XIV ENCONTRO ANUAL DE TECNOLOGIA DA INFORMAÇÃO
ANAIS DO XIV ENCONTRO ANUAL DE TECNOLOGIA DA INFORMAÇÃO - EATI
| ANO 12, N. 1 | NOVEMBRO / 2023

Desenvolvimento de uma Ferramenta para Reconhecimento de Entidades Nomeadas em Certificados de Atividades Complementares de Curso utilizando spaCy

Bernardo Gularte Kirsch[1], Árton Pereira Dorneles[1]

[1] Curso de Bacharelado em Ciência da Computação do Instituto Federal de Educação, Ciência e Tecnologia Farroupilha (IFFar) – Frederico Westphalen – RS – Brasil

bernardogulartekirsch@gmail.com, arton.dorneles@iffarroupilha.edu.br

Abstract. *This work proposes the development of a Python tool to recognize entities named in certificates for complementary course activities. The tool was implemented with the help of the spaCy library and evaluated using a corpus of certificate data from courses in the IFFar/FW information and communication axis. After conducting computational experiments, the results obtained demonstrate through metrics that the proposed tool is promising for recognizing names, titles, periods and workload of a certain class of certificates for complementary course activities.*

Resumo. *Este trabalho propõe o desenvolvimento de uma ferramenta em Python para realizar o reconhecimento de entidades nomeadas de certificados de atividades complementares de curso. A ferramenta foi implementada com o auxílio da biblioteca spaCy e avaliada por meio de um corpus de dados de certificados provenientes dos cursos do eixo de informação e comunicação do IFFar/FW. Após a condução de experimentos computacionais, os resultados obtidos demonstram por meio de métricas que a ferramenta proposta é promissora para reconhecer nomes, títulos, períodos e carga horária de uma determinada classe de certificados de atividades complementares de curso.*

1. Introdução

Com o avanço da tecnologia e da internet, a quantidade de dados produzidos e armazenados nas mais diversas aplicações tem crescido significativamente nos últimos anos, envolvendo principalmente as mídias de texto, áudio, imagem e vídeo. Além do armazenamento desses formatos se constituir em um desafio por si só, ainda existe o desafio de processá-los para extrair informações relevantes para cada aplicação como, por exemplo, a identificação de pessoas, organizações, locais, data e hora presentes nos dados brutos de cada mídia. Esse processo de extração, quando realizado de forma manual por um operador humano, se torna oneroso e repetitivo, exigindo um significativo investimento de tempo, além de potencializar a ocorrência de erros.

Como uma proposta de solução para esse problema de extração de dados, foi apresentado em 1996, na 6ª edição do evento Message Understanding Conference (MUC-6), uma técnica de Processamento de Linguagem Natural denominada Reconhecimento de Entidades Nomeadas - REN (do inglês, Named Entity Recognition - NER) que, na ocasião, tinha como exemplo de aplicação principal a extração automática de informações relevantes em mensagens militares. Desde então, esta técnica tem sido utilizada para extração de dados nas mais diversas áreas. O processo de extração de dados utilizando REN consiste em uma implementação de software com o objetivo de identificar um conjunto específico de entidades

no texto como, por exemplo, nomes de pessoas, organizações, locais, data, hora, entre outras informações de interesse de cada aplicação. Para essa implementação, normalmente é utilizado um *corpus* de dados particular da aplicação para treinamento da ferramenta, buscando maior êxito e precisão no reconhecimento das entidades de interesse. Uma vez que a ferramenta tenha sido treinada, é possível utilizá-la para obter maior agilidade no processo de extração, otimizando o uso de tempo e de recursos de uma determinada aplicação.

Na literatura científica são encontrados vários trabalhos utilizando REN que buscam a extração automática de informações em diversos formatos de arquivos, mas, em particular, os arquivos do tipo PDF (*Portable Document Format*) se destacam como um dos principais formatos de interesse para extração de informações por ser muito utilizado para intercâmbio e disseminação de diversas fontes de informações na internet, como por exemplo, registros do Diário Oficial da União, documentos jurídicos, editais, notas fiscais eletrônicas, textos legislativos e documentos de coleta de dados como formulários e questionários. Para todas essas fontes é possível encontrar propostas correspondentes para extração de informações utilizando REN. Arquivos no formato PDF também são muito utilizados para criar documentos de certificados utilizados para comprovar, por exemplo, a conclusão de um curso de curta duração ou participação em um evento científico. Esse tipo de certificado é bastante utilizado para validação de atividades complementares de cursos de instituições de ensino no Brasil e, até onde se sabe, não existe nenhuma proposta de ferramenta na literatura com o objetivo de extrair informações de forma automatizada em certificados armazenados em PDF.

