Relatório Técnico Parcial 3 do Projeto KnEDLe / NIDO

Tatiana Franco Pereira Matheus Stauffer Viana de Oliveira
Isaque Alves Vinícius R. P. Borges Thiago de Paulo Faleiros
Fabrício A. Braz Nilton Correia Silva Carolina Alves Okimoto
Teófilo E. de Campos

Universidade de Brasília

http://nido.unb.br

6 de setembro de 2021

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 um fruto de uma parceria entre a Universidade de Brasília, FAPDF e FINATEC[1]. Trata-se do relato das atividades e resultados produzidos na terceira fase (*release 3*) do Projeto, de 31/12/2020 a 30/06/2021.[2]

2 Viabilidade Técnica

Durante a terceira *Release* do projeto, o foco foi na reestruturação e organização do time e do projeto. Com a saída de alguns membros e a chegada de outros, novas ideias e uma nova cultura pode ser vislumbrada. Um fator que foi particularmente impactante neste período foi o retorno do time do formado por membros da Engenharia de Software da Faculdade do Gama (FGA), liderados pelos professores Fabrício e Nilton. Esses novos membros introduziram ao projeto ferramentas de gestão de dados (MetaFlow) e uma melhor ferramenta para gestão de *sprints* (ZenHub), a qual passou a ser adotada pelo projeto como um todo a partir do final desta fase.

[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]Além dos autores deste relatório, o trabalho realizado no referido período contou com as participações dos seguintes bolsistas: José Reinaldo Neto, Khalil Carsten do Nascimento, Renato Avellar Nobre, Leonardo Maffei da Silva, Lívia Gomes Costa Fonseca, Pedro Henrique Luz de Araujo, Frederico Guth. Além desses bolsistas, também contamos com a participação dos seguintes voluntários: Lindeberg Pessoa Leite, Patricia Medyna Lauritzen de Lucena Drumond.

O ZenHub é uma ferramenta de gestão de projetos poderosa, que permite a integração de utilização das *issues* do GitHub 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*.

2.1 *Objective and Key Results (OKR)*

Durante a última *Release* observou-se uma manutenção da situação dos riscos reportada anteriormente, com destaque para os atrasos do repasse financeiro pela FAPDF, situação recorrente desde o início do projeto. Logo, ao longo da *Release*, houve negociações com os representantes da referida fundação, redundando em readequações no plano de trabalho promovidos pela coordenação do projeto. Esse risco se materializou no bloqueio proporcionado pela ausência de equipamentos condicionantes a realização da pesquisa.

Essa Release foi marcada pela alta rotatividade dos membros do projeto, em razão de encerramento de contratos de membros motivada por formatura de graduandos, além da substituição do Coordenador.

Por outro lado, alguns riscos não precisaram ter uma atuação, pois se mantiveram em nível baixo, como o **Não cumprimento das metas (OKR)** e a **Falta de controle/gestão das atividades**. Já o risco **Dificuldades de adesão à metodologia** necessitou receber de mais atenção nas primeiras *Sprints*, considerando a entrada dos novos membros. Isso demandou algumas adaptações dessa nova equipe à cultura e aos ritos do projeto. Vale ressaltar que tal fato possa acontecer novamente com as mudanças na metodologia e a chegada de novos bolsistas, mas que pode ser minimizado a partir de treinamentos e reuniões de alinhamento.

2.2 Gestão de Riscos

Na *Release* 3, não houve a necessidade de manter um acompanhamento muito próximo dos riscos, portanto mantivemos atualizações sempre que foi necessário. Alguns riscos identificados se mantiveram alto durante todo o projeto como o repasse de verba da FAPDF, que desde o início se manteve em atraso. A falta dos repasses impactou diretamente pela carência de equipamentos, atrasos nos trabalhos, e escassez de recursos humanos. Porém, é importante salientar que a FAPDF sempre esteve aberta para o diálogo e, a medida do possível, recebeu as solicitações enviadas.

Outros pontos de atenção foram levados em consideração durante essa *Release*. Um ponto importante foi a grande rotatividade durante a execução com colaboradores no fim do contrato, novos times chegando ao projeto, e o Coordenador que foi substituído no final desse período.

Por outro lado, alguns riscos não precisaram ter uma atuação pois se mantiveram baixo, como: **Não cumprimento das metas (OKR)**, **Falta de controle/gestão das atividades**. Já o risco **Dificuldades de adesão à metodologia** teve que receber um pouco mais

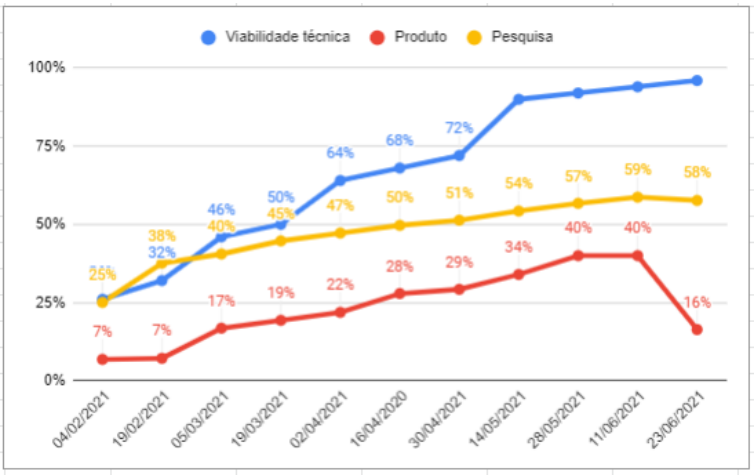


Figura 1: Evolução das três grandes áreas do projeto nesta *release* desde que a técnica OKR foi implementada. O eixo das ordenadas representa o percentual de objetivos alcançados, o qual foi computado a partir da coleção de resultados-chave necessários para cada objetivo. A queda na curva de produto se justifica pela inserção de novos objetivos que não foram cumpridos até a referida data.

de atenção nas primeiras *Sprints* devido a adesão de novos colaboradores no projeto.

A Figura 2 apresenta a evolução dos riscos desde o início do projeto. A *Release* 3 que teve coletar a partir da *Sprint* 7, onde apresenta um controle constante dos riscos.

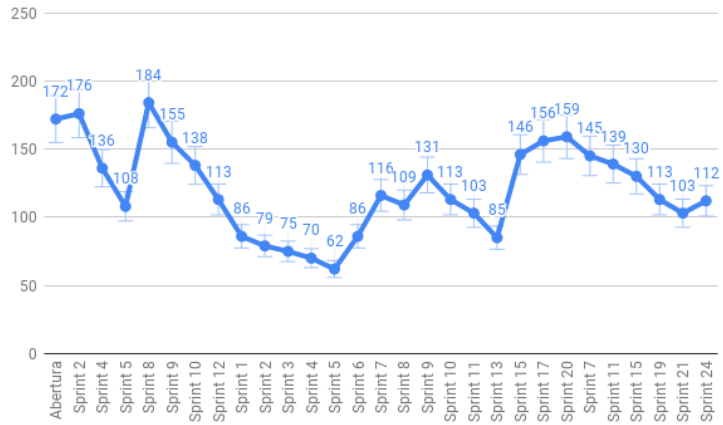


Figura 2: *Risk burndown* Geral da *Release 3*.

3 Resultados

O projeto KnEDLe possui particularidades, em especial, o desenvolvimento e inovação por meio da pesquisa aplicada em Aprendizado de Máquina. Por isso, nem todas as práticas da Engenharia de Software foram contempladas. Assim, foi necessário inserir na *Release 3* uma etapa de Revisão da Literatura, em que foi aproveitado o trabalho de conclusão de curso do bolsista Isaque Alves, que buscou e analisou as práticas e métodos utilizadas pelo mercado e pela academia. A ideia foi entender essas particularidades, buscando-se casos de estudos de empresas já consolidadas no mercado para auxiliar no gerenciamento desse projeto. Esse material está descrito em um artigo em preparação para submissão a um periódico.

A aplicação do OKR auxiliou na organização e na transparência do projeto. Durante essa *release*, a aplicação e utilização foi um experimento para verificar se a equipe se adaptaria. Essa metodologia foi bem adotada, mas ainda precisa de melhorias para a próxima *release*, como verificar e definir as atividades da *sprint* com um foco maior em atingir os objetivos e as atividades chaves.

Mesmo com os riscos mencionados, o projeto conseguiu evoluir e a primeira rodada do processo de anotação foi garantida, passando pela curva de aprendizado de desenvolvimento. O projeto prosseguiu em um ritmo mais lento do que o planejado originalmente, mas não houve interrupção, obtendo-se resultados satisfatórios durante essa *release*.

