas", hoje tenta-se "empilhar dados". Esse comportamento desencadeou uma corrida por dados em várias corporações, levantando preocupações sobre privacidade e concentração de poder. É um fato conhecido, no entanto, que é possível aprender com muito menos amostras: os humanos mostram uma capacidade de generalização muito melhor do que a inteligência artificial atual. Para alcançar tal feito, é necessário um melhor entendimento de como funciona a generalização, em particular em redes neurais profundas. No entanto, a prática do aprendizado de máquina moderno ultrapassou seu desenvolvimento teórico. Em particular, "medidas tradicionais de complexidade do modelo lutam para explicar a capacidade de generalização de grandes redes neurais artificiais". Ainda não existe uma nova teoria geral de aprendizagem estabelecida que lide com esse pseudo-paradoxo. Em 2015, Naftali Tishby e Noga Zaslavsky publicaram uma teoria seminal da aprendizagem baseada no conceito teórico da informação do princípio do gargalo com o potencial de preencher essa lacuna. Esta dissertação teve como objetivo investigar os esforços usando o princípio do gargalo de informação para explicar as capacidades de generalização de redes neurais profundas, consolidá-las em um resumo abrangente e analisar sua relação com a teoria atual de aprendizado de máquina.*

## 8.4 Pesquisa em aprendizado por supervisão fraca

Nesta última década, o aprendizado profundo tem alcançado sucesso considerável em tarefas de classificação de dados textuais, evitando grande parte do processo manual de engenharia de características. Em dados de domínio específico, como os do Diário Oficial do Distrito

[17] `https://teodecampos.github.io/fred_guth/guth_msc_unb_final_2022.pdf`

Federal, a elaboração de características pode ser difícil e custosa. Para aproveitar características latentes nos dados, o processamento em redes neurais profundas são compostas por arquiteturas complexas que requerem uma grande quantidade de dados. Com isso, chega-se novamente à necessidade de dados rotulados. Mesmo técnicas de aprendizado ativo podem ser insuficientes para suprir a faminta necessidade de dados rotulados que esses modelos exigem.

Para realmente reduzir os encargos da anotação de dados de treinamento, uma estratégia interessante, e promissora para o contexto deste projeto, é recorrer a fontes mais baratas de dados rotulados. Acredita-se na possibilidade de usar técnicas baseadas em distância para encontrar documentos similares aos já anotados, aproveitar modelos já treinados e que também foram úteis para a estratégia de aprendizado ativo, ou mesmo utilizar heurísticas e regras previamente conhecidas por especialistas do domínio.

Essas estratégias de automatização da rotulação são propostas recentes, mas já apresentam frameworks programáticos para essas atividades. Especificamente, os usuários codificam fontes de supervisão fracas, por exemplo, heurística, bases de conhecimento e modelos pré-treinados, na forma de funções de rotulagem. As funções de rotulagem são rotinas definidas pelo usuário e que são capazes de fornecer rótulos para algum subconjunto dos dados. Essas funções devem ser variadas de modo que, coletivamente, geram um grande conjunto de rótulos de treinamento.

Os frameworks para o aprendizado fraco possibilitam a incorporação das várias funções de rotulação. É esperado que os rótulos gerados por essas funções de rotulação sejam ruidosos e conflitantes. Para tratar esse problema, os frameworks possuem agregadores de funções de rotulagem que produzem dados rotulados baseados no voto (ou rotulação) fornecido por essas funções.

As funções de rotulação são codificações variadas. Em aplicações práticas, um especialista do domínio pode criar expressões regulares que indicam a rotulação de alguns documentos. Algumas estratégias utilizam bases externas para encontrar alguma relação de similaridade entre os dados não rotulados e seus supostos rótulos. Ou mesmo, pode-se utilizar modelos pré-treinados, que não apenas indiquem os rótulos, mas que possam realizar tarefas, como por exemplo, detecção de entidades nomeadas, e essas entidades auxiliar na indicação do rótulo.

Portanto, acredita-se que a supervisão fraca ofereça uma direção promissora para aumentar o volume de dados rotulados, e que também diminua o esforço humano.

## 8.5 Pesquisa em segmentação de texto

Os modelos de linguagem são considerados o atual estado da arte para resolução de diversos problemas relacionados à extração de conhecimento de dados textuais. Pode-se destacar os modelos utilizados para detecção de entidades nomeadas, que possibilitam a extração automática de entidades de bases textuais com acurácia superior aos modelos antecedentes. No entanto, tais modelos necessitam de uma entrada com um limitado número de sentenças que constituem um bloco textual a ser processado.

No contexto de documentos oficiais, o Diário Oficial do Distrito Federal realiza a organização das informações por meio de seções que compreendem diversos tipos de publicações, como por exemplo: atos de pessoal; extratos de contratos; termos aditivos; extratos de edi-

tais entre outros. Nessa perspectiva, a criação de modelos para a extração informações de cada uma dessas publicações necessita que seja realizada a segmentação dos blocos textuais relativos à cada tipo publicação.