Nesse contexto, este trabalho propõe uma ferramenta desenvolvida em Python para realizar o reconhecimento de entidades nomeadas de certificados de atividades complementares de curso utilizando como base de treinamento um *corpus* de dados de certificados provenientes dos cursos do eixo de informação e comunicação do IFFar/FW. Desta forma, espera-se que o desenvolvimento deste projeto possa contribuir não somente com o desenvolvimento científico na área de computação, mas também em fornecer uma ferramenta que possa ser utilizada para agilizar, facilitar e minimizar erros de entrada de dados em sistemas informatizados para validação e gerenciamento de atividades complementares de curso.

O restante do trabalho está organizado como segue. Na Seção 2 é apresentado o referencial teórico do trabalho. Na Seção 3, a metodologia do projeto da ferramenta proposta é detalhada. A Seção 4 apresenta experimentos computacionais e resultados, e por fim, na Seção 5, são apresentadas considerações finais e opções de trabalhos futuros.

2. Referencial Teórico

Nesta seção são apresentados os principais conceitos e tecnologias necessárias para a compreensão deste trabalho, bem como um conjunto de trabalhos relacionados.

2.1. Reconhecimento de Entidades Nomeadas (REN)

De acordo com SILVA (2020), o Reconhecimento de Entidades Nomeadas (REN) é uma técnica de Processamento de Linguagem Natural (PLN) que surgiu em 1996, na 6ª edição do evento Message Understanding Conference (MUC-6), e teve seu foco na extração automática de informações em mensagens militares. Entidades nomeadas são artefatos de texto presentes em documentos descritos em linguagem natural. O processo de Reconhecimento de Entidades Nomeadas busca identificar e classificar esses artefatos de acordo com as necessidades de

cada aplicação. Um sistema de REN recebe como entrada um texto livre, não estruturado, e devolve como saída um conjunto de textos estruturados destacando as entidades de interesse.

As entidades nomeadas podem ser classificadas em categorias pré-definidas como Pessoa, Organização, Local, Data, Moeda, entre outras, esse tipo de entidade é chamado de entidade pré-construída (*Prebuilt Entity*). As entidades podem ter classificações específicas de uma determinada aplicação, mas para que isso seja possível é necessário que um algoritmo de aprendizado de máquina seja aplicado para treinamento prévio destas entidades.

Em resumo, um sistema de REN possui quatro etapas: pré-processamento, identificação de palavras pertinentes, classificação e pós-processamento. Na primeira etapa, a etapa de pré-processamento, o texto é corrigido e preparado para análise. Na etapa de identificação de palavras pertinentes, as palavras candidatas a entidades são identificadas. Na etapa de classificação, as palavras candidatas são classificadas em suas respectivas categorias e por fim, na última etapa, a etapa de pós-processamento, é onde as entidades são refinadas e, se necessário, agrupadas em entidades compostas ou relacionadas.

Atualmente, existem várias bibliotecas e sistemas de REN de uso geral que estão disponíveis para extração de dados em língua portuguesa e fornecem interfaces de programação em diversas linguagens de programação. Na linguagem Python, existe o sistema NERP-CRF, que possui código aberto e utiliza aprendizado de máquina. Em Java, foi proposto o LanguageTasks que é livre para fins acadêmicos e pago para fins comerciais. Este é capaz de reconhecer e classificar entidades por meio de um ambiente web. Outro sistema implementado em Java é o Apache OpenNLP que suporta tokenização e análise sintática. Desenvolvido em C++, o sistema FreeLing também possui código aberto e tem como foco o reconhecimento em textos na língua portuguesa, assim como o sistema 'Palavras' (FONSECA et. al, 2015). Outro sistema que fornece uma interface de programação em Python é o spaCy. Trata-se de uma biblioteca de código aberto mas com foco no uso em ambientes de produção. Ela foi desenvolvida em Cython e possui um bom desempenho em tarefas de extração de informações de textos em grande escala. Ela possui diversos *pipelines* treinados para diferentes idiomas, incluindo o Português, possui um sistema de treinamento pronto para produção, tokenização, componentes para REN com fácil extensão e personalização, marcação de classes de gramática, segmentação de frases, classificação de texto, lematização, análise morfológica e vinculação de entidades (SPACY, 2023).