3.1 Produto

Esta seção resume o progresso realizado nas tarefas que foram centradas no desenvolvimento de ferramentas de software que podem ser usadas para extrair informações (Seção 3.1.1), anotar dados (Seção 3.1.2), visualizar e interagir com os dados de uma maneira intuitiva (Seção 3.1.3).

3.1.1 DODFMiner

O primeiro produto do Projeto KnEDLe, o DODFMiner, foi desenvolvido na forma de uma biblioteca da linguagem Python. A biblioteca foi concebida tanto com o intuito de ser usada em scripts Python como programa executável em terminal. Atualmente ela possui duas principais funções: executar o download dos diários oficiais do Distrito Federal e extrair tanto a informação textual quanto informações relativas a publicações.

Seguindo a lógica de suas funcionalidades, os módulos do DODFMiner são implementados de maneira similares. A ilustração dos módulos do DODFMiner está apresentado 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* contém todas as funcionalidades para extração das publicações e segmentos das informações textuais anteriormente extraídas pelo pure. A biblioteca pode ser encontrada no repositório do github: `https://github.com/UnB-KnEDLe/DODFMiner` e a documentação para a biblioteca pode ser encontrada: `https://dodfminer.readthedocs.io/en/stable/`.

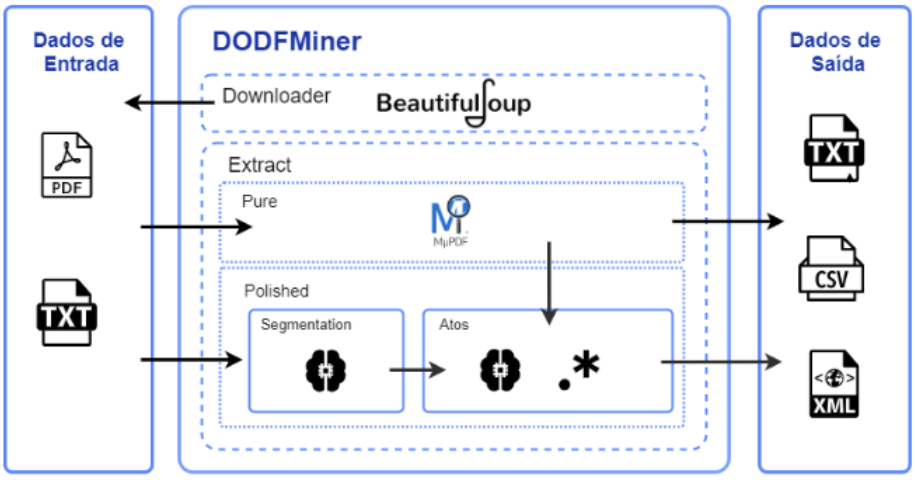


Figura 3: Ilustração da arquitetura dos componentes da ferramenta DODFMiner.

3.1.2 Anotação de Textos

Na *Release* 2, o processo de anotação de documentos do Diário Oficial do Distrito Federal (DODF) foi iniciado, tendo os próprios membros bolsistas do projeto como os anotadores. Um conjunto de 100 DODFs foi definido pela equipe de extração para que fossem anotados, obtendo-se assim o *corpus* Padrão Ouro. Em seguida, a equipe de anotação realizou uma análise preliminar na frequência dos atos em DODFs, apresentada na Tabela 1. Tal informação é importante para alocar mais anotadores aos atos mais frequentes no documentos, de forma a balancear o esforço de anotação entre os voluntários. Considera-se que documentos com a presença, em média, de mais de 20 atos como "Alta"; entre 5 e 19 atos por documento como "Média"; e, documentos com menos de 5 atos como "Baixa".

Tabela 1: Frequência média para cada ato considerando dois documentos do Diário Oficial do Distrito Federal: edições de 13 de Agosto e 21 de Agosto de 2020.

Ato	Frequência Média	Ocorrência	Anotadores alocados
Nomeação	53	Alta	3
Exoneração	49	Alta	3
Retificação	38	Alta	3
Substituição de função	37	Alta	3
Ato Tornado Sem Efeito	16	Média	2
Cessão	9	Média	2
Reversão	1	Baixa	2
Abono de Permanência	1	Baixa	2

Inicialmente, um treinamento foi fornecido pela equipe de anotação por meio de vídeos gravados explicando o uso da ferramenta NidoTat[3], com o apoio de um tutorial de anotação[4], detalhando as entidades nos atos de pessoal do DODF. Uma equipe no Microsoft Teams

[3] `http://nido.cic.unb.br/`

[4] `https://github.com/UnB-KnEDLe/tutorial_annotation_teamtat`

foi criada pela equipe de anotação para tratar exclusivamente de assuntos relacionados ao processo de anotação. Um conjunto de 3 DODFs foram disponibilizados para a ambientação de todos os anotadores com o NidoTat e seus recursos de interação. A anotação consiste em um conjunto de DODFs, aqui denominado informalmente como *Batch* e é constituída por duas fases:

- **Fase de rotulação:** um anotador A rotula os blocos de texto correspondentes ao ato X em um conjunto de DODFs $\mathcal{D}_1$, como também suas entidades;
- **Fase de revisão:** o anotador A revisa (podendo corrigir rotulações incorretas) as entidades e os blocos de texto associados ao ato X em um conjunto de DODFs $\mathcal{D}_2$, que foram rotulados por um outro anotador B. Em seguida, para cada bloco de texto relacionado ao ato X, o anotador A cria uma relação entre esse bloco e suas respectivas entidades.

O processo de anotação "de fato" começou ainda na *Release 2* com os membros do KnEDLe com a definição de um conjunto de 33 DODFs. Aproximadamente 6 documentos foram alocados para cada anotador de forma que não ocorreram sobreposição de anotadores rotulando um mesmo ato em um mesmo documento. Tal estratégia visa maximizar a anotação de documento, considerando a grande quantidade de documentos, atos e entidades, em contrapartida aos poucos anotadores. Em resumo, o processo de anotação utilizando os membros do projetos progrediu em ritmo lento, sofrendo algumas paralisações, devido aos seguintes fatores:

- Os membros do KnEDLe não são especialistas em linguagem jurídica, como também não estavam familiarizados com as diferentes entidades e atos presentes no DODF. Isso gerava dúvidas na rotulação das entidades, muitas vezes demandando consultas à equipe de anotação para solucioná-las.
- Os bolsistas tinham que destinar metade do seu tempo para o processo de anotação, enquanto que a outra metade era utilizada para as atividades de pesquisa no projeto (além das adividades de seus cursos). Assim, os membros tiveram seus trabalhos paralelizados, ocasionando queda de produtividade, tanto na anotação, quanto em suas pesquisas;
- Alguns anotadores tiveram que deixar o processo de anotação no decorrer de sua execução devido a problemas de saúde e à necessidade de priorizar outras atividades dentro do projeto. Especificamente, os alunos de pós-graduação tinham que priorizar o cumprimento de prazos no Programa de Pós-Graduação em Informática (PPGI) do CIC/UnB, como a escrita de artigos e qualificações. Tal fato acarretou na realocação do documentos e de atos aos anotadores remanescentes, aumentando o esforço de anotação e atrasos nos prazos de entrega das anotações;
- Problemas e/ou ausência de funcionamento no NidoTat ocasionados por quedas de energia no CIC/UnB ou pela configuração de *hardware* do servidor não ser apropriada para hospedar o NidoTat, executando com documentos grandes e com anotação colaborativa. Vale ressaltar que o servidor que hospeda o NidoTat é de propriedade do

CIC/UnB. Devido aos atrasos no repasse pela FAPDF, não foi possível adquirir um servidor de adequado para a aplicação das tarefas.

Como o processo de anotação estava prejudicando o progresso das atividades de pesquisa do KnEDLe, no início da *Release 3*, a equipe de anotação decidiu delegar a tarefa de anotação do *corpus* padrão ouro do DODF para voluntários externos ao projeto. Foram recrutados vários voluntários dentre estudantes do CIC e a equipe de anotação do KnEDLe ficou responsável por gerenciar e apoiar esse grupo de alunos. Foi usada a plataforma Microsoft Teams para esse acompanhamento. Um total de 23 voluntários se inscreveram para participar do processo de anotação, com o compromisso de contribuir durante o semestre letivo, que se encerrou no final de maio de 2021. Essa quantidade foi considerada satisfatória pela equipe de anotação do KnEDLe. Inicialmente, foram aleatoriamente atribuídos os atos pelos quais os anotadores voluntários seriam responsáveis para rotular os 67 documentos restantes. É importante notar que os membros do KnEDLe rotularam 33 documentos, do total de 100 documentos que devem constituir o corpus do DODF. Para definir os prazos de entrega das anotações, foram definidos 4 *batches* (lotes) de documentos, como mostra a Tabela 2.