Nesse sentido, foi realizado o levantamento bibliográfico a fim de identificar os algoritmos e abordagens consideradas o estado da arte relacionadas ao problema de segmentação de texto. Dentre as técnicas encontradas a maioria fez uso de aprendizado supervisionado para identificação dos segmentos, no entanto o viés de segmentação adotado em todos os trabalhos foi de identificação de tópicos por meio do contexto textual.

Levando em consideração as técnicas encontradas, foi proposta uma nova abordagem análoga aos modelos de reconhecimento de entidade nomeada. Visto que no caso do Diário Oficial não é possível realizar a segmentação por meio do contexto semântico, pois a linguagem utilizada é especializada e compartilha de um vocabulário similar entre as seções. Foram realizados experimentos preliminares utilizando modelos baseados em CRF-LSTM, os quais obtiveram resultados promissores no que diz respeito à identificação dos segmentos relacionados aos atos de pessoal. Como trabalhos futuros serão realizados novos experimentos utilizando algoritmos baseados em mecanismos de atenção

# 9 Considerações Finais

O projeto tem evoluído com sucesso. O grupo tem apresentado progresso, não somente em termos de pesquisa, mas também em termos da implementação de ferramentas. O protótipo desenvolvido (versão para testes) seguem as especificações contidas na Seção II do DODF que são do interesse da Secretaria de Fiscalização de Pessoal do GDF e foi apresentado no IV Whorkshop do Kendle.

# 10 Equipe

Além dos autores deste relatório, o trabalho realizado no referido período contou com as participações dos seguintes bolsistas: Daniel de Sousa Oliveira Melo Veras, Ian filipe Pontes Ferreira, Thais Rebouças Araujo, Maicon Rodrigues Queiroz, Vitor De Oliveira Araujo Araruna, Jonatas Gomes Barbosa da Silva, Rafael Amaral Soares, Vitor Vasconcelos de Oliveira, Larissa Santana de Freitas Andrade, Felipe Xavier Barbosa da Silva, Manuela Matos Correia de Souza, Gabriel Mendes Ciriatico Guimaraes, Gabriel da Silva Corvino Nogueira, Matheus Stauffer Viana de Oliveira e Tatiana Franco Pereira. Também contamos com o trabalho dos alunos de pós-graduação Micael Filipe Ribeiro de Lima, Lucelia Vieira, José Reinaldo Neto.

# Referências


[1] Pedro Henrique Luz de Araújo. Domain-specific datasets for document classification and named entity recognition. Dissertação (mestrado em informática), Universidade de Brasília, Brasília, 2021.

[2] BAVALPREET. Ner using crf, Mar 2020.

[3] Aloisio Dourado, Frederico Guth, and Teofilo de Campos. Data augmented 3d semantic scene completion with 2d segmentation priors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3781–3790, January 2022.

[4] Frederico Guth. The emergence of an information bottleneck theory of deep learning. Dissertação (mestrado em informática), Universidade de Brasília, Brasília, 2022.

[5] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[6] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, page 282–289, 2001.

[7] Pedro H. Luz de Araujo and Teófilo E. de Campos. Topic modelling brazilian supreme court lawsuits. In *International Conference on Legal Knowledge and Information Systems (JURIX)*, Frontiers in Artificial Intelligence and Applications, pages 113–122, Prague, Czech Republic, December 9-11 2020. IOS Press.

[8] Pedro H. Luz de Araujo, Teófilo E. de Campos, Fabricio Ataides Braz, and Nilton Correia Silva. Victor: a dataset for Brazilian legal documents classification. In *Language Resources and Evaluation Conference (LREC)*, Marseille, France, May 2020.

[9] Pedro H. Luz de Araujo, Teófilo E. de Campos, and Marcelo Magalhaes Silva de Sousa. Inferring the source official texts: can SVM beat ULMFiT? In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), Evora, Portugal, March 2-4 2020. Springer.

[10] madmaze. pytesseract 0.3.9, Feb 2022.

[11] Christopher Marshall. What is named entity recognition (ner) and how can i use it?, Dec 2019.

[12] José S. Neto and Thiago Faleiros. Deep active-self learning applied to named entity recognition. In *Anais da X Brazilian Conference on Intelligent Systems*, Porto Alegre, RS, Brasil, 2021. SBC.

[13] Silva Neto, José Reinaldo da Cunha Santos Aroso, et al. Deep active learning approaches to the task of named entity recognition. Dissertação (mestrado em informática), Universidade de Brasília, Brasília, 2021.

[14] Paul R Niven and Ben Lamorte. *Objectives and key results: Driving focus, alignment, and engagement with OKRs*. John Wiley & Sons, 2016.

[15] Jorge Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.

[16] Marek Rei, Gamal K. O. Crichton, and Sampo Pyysalo. Attending to characters in neural sequence labeling models. *arXiv preprint arXiv:1611.04361*, 2016.

[17] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

[18] Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*, 2017.