2.2. Certificados de Atividades Complementares

Conforme SOUTHIER e DORNELES (2022) as atividades complementares de curso constituem um requisito obrigatório para a formação acadêmica e profissional de um estudante em diversas instituições de ensino do país. Especificamente, no Instituto Federal Farroupilha (IFFar), Campus Frederico Westphalen, as atividades podem ser realizadas em diversas categorias, como participação em eventos, realização de cursos de curta duração, estágios, bem como atividades de pesquisa, ensino e extensão. Além disso, cada curso da instituição define uma carga horária mínima de atividades que o estudante precisa cumprir e comprovar mediante a apresentação de certificados na coordenação de curso.

Ao receber um certificado, além do tipo de atividade, o coordenador de curso precisa extrair quatro informações principais: nome do estudante, título da atividade, período de realização e carga horária. Cada uma destas informações pode ser considerada uma entidade nomeada diferente. Na Figura 1 é apresentado um exemplo de um certificado onde estas quatro entidades estão destacadas em amarelo.

Figura 1. Exemplo de um certificado de participação em evento

2.3. Trabalhos Relacionados

Até o momento, não há registro na literatura acadêmica de uma proposta que de uma ferramenta que extrai informações relevantes de certificados de atividades complementares de curso, mas há trabalhos semelhantes que extraem informações de outras fontes. Como o trabalho desenvolvido por ALLES (2018) que teve como proposta a utilização de aprendizado supervisionado para extração de entidades nomeadas do Diário Oficial da União (DOU), onde é apresentado um estudo explorando os conceitos e aplicações de 4 ferramentas de processamento de linguagem natural, a OpenNLP e a CoreNLP para reconhecimento de entidades nomeadas e a NLTK e a Syntaxnet para reconhecimento morfossintático.

Para explorar os conceitos de ferramentas que realizam PLN, o autor propõe uma metodologia que consiste na construção de um *corpus* específico para auxiliar no reconhecimento de entidades do DOU, realizado o seu processamento visando o entendimento linguístico das palavras em um texto, e em seguida, a quantidade de entidades presentes no texto é comparada com a quantidade de entidades reconhecidas e a qualidade das entidades reconhecidas é verificada. O *corpus* foi construído com a OpenNLP, utilizando aprendizado supervisionado, para que fosse elaborada uma proposta de construção de um corpus específico para extrair entidades nomeadas com melhor qualidade, comparando com os resultados dos *corpus* disponíveis. Os resultados quantitativos e qualitativos obtidos pelo DOU-Corpus se mostraram superiores em comparação com outras estratégias da literatura.

Outro trabalho é o desenvolvido por FONSECA et. al (2015), onde é apresentada a construção de um modelo para o reconhecimento de entidades nomeadas utilizando o NameFinder, uma classe contida no OpenNLP, que tem como objetivo reconhecer e classificar entidades nomeadas para o Português, dada a inexistência de um modelo para língua portuguesa no OpenNLP. Foram utilizados os *corpus* Amazônia e Harem para treinar e

avaliar o modelo considerando 10 categorias: Pessoa, Local, Organização, Acontecimento, Obra, Abstração, Coisa, Tempo, Valor e Outro. Na avaliação do modelo é comparado o número de entidades anotadas do Harem com as encontradas e classificadas corretamente pelo modelo considerando as 10 categorias, e apresentando os resultados de *precisão, recall* e *f-measure* para cada uma. Também é apresentada uma comparação da *precisão, recall* e *f-measure* das categorias Pessoa, Local e Organização no OpenNLP em comparação com os resultados de outras ferramentas como NERP-CRF, LTASK, Freeling e PALAVRAS. Os resultados apresentados no trabalho são compatíveis com os demais modelos, podendo ser ligeiramente superior em alguns aspectos.