Tabela 2: Descrição dos batches de DODFs construídos com trabalho dos voluntários de anotação de textos.

Nome do Batch	Quantidade de documentos	Início	Fim
Batch 1	21	03/03/2021	30/03/2021
Batch 2	20	31/03/2021	não finalizada
Batch 3	20	03/05/2021	não finalizada
Batch de Validação	6	13/04/2021	30/04/2021

Cada anotador voluntário recebeu aproximadamente entre 5 a 10 documentos em cada *batch*. A Tabela 2 também apresenta o período de anotação para cada um dos *batches*. Cada *batch* de documentos era carregado em projetos separados no NidoTat, sendo possível ter dois projetos em andamento simultaneamente, uma vez que eles são independentes entre si.

O *batch* 1 transcorreu sem problemas, sendo que a fase de rotulação demorou uma semana a mais do que o previsto devido às diversas dúvidas dos alunos ao identificar corretamente as entidades dos atos. No *batch* 2, a fase de rotulação também transcorreu normalmente, mas a presença de alguns documentos com uma imensa quantidade de anotações afetou o início da fase de revisão, pois o NidoTat apresentou problemas ao carregar as anotações para revisão. Com isso, a própria equipe de anotação do KnEDLe ficou responsável por finalizar a fase de revisão do batch 2. Assim, para não atrasar o processo de anotação com esses voluntários, o *batch* de validação foi iniciado, sendo de fundamental importância para avaliar a qualidade do *corpus* do DODF. Por último, o *batch* 3 foi iniciado, mas como restava pouco tempo para o final do semestre, apenas a fase de rotulação foi concluída, fazendo que a equipe de anotação do KnEDLe ficasse responsável pela revisão das anotações.

No final da *Release* 3, a equipe de anotação do KnEDLe concentrou seus esforços em analisar cuidadosamente todas as anotações criadas pelos voluntários externos ao projeto na disciplina, como também das anotações ainda incompletas realizadas na *Release* 2 pelos membros do KnEDLe. Espera-se que até o final de agosto, já no início da *Release* 4, o *corpus*

do DODF seja disponibilizado para todos os membros do projeto KnEDLe. Detalhes sobre as pesquisas relacionadas com as métricas de validação das anotações a nível de atos e entidades são fornecidas na Subseção 3.2.5.

3.1.3 VisNote

O VisNote foi inicialmente desenvolvido para explorar e anotar dados textuais de Diários Oficiais do Governo. Durante o terceiro semestre deste projeto, tal ferramenta foi adaptada para auxiliar a tarefa de revisar as anotações que foram feitas utilizando o NidoTat. Para esse propósito, uma das novas funcionalidades implementadas foi a extração dessas anotações em arquivos CSV, uma vez que o NidoTat armazena as anotações em arquivos XML. A página que oferece essa funcionalidade pode ser vista na Figura 4. Fora a possibilidade de inserção dos dados, a estrutura global do VisNote sofreu poucas alterações. Assim como em sua primeira versão, o usuário pode fazer o *download* do conjunto revisado de dados textuais posteriormente. Devido a necessidade de revisão das anotações a nível de relação e a nível de entidades, a parte majoritária das mudanças está concentrada nas ferramentas de interação com o usuário.

Figura 4: Página do VisNote responsável pela inserção das anotações a serem revisadas.

Atualmente, o usuário pode optar por visualizar as anotações referentes a todos os tipos de atos, ou selecionar um subconjunto de atos a serem revisados. O conteúdo presente nas tabelas e nas representações gráficas é atualizado conforme tal escolha. As guias posicionadas mais à esquerda, permitem que o usuário analise rapidamente os atributos chaves das anotações que estão sendo revisadas. Tais atributos são organizados em duas tabelas, uma a nível de entidade e outra a nível de entidades. O objetivo é possibilitar que o usuário

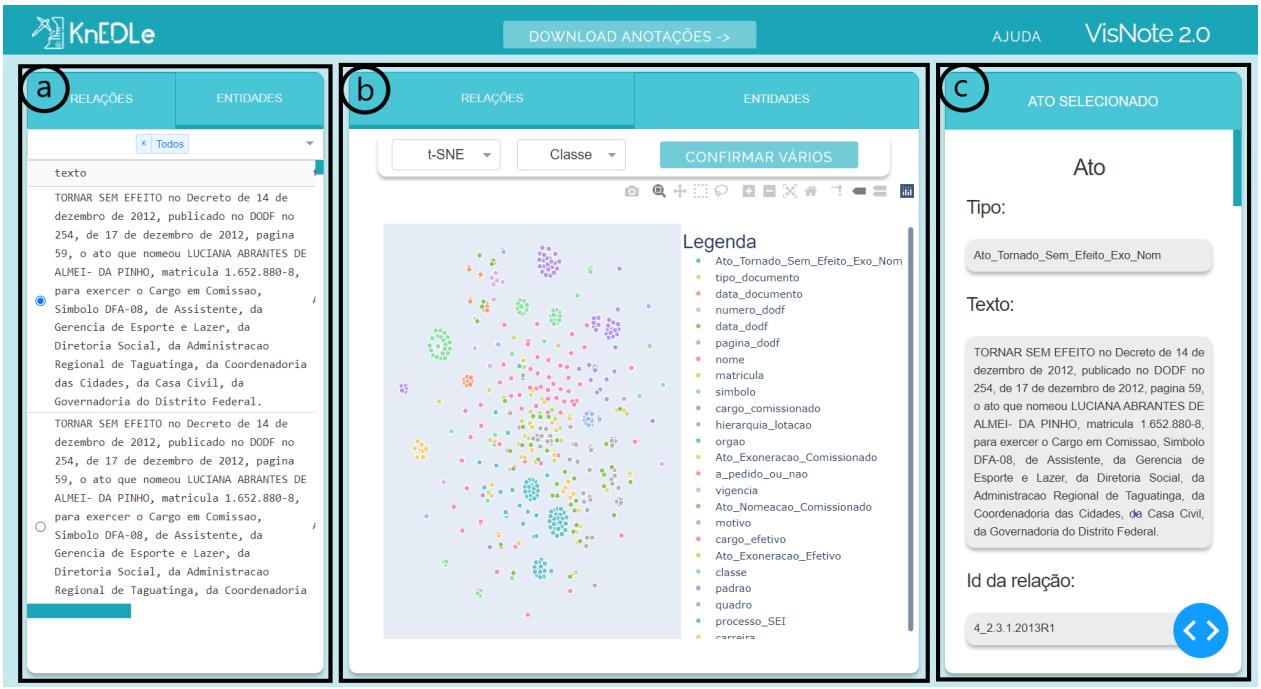
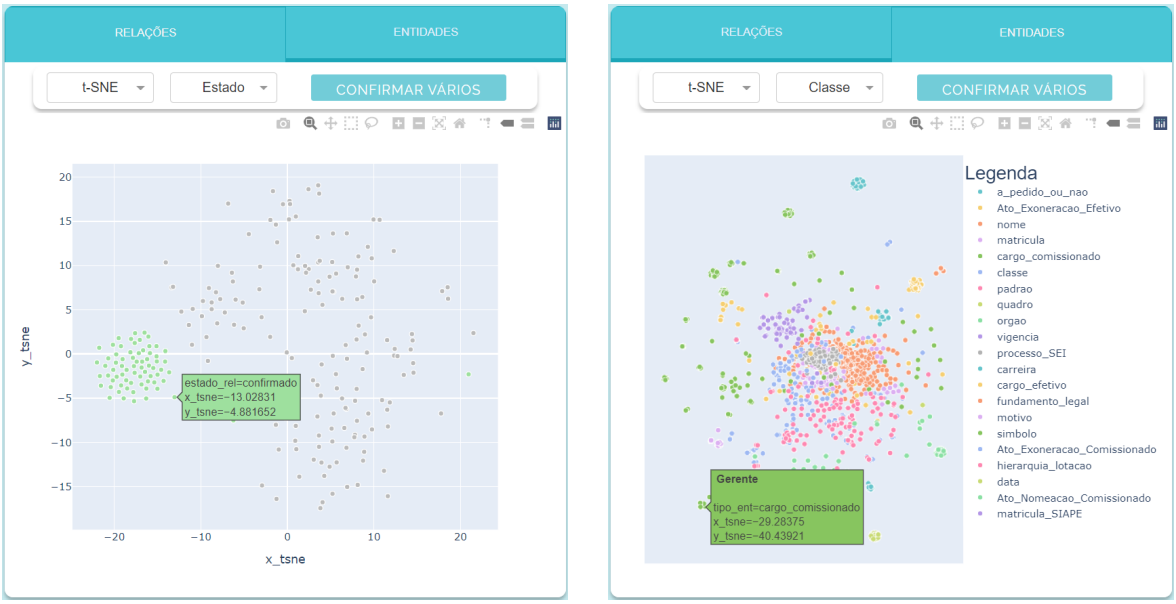


Figura 5: Página do VisNote responsável pela parte de revisão das anotações. (a) Tabelas com informações principais sobre as entidades e suas relações. (b) Representações gráficas das anotações a nível de relação e a nível de entidade, geradas a partir de redução de dimensionalidade. (c) Painel de controle para a análise mais detalhada das anotação e o contexto em que elas se encontram.

analise o conteúdo textual da anotação, o tipo que foi anotado e o seu estado atual sem precisar navegar por todos os pontos presentes nas representações gráficas. Ao selecionar uma determinada linha, se obtém mais informações sobre o ato apresentado no painel de controle.

Como pode ser visto na Figura 6, o VisNote possibilita a interação ativa com as representações gráficas tanto a nível de relações quanto a nível de entidades. Cada ponto representa uma anotação e sua coloração pode ser feita de acordo com os tipos que foram anotados ou de acordo com os estados de revisão em que se encontram as anotações. Ao clicar em um ponto, se obtém mais informações sobre o ato representado pelo mesmo através do painel de controle. Além disso, o usuário pode fazer uso da posição relativa dos pontos, uma vez que pontos posicionados aproximadamente podem compartilhar de padrões semelhantes, como exemplificado na Figura 7. No layout gerado para representar as entidades anotadas, tem-se a opção de selecionar visualizar as instâncias de apenas um tipo de entidades, funcionalidade esta que auxilia o usuário a navegar pelas grande quantidade de anotações a nível de entidades, a qual costuma ser até dez vezes maior do que a quantidades de anotações a nível de relações.

Outro recurso que visa tornar o processo de revisão uma tarefa menos custosa é a opção de selecionar várias anotações para serem confirmadas de uma só vez. Seu intuito é agilizar o processo de revisão, possibilitando que o usuário utilize o painel de controle apenas para a visualização do contexto no qual as anotações estão inseridas. Ao registrar as anotações de entidades que devem ser apagadas, corrigidas ou revisadas posteriormente, todas as demais



(a) (b)

Figura 6: Exemplos de projeções geradas pelo VisNote: (a) representação gráfica de anotações a nível de relação, colorida de acordo com o estado do processo de revisão; (b) representação gráfica a nível de entidade, colorida de acordo com a classe das anotações.

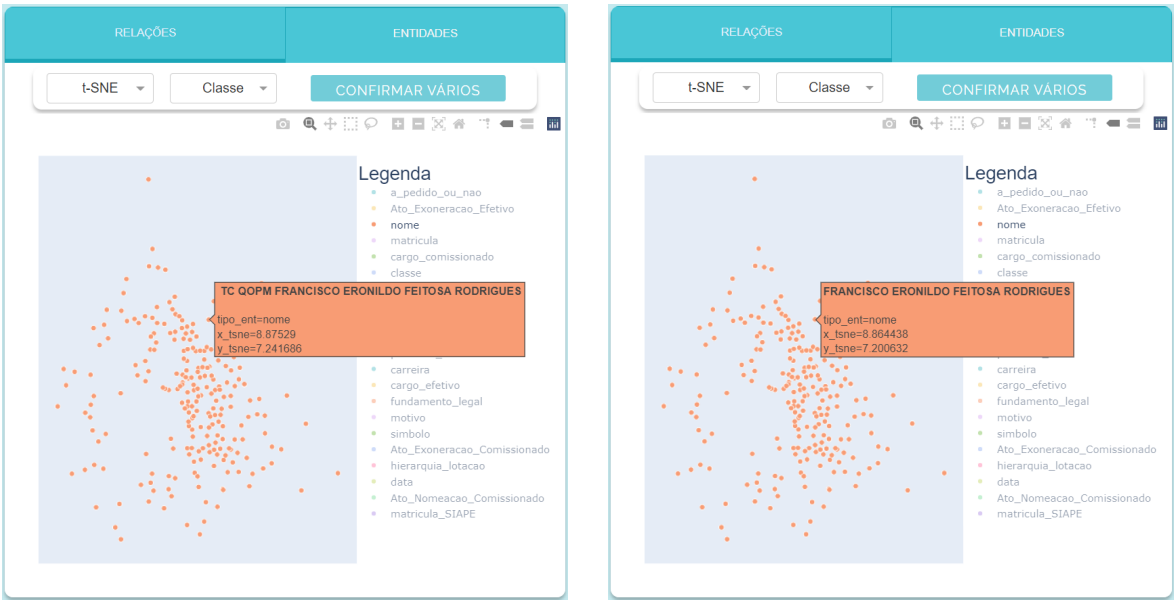


Figura 7: Exemplos de revisão de anotação utilizando a representação gráfica das anotações a nível de entidade.

anotações podem ser confirmadas com um único botão. Isso é possível ao se utilizar o botão "confirmar vários", presente tanto no layout das relações quanto no layout das entidades. Caso este método seja utilizado para confirmar as anotações a nível de relações, todas as

entidades que estavam no estado "nao_confirmado" e que pertencem às relações selecionadas também terão seu estado atualizado. As anotações de entidades que estiverem nos estados "em_duvida" ou "deletar", só podem ser alteradas por meio do painel de controle.

A principal função do painel de controle é disponibilizar ao usuário mais informações sobre o ato ligado à entidade ou relação que foram selecionadas. Primeiro, ele apresenta o tipo do ato que é representado pela relação em destaque, seguido pelo seu texto completo e número de identificação da relação. Abaixo encontra-se uma lista com todas as entidades que são ligadas pela relação em destaque, em blocos que contém o tipo da entidade, seu texto, e quatro botões que são utilizados para registrar qual categoria o usuário atribui às anotações sendo revisadas.

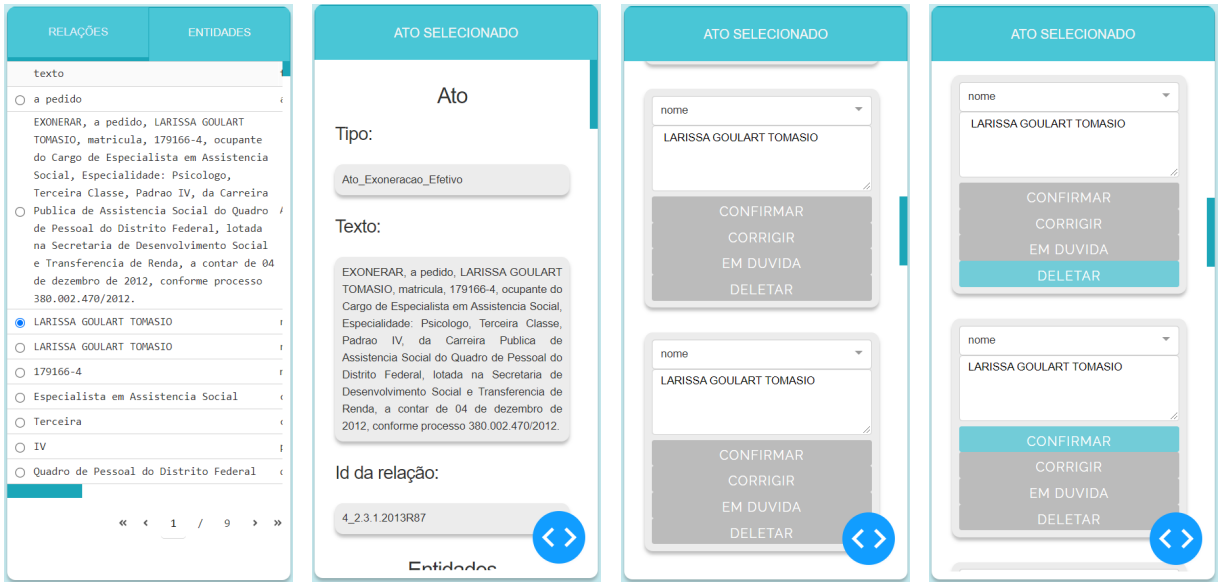
A fim de representar o estado em que se encontra determinada anotação, foram determinadas quatro categorias: "deletar", "em_dúvida", "corrigido", "confirmado" e "nao_confirmado". O estado inicial de todas as anotações é "nao_confirmado". Para registrar que uma anotação está conforme o necessário, o usuário deve alterar o estado da relação para "confirmado", clicando no botão "confirmar". Caso o tipo da entidade esteja errado, ou o texto requeira alguma correção, o usuário é capaz de realizar as modificações necessárias, tendo em vista que a parte que apresenta as informações referentes às entidades é editável. Posteriormente, basta selecionar a opção "corrigir" que as informações serão atualizadas conforme o desejado, e inseridas na categoria "corrigido". Por fim, o usuário pode optar por "deletar" uma anotação em casos de anotações duplicadas, anotações que estão fora do que deveria ter sido anotado ou demais situações adversas. Os estados em que se encontram todas as entidades e as relações às quais elas pertencem podem ser visualizados tanto no layouts quanto nas tabelas e no painel de controle, os quais são atualizados frequentemente. A Figura 8 demonstra como pode ser realizada a revisão de anotações utilizando a tabela de entidade e o painel de controle.

Os próximos passos do desenvolvimento do VisNote incluem experimentos visando a validação da qualidade das representações visuais geradas pela ferramenta e o quão fielmente elas retratam as relações de similaridade entre as anotações. Ademais, estão sendo conduzidos estudos relacionados a utilização de técnicas de aprendizado ativo para a sugestão automática de anotações a serem corrigidas.

3.1.4 KnEDLe Contratos

Foi feita uma revisão bibliográfica que partiu da escolha de trabalhos que utilizaram conjuntos de dados para a elaboração de uma linha do tempo, com o objetivo de visualizar as fases de um processo licitatório feito pelo GDF ao longo do tempo. Para a elaboração da ferramenta um dos focos principais foi o estudo de licitações públicas.

O primeiro grande desafio foi entender um processo licitatório. Esse entendimento foi obtido com o auxílio do documento público e fornecido pelo *GDF* – veja o documento citado em [3]. O objetivo desse documento é elucidar os principais fundamentos normativos relativos às licitações públicas. Com isso, foi possível analisar mais a fundo amostras de atos de licitações publicados no Diário Oficial do Distrito Federal. Foram identificadas as formas como esses atos são estruturados, os dados que cada tipo de ato apresenta, definindo padrões em como esses dados são apresentados e destacado alguns tipos de atos mais comuns nesses processos.



(a) (b) (c) (d)

Figura 8: Exemplos de revisão de anotação utilizando a tabela de entidade e o painel de controle: (a) identificação de anotações que parecem ser duplicadas; (b) informações sobre a relação a qual essas entidades pertencem; (c) informações sobre as entidades antes de alterar seus estados; (d) informações sobre as entidades depois de serem revisadas e terem seus estados alterados.

Após algumas análises minuciosas em cima de padrões, contamos com o auxílio de expressões regulares (*regex*) para a coleta de dados importantes com o foco em rastreamento, caracterização e identificação de cada ato, como o tipo do ato, o número do processo, a data da publicação no Diário Oficial, nome das empresas que participaram da licitação, empresas vencedoras, itens vencedores, itens fracassados e o conteúdo do próprio ato.

A grande maioria das licitações possui um determinado número de processo. A partir desse número é possível rastrear os atos no Diário Oficial do Distrito Federal que diz respeito a essa licitação, e acompanhá-la com o passar do tempo, além de analisar como ela se comporta até seu encerramento ou cancelamento. Com isso em mente foi possível começar a desenhar ideias de como poderíamos apresentar os dados coletados de uma forma clara e iterativa. A partir disso chegamos a conclusão de que uma linha do tempo (*timeline*) poderia apresentar esse rastreio e visualização da melhor forma.

SECRETARIA DE ESTADO DE EDUCAÇÃO

SUBSECRETARIA DE ADMINISTRAÇÃO GERAL
COMISSÃO PERMANENTE DE LICITAÇÃO

AVISO DE ABERTURA
PREGÃO ELETRÔNICO SRP Nº 06/2021 - (UASG 450432)

Objeto: Registro de preços para eventual aquisição de gêneros alimentícios perecíveis - Filé de Peixe Congelado de TILÁPIA-, para o Programa de Alimentação Escolar do Distrito Federal (PAE-DF) conforme condições, quantidades e exigências estabelecidas no Edital e seus Anexos. Total de Itens: 08- Valor total estimado: R$ 19.900.737,72 (dezenove milhões, novecentos mil setecentos e trinta e sete reais e setenta e dois centavos). PROCESSO Nº 00080-00198062/2020-14-Cadastro das Propostas: a partir de 12/01/2021. Abertura das Propostas: 22/01/2021, às 10h, horário de Brasília. O Edital estará disponível nos endereços eletrônicos: www.comprasgovernamentais.gov.br e http://www.se.df.gov.br/pregao-eletronico-sistema-de-registro-de-precos/.

Brasília/DF, 11 de janeiro de 2021
REGINA RODRIGUES PORTO
Pregoeira

Figura 9: Exemplo de um aviso de abertura, ato bastante comum, onde o objetivo é informar o surgimento de um processo licitatório

Figura 10: Visualização de um ato na timeline

Até o presente momento, estamos rastreando 16 atos distintos, são eles:

- Aviso de Abertura de Licitação,
- Aviso de Homologação e Adjudicação,
- Aviso de Homologação e Convocação,
- Aviso de Resultado,

- Aviso do Resultado de Julgamento,
- Aviso de Declaração de Vencedor,
- Aviso de Julgamento,
- Aviso de Julgamento de Habilitação,
- Aviso de Julgamento de Recurso Administrativo,
- Aviso de Suspensão de Licitação,
- Aviso de Adiamento de Licitação,
- Aviso de Reabertura,
- Aviso de Licitação,
- Extratos de Contrato,
- Aviso de cancelamento e
- Aviso de Alteração.

3.2 Pesquisa

Esta seção resume os principais avanços do projeto em termos de pesquisas que exploram técnicas visando propor algo que avance o estado-da-arte nas áreas relacionadas a este projeto. Algumas das seções abaixo descrevem estudos que foram realizados para dar suporte às escolhas feitas no desenvolvimento dos produtos descritos nas seções anteriores. Outras seções descrevem estudos ou experimentos cujos resultados poderão futuramente ser transferidos a produtos deste projeto.

3.2.1 Artigo científico: *Deep Active-Self Learning Applied to Named Entity Recognition*

Esta seção relata os resultados do aluno de mestrado José Reinaldo da Cunha. Em especial, é descrito brevemente o artigo submetido e aceito para *10th Brazilian Conference on Intelligent System* (BRACIS 2021). O trabalho condiz com experimentos para o problema NER utilizando aprendizado ativo com *self-training* em modelos de aprendizagem profunda.

Aprendizagem profunda tem sido o estado da arte para uma variedade de tarefas desafiadoras em processamento de linguagem natural. Porém, para alcançar bons resultados é necessário uma grande quantidade de dados rotulados. Técnicas de *Deep active learning* foram projetadas para reduzir a quantidade de dados anotados para o treinamento de tais modelos. Entretanto, essas técnicas não apresentam resultados satisfatórios para problemas que exigem um completo conjunto de treinamento. Como solução para este problema, foi investigado técnicas baseadas *Active-self learning* que empregam auto rotulagem usando o modelo já treinado para ajudar a aliviar o custo da anotação de todo o conjunto de dados em tarefa de reconhecimento de entidades nomeadas (NER).

Como parte do trabalho realizado pelo aluno, foram feitos experimentos que indicaram que a proposta foi capaz de reduzir a anotação manual. Também foram investigadas técnicas de parada antecipada que não dependem do conjunto de validação, que efetivamente reduz ainda mais o custo de anotação.

3.2.2 Análise de documentos usando informações textuais e visuais

Documentos são constituídos por elementos estruturais dispostos tanto espacial quanto hierarquicamente seguindo algum *layout* para facilitar compreensão pelos humanos. Análise de documentos combinando informações textuais e visuais é um tópico recente mas já bastante ativo. A ideia é combinar técnicas de processamento de linguagem natural com visão computacional, agregando análise textural com análise da disposição da página (layout). A aluna Patrícia Drumond (doutoranda) tem focado sua pesquisa nessa área, com a realização de uma revisão bibliográfica e de experimentos preliminares para reproduzir resultados de artigos considerados como o estado-da-arte. O foco desse trabalho tem sido na elaboração de modelos de linguagem com contexto visual. Isso pode ser usado para diversas aplicações pertinentes a este projeto, desde reconhecimento de entidades nomeadas à classificação de seções. Para a análise de layout pode ser utilizado um método de segmentação da imagem de páginas, separando seus elementos estruturais e classificando-os, por exemplo, em título de seção, parágrafos, figuras, tabelas, expressões matemática, carimbos, assinaturas, etc. Técnicas baseadas em redes neurais profundas (deep learning) estão sendo avaliadas para a extração de informações tanto textuais quanto visuais.

Além desse trabalho preliminar o projeto já conta com resultados em classificação de páginas com fusão de informação textual e visual. O bolsista Pedro H. Luz de Araujo (mestrando) liderou a pesquisa e a escrita do artigo intitulado *Sequence-aware multimodal page classification of Brazilian legal documents*, que apresenta exaustivos experimentos numa grande base de dados. Seus resultados mostram que a combinação dessas três fontes de informação: texto, imagem e informação de sequencia de páginas. Esse artigo foi submetido para o *International Journal on Document Analysis and Recognition* e está sendo revisado.

Além disso, o bolsista Pedro H. Luz de Araujo concluiu a escrita da sua dissertação de mestrado, a qual já foi defendida (vide [12]). A íntegra do texto, bem como todas as publicações relacionadas, vídeos de apresentação, bases de dados criadas, programas (código fonte) e parâmetros, estão todos disponíveis a partir desta página: `http://cic.unb.br/~teodecampos/peluz/`.

3.2.3 Suporte Visual para Tarefas de Classificação e Anotação

As técnicas de visualização interativa implementadas no VisNote foram inspiradas em trabalhos como o ATR-Vis [13], o qual apresenta uma abordagem visual orientada ao usuário para a recuperação de conteúdo do Twitter. Os dados extraídos são posteriormente utilizados em outras tarefas de análise de dados. Para que a ferramenta seja acessível e utilizável por pessoas leigas, o processo de extração de informação e as estratégias relacionadas a ele foram integradas a uma interface visual. Os autores concluem que a interface interativa fornece as ferramentas necessárias para a obtenção de uma coleção de dados extraídos do Twitter confiável, e que responde a necessidades de informação específicas, por usuários não

especialistas.

Heimerl et al. [6] comparam três modelos de anotação de dados textuais, todos explorando o aprendizado ativo de forma direta ou indireta como parte de um processo de anotação visual interativa semi-guiada. O Método Básico é bastante direto, fornecendo uma interface baseada em texto, enquanto os métodos Visual e Orientado ao Usuário fornecem acesso visual interativo ao estado do classificador. O terceiro oferece graus de liberdade adicionais em relação à seleção dos documentos a serem anotados.

Ao analisar os resultados da avaliação quantitativa, nota-se que os participantes tendem a ser mais rápidos ao usar o Método Básico. Por outro lado, o Método Orientado ao Usuário manifestou uma necessidade inicial de se dedicar mais tempo para o treinamento do usuário. Os comentários realizados pelos usuários também indicam que analisar e anotar um número considerável de documentos sequencialmente, conforme exigido pelo Método Básico, é cansativo e os participantes ficam entediados rapidamente. Por fim, afirma-se que obter um *feedback* visual é o preferido pela maioria dos participantes. Pode-se notar que embora o suporte visual alivie a carga do processo de anotação, ele deve ser intuitivo o suficiente para evitar o efeito contrário ao que se deseja.

Como parte de uma colaboração com pesquisadores do projeto StaViCTA [5], Kucher et al. [10] introduziram uma abordagem para suporte visual de anotação de texto e classificação no contexto de análise de posicionamento de dados textuais. Seu sistema, denominado ALVA, inclui a anotação de várias categorias de postura, um classificador com uma abordagem de aprendizagem ativa e várias interfaces visuais para auxiliar na análise visual exploratória dos dados anotados. As visualizações permitem que usuário obtenha uma visão geral do conjunto de dados, análise a ocorrência de múltiplas categorias, identifique casos interessantes relacionados que pertencem a várias categorias e compare as anotações feitas para os mesmos enunciados. Os autores também introduziram uma nova representação visual, chamada CatCombos, a qual foi projetada para representar os registros de anotações individuais e combinações de categorias de postura.

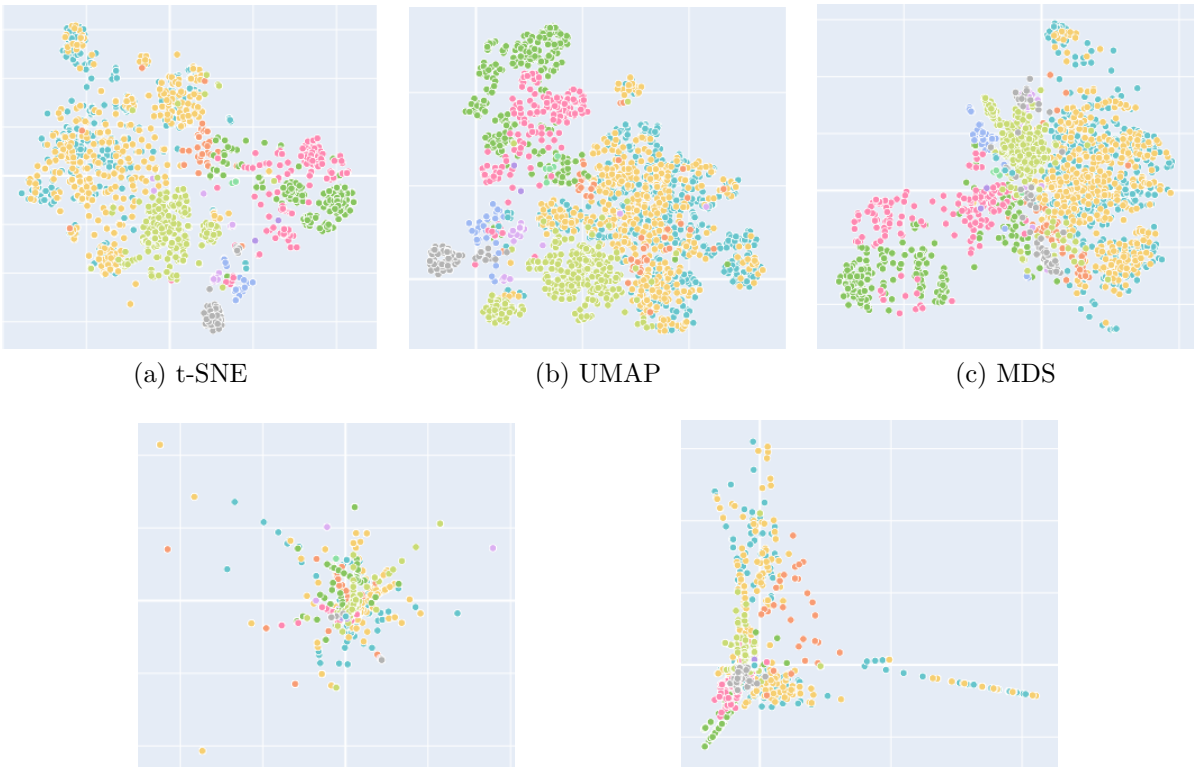
3.2.4 Projeções Multidimensionais

Os vetores de contagem de palavras utilizados para a representação de dados textuais, como por exemplo o *Term Frequency-Inverse Document Frequency* (TF-IDF) [7], costumam possuir centenas de dimensões. A fim de se gerar representações de baixa dimensão a partir deles, pode-se utilizar projeções multidimensionais. Também conhecida por redução de dimensionalidade, tal técnica busca representar dados multidimensionais em espaços de baixa dimensão, ao mesmo tempo em que mantém as partes mais relevantes de sua estrutura [19]. Até o momento, a ferramenta VisNote incorpora visualizações baseadas nas técnicas *t-Stochastic Distributed Neighbor Embedding* (t-SNE) [22] e *Uniform Manifold Approximation and Projection* (UMAP) [11], as quais reduzem a dimensionalidade das representações TF-IDF para espaços bidimensionais, permitindo assim a geração dos layouts 2D.

Nonato et al. [14] buscaram reunir os principais aspectos que devem ser levados em consideração ao integrar projeções multidimensionais (MDP) a ferramentas de análise visual em sua pesquisa. Como fruto de seu estudo, são fornecidas análises e taxonomias detalhadas

[5] `https://cs.lnu.se/stavicta/`

quanto à organização de técnicas de MDP, junto a discussões sobre suas influências em propriedades de percepção visual e outros fatores humanos. Além disso, aborda-se os diferentes tipos de distorções que podem resultar das projeções e seu impacto em diferentes tarefas analíticas realizadas na exploração de dados multidimensionais. Também são apresentadas técnicas de enriquecimento de layout para mitigar os efeitos das distorções do MDP. Ressalta-se a importância de fornecer informações relevantes a fim de auxiliar o usuário na tomada de decisões, de tal forma que não se dependa inteiramente do que é exposto nas visualizações.



(a) t-SNE (b) UMAP (c) MDS

(d) LLE (e) ISOmap

Figura 11: Exemplos de projeções obtidas a partir da implementação de diferentes técnicas em um conjunto de dados textuais extraídos de documentos do DODF.

Com o intuito de avaliar de forma quantitativa as distorções das técnicas de projeção multidimensional implementadas no VisNote, utilizou-se um conjunto de dados que consiste em 1824 atos pertencentes a 12 classes distintas. Eles foram obtidos a partir de cinco documentos do DODF anotados por membros do projeto KnEDLe. Ademais, a fim de comparar a qualidade das representações gráficas obtidas pelos métodos t-SNE e UMAP, foram implementadas mais três técnicas de projeção multidimensional em nosso conjunto de dados: *Isometric Feature Mapping* (Isomap) [20], *Locally Linear Embedding* (LLE) [18] e *Multidimensional Scaling* (MDS) [21]. O LLE busca realizar uma projeção que preserve as distâncias entre as instâncias a nível local. O MDS visa uma representação de baixa dimensão dos dados que respeite as distâncias do espaço de alta dimensão original. Por fim, o Isomap visa projetar um espaço de baixa dimensão que mantenha as distâncias geodésicas entre todos os pontos. Com isso, foram geradas cinco representações gráficas distintas, conforme apresen-

tado na Figura 11. Cada ponto representa um ato, e sua cor representa a respectiva classe do ato.

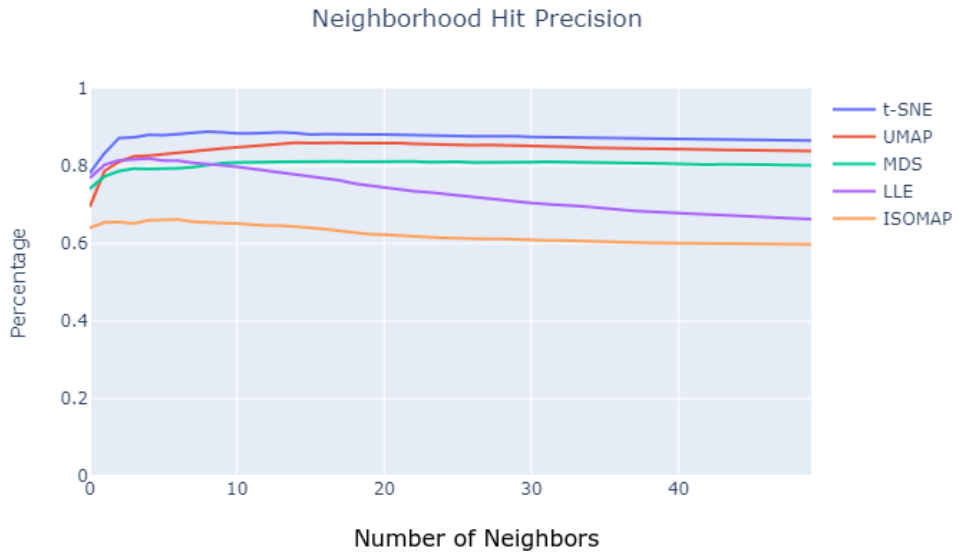


Figura 12: *Neighborhood Hit Precision* das projeções obtidas a partir de diferentes técnicas no mesmo conjunto de dados.

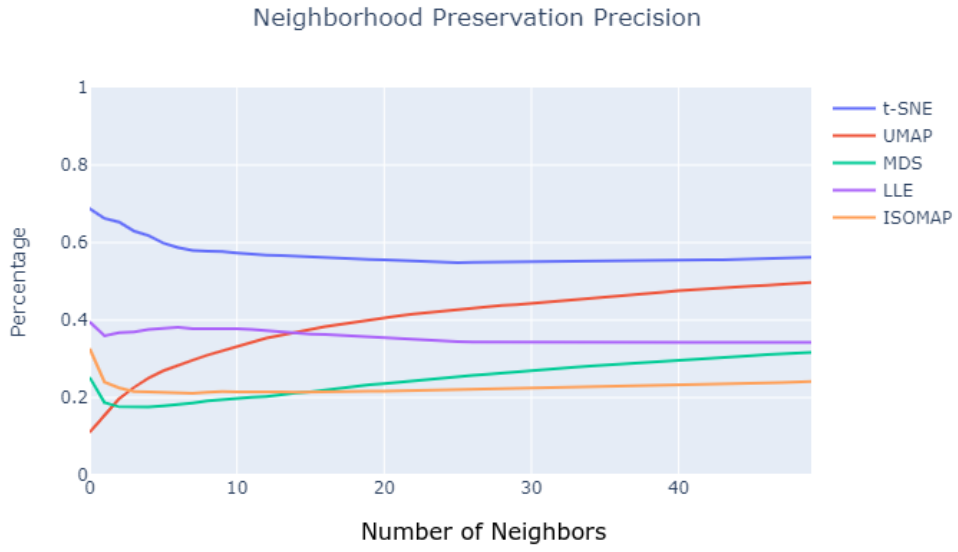


Figura 13: *Neighborhood Preservation Precision* das projeções obtidas a partir de diferentes técnicas no mesmo conjunto de dados.

As medidas *Neighborhood Hit* [17] e *Neighborhood Preservation* [16] foram implementadas

para avaliar cada projeção apresentada na Figura 11. Tais métricas são uma fácil avaliação das qualidades das representações gráficas, com foco no objetivo visual das projeções, que é ter instâncias semelhantes posicionais próximas umas das outras. Para ambas as medidas, é necessário procurar pelos n vizinhos mais próximos de cada instância do conjunto de dados, tanto nos espaços de alta dimensão quanto nos espaços de baixa dimensão. *Neighborhood Hit* mede a porcentagem de instâncias de mesma classe que foram preservadas quando projetadas do espaço de alta dimensão para o espaço de baixa dimensão. Enquanto *Neighborhood Preservation* tem conceito semelhante mas verifica a preservação dos vizinhos em si, independente da classe. Conforme mostrado nas Figuras 12 e 13, ambas as medidas indicam que o t-SNE é a técnica que melhor preserva as relações de similaridade entre os documentos do DODF ao projetá-los, acompanhado estreitamente da técnica UMAP.

3.2.5 Critérios de Avaliação do Corpus KnEDLe

A motivação em se construir um conjunto de dados em um padrão ouro [23] é fornecer dados rotulados confiáveis para os modelos de aprendizado de máquina e aprendizado profundo em consideração no âmbito do projeto KnEDLe. Nesse sentido, para verificar a qualidade as anotações dos textos e das entidades, é necessário um procedimento objetivo de avaliação.

Os dados coletados dos documentos do DODF se caracterizam por terem estrutura complexa, com muitos rótulos e anotações aninhadas. Esse cenário aponta para a necessidade de um critério elaborado de avaliação de similaridade entre anotações. Especificamente, a estrutura dos textos indica critérios de concordância inter-anotador com relação aos rótulos e às sequências de textos.

Ontañón et al. [15] descrevem diversas funções para o cálculo da similaridade e da distância em relação a dados estruturados. Dentre elas, as escolhas imediatas para sequências de textos são aquelas que se baseiam em representações vetoriais, como *distância Cosseno* e *distância de Minkowski*. Outra possibilidade consiste em empregar o *índice de Jaccard*, proposto para comparação de conjuntos matemáticos e usualmente utilizado em dados de imagens, mas com adaptações bem sucedidas para dados textuais [4, 1].

Considerando as sequências de textos, um outro aspecto importante se trata da avaliação com respeito aos rótulos assinalados em cada anotação. Existem algumas métricas bem estabelecidas, descritas por Krippendorff [9]. Os pontos positivos apresentados são o *critério Kappa*, de Cohen, o qual define a chance da concordância observada entre dois anotadores se seus comportamentos são estatisticamente não-relacionados, e o *critério Alpha*, de Krippendorff, que é uma medida multi-propósito de confiabilidade usada em diferentes domínios.

Outra estratégia válida se pauta em fazer rodadas de anotação específicas para a validação. Os dados obtidos de rodadas de validação são então comparados com os dados originalmente anotados usando alguma métrica adequada. Esse procedimento já foi usado com sucesso em domínios que lidam com aprendizado de máquina [2]. Ademais, pode-se citar uma avaliação extrínseca dos resultados de tarefas de Processamento de Linguagem Natural feitas junto aos dados coletados [5, 8].

3.2.6 UI/UX para *KnEDLe Contratos*

User eXperience (UX) trata da relação entre uma pessoa/usuário e um determinado pro-

duto/serviço e contempla desde o interesse, as pesquisas e a compra de um produto, pelo usuário. Dessa forma, o principal papel do *UX Designer* é se preocupar com cada etapa com a qual o usuário interage com o produto ou serviço; e fazer com que essa interação ocorra da forma mais tranquila possível.

Já *UI* significa *User Interface* ou Interface do Usuário em português. Nesse sentido, representa tudo aquilo que é utilizado na interação com um produto, sendo a intermediária visual entre o homem e a máquina. Dessa forma, o *UI Designer* é responsável principalmente pela criação de interfaces funcionais, as quais permitem que usuário navegue intuitivamente por toda sua jornada.

O perfil do usuário para a elaboração da ferramenta *KnEDLe Contracts* foi traçado na parte inicial do projeto – o objetivo é auxiliar o governo na auditoria de licitações públicas.

Em relação à interface, foi utilizado o conceito de prototipação de alta fidelidade do mundo do UI. Um protótipo de alta fidelidade deve se aproximar ao máximo dos aspectos visuais e funcionais do produto final, incluindo o conteúdo, fluxo de navegação e interações.

A apresentação das interfaces do módulo do KnEDLe Contratos está ilustrada na Figuras 14. Nas figuras 15, 16, 17 estão alguns detalhes que foram definidos no projeto da interface da ferramenta. Esses detalhes estão relacionados a definição da tipografia, ícones, botões e caixas de entradas (*inputs*). Maiores detalhes da interface e a arquitetura da ferramenta visual estão descritos na seguinte página: `https://unb-knedle.github.io/timeline-contratos/`

4 Considerações finais

Neste período, o projeto avançou em diversas áreas, principalmente no que se trata da aquisição de uma grande quantidade de dados detalhadamente rotulados, graças à participação de diversos colaboradores, sob orientação dos bolsistas da equipe de anotação de dados. O projeto tem atingido suas metas, apesar de adversidades causadas pela pandemia de CoViD e pelo grande atraso nos repasses de recursos. Ao final do referido período, houve um repasse integral e a equipe finalmente foi constituída de acordo com o número de colaboradores previsto no plano de trabalho e alguns dos equipamentos computacionais foram adquiridos. Com isso, espera-se que haja um salto na produtividade do projeto no próximo semestre, tanto no desenvolvimento de pesquisa quanto nos produtos.

Figura 14: Protótipo final, temos a página *Home*, *Timeline* e *Not Found*.

Figura 15: Tipografia definida para o projeto, trata-se da familia de fontes *Montsserrat*.

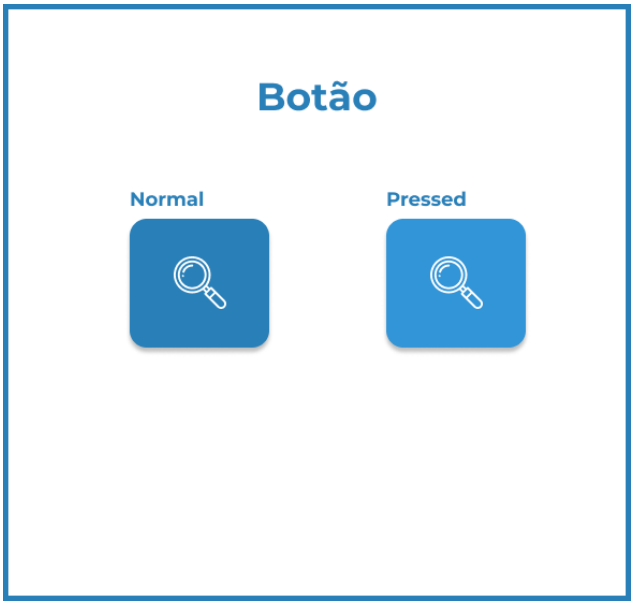


Figura 16: Botão responsável por mandar um número de processo para o *backend*.

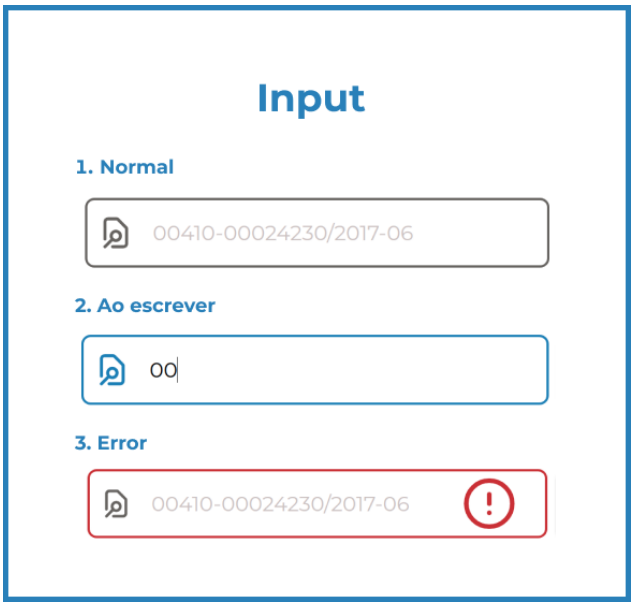


Figura 17: Os 3 estados do *input*, quando há erro, é exibido um *tooltip*.

Referências

[1] M. Afzal, F. Alam, K. Malik, and G. Malik. Clinical context–aware biomedical text summarization using deep neural network: Model development and validation. *Journal of Medical Internet Research*, 22, 2020.

[2] Victoria Bobicev and Marina Sokolova. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 97–102, Varna, Bulgaria, September 2017. INCOMA Ltd.

[3] Lucas Rocha Furtado. Questões prático-operacionais de licitações públicas para servidores.

[4] Debasis Ganguly, Dipasree Pal, Manisha Verma, and Procheta Sen. Overview of rcd-2020, the fire-2020 track on retrieval from conversational dialogues. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 33–36, New York, NY, USA, 2020. Association for Computing Machinery.

[5] Nathan S. Hartmann, Erick R. Fonseca, Christopher D. Shulby, Marcos V. Treviso, Jéssica S. Rodrigues, and Sandra M. Aluísio. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 122–131, Porto Alegre, RS, Brasil, 2017. SBC.

[6] Florian Heimerl, Steffen Koch, Harald Bosch, and Thomas Ertl. Visual classifier training for text document retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2839–2848, 2012.

[7] L. Jing, H. Huang, and Hong bo Shi. Improved feature selection approach tfidf in text mining. *Proceedings. International Conference on Machine Learning and Cybernetics*, 2:944–946 vol.2, 2002.

[8] Vojtěch Kovář, Miloš Jakubíček, and Aleš Horák. On evaluation of natural language processing tasks - is gold standard evaluation methodology a good solution? In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, pages 540–545, Rome, Italy, 02 2016.

[9] klaus krippendorff. Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5:93–112, 04 2011.

[10] Kostiantyn Kucher, Carita Paradis, Magnus Sahlgren, and Andreas Kerren. Active learning and visual analytics for stance classification with alva. *ACM Trans. Interact. Intell. Syst.*, 7(3), October 2017.

[11] John Healy Leland McInnes and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426v3*, 2020.

[12] Pedro Henrique Luz de Araujo. Domain-specific datasets for document classification and named entity recognition. Master's thesis, Universidade de Brasilia, July 2021. Related resources available from `https://cic.unb.br/~teodecampos/peluz/`.

[13] Raheleh Makki, Eder Carvalho, Axel J. Soto, Stephen Brooks, Maria Cristina Ferreira De Oliveira, Evangelos Milios, and Rosane Minghim. Atr-vis: Visual and interactive information retrieval for parliamentary discussions in twitter. *ACM Trans. Knowl. Discov. Data*, 12(1), February 2018.

[14] L. G. Nonato and M. Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2650–2673, 2019.

[15] Santiago Ontañón. An overview of distance and similarity functions for structured data. *Artificial Intelligence Review*, 53(7):5309–5351, 2020.

[16] Fernando V Paulovich and Rosane Minghim. Hipp: a novel hierarchical point placement strategy and its application to the exploration of document collections. *IEEE transactions on visualization and computer graphics*, 14(6):1229–36, 2008.

[17] Fernando V. Paulovich, Luis G. Nonato, Rosane Minghim, and Haim Levkowitz. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, 14(3):564–575, May 2008.

[18] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[19] Dominik Sacha, Leishi Zhang, Michael Sedlmair, John A. Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C. North, and Daniel A. Keim. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):241–250, January 2017.

[20] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[21] Jengnan Tzeng, Henry Horng-Shing Lu, and Wen-Hsiung Li. Multidimensional scaling for large genomic data sets. *BMC Bioinformatics*, 9(179), 2008.

[22] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[23] Lars Wissler, Mohammed Almashraee, Dagmar Monett, and Adrian Paschke. The gold standard in corpus annotation. In *5th IEEE Germany Student Conference*, Passau, Germany, 06 2014.