3. Metodologia

Esta seção apresenta detalhes sobre a metodologia utilizada para o projeto da ferramenta proposta, bem como delimita o escopo dos materiais de dados da pesquisa.

3.1. Delimitação de Escopo de Certificados e Entidades Nomeadas de Interesse

Devido a existência de inúmeros tipos de certificados de atividades complementares representados em formatos de mídia diversos é necessário delimitar o escopo desta pesquisa. Desta forma, assume-se que um certificado de atividade complementar está representado no formato de um arquivo PDF e que as informações de interesse estão disponíveis em sua primeira página. Além disso, um certificado pode ter as informações organizadas em qualquer *layout* mas devem estar disponíveis em algum objeto de texto no arquivo PDF. Essa delimitação implica que estão excluídos do escopo desta pesquisa certificados cujas informações estão representadas por meio de um objeto de imagem inserido no arquivo PDF.

Em relação às entidades nomeadas de interesse que precisam ser extraídas de um certificado, são consideradas quatro categorias: NOME, TITULO, CARGA e PERIODO. A categoria NOME especifica um ou mais nomes atribuídos a certificação da atividade realizada. A categoria TITULO se refere ao título da atividade, podendo ser o nome de um evento ou curso, por exemplo. A categoria CARGA se refere a carga horária da atividade. Por fim, a categoria PERIODO se refere ao período em que a atividade certificada foi realizada, podendo ser, por exemplo, a data de conclusão ou as datas de início e final da atividade. No certificado da Figura 1, por exemplo, as entidades NOME, TITULO, CARGA e PERIODO correspondem, respectivamente a "João da Silva Lemos", "XI Encontro Anual de Tecnologia da Informação (EATI)", "25/01/2021 e 28/01/2021" e "16 horas".

3.2. Procedimento para Converter um Certificado no Formato PDF em Texto

Para que seja possível extrair as entidades nomeadas de interesse de um certificado no formato PDF é necessário, primeiramente, convertê-lo em um formato de texto. Para realizar esse processo de conversão, utiliza-se o utilitário de linha de comando denominado Pdftotext, o qual vem incluído com a biblioteca de renderização de arquivos PDF Poppler (POPPLER, 2023). Após a conversão com o utilitário, o texto é convertido em uma única linha por meio da remoção de quebras de linhas, remoção de espaços duplicados e substituição de aspas duplas por aspas simples. Desta forma, a aplicação do procedimento descrito nesta seção no certificado da Figura 1 resulta no certificado em formato texto exibido na Figura 2.

```
CERTIFICADO Certificamos para os devidos fins que João da Silva Lemos participou do XI
Encontro Anual de Tecnologia da Informação (EATI), ocorrido entre 25/01/2021 e
28/01/2021, perfazendo um total de 16 horas. André Fiorin Coordenação do Curso de
Sistemas para Internet-IFFar Solange Pertile Coordenação do curso Sistemas de
```

```
Informação-UFSM Este certificado foi entregue a João da Silva Lemos e registrado à fl:
XX do livro respectivo número YY sob o número de registro ZZ. Chave de Verificação:
BBBB.BBBB.BBBB.BBBB.BBBB      Verificação:      www2.fw.iffarroupilha.edu.br/autenticacao
Frederico Westphalen, 28 de janeiro de 2021.
```

Figura 2. Exemplo de um certificado convertido em formato texto

3.3. Procedimento de Treinamento com a Biblioteca spaCy

A biblioteca spaCy fornece recursos para a criação de um modelo para extração de entidades nomeadas específicas da aplicação por meio de um processo de treinamento. Este processo requer a especificação das entidades nomeadas que serão aprendidas, a rotulação de um conjunto de dados de treinamento e a escolha de parâmetros de treinamento.

Para este trabalho o processo de treinamento com o spaCy foi configurado para reconhecimento de quatro entidades: NOME, TITULO, CARGA e PERIODO conforme definido na Seção 3.1. O processo de rotulação de dados envolve criar uma estrutura de dados específica para cada certificado presente no conjunto de dados de teste. Essa estrutura identifica a posição em que cada entidade nomeada é encontrada no texto de um certificado. Para exemplificar o procedimento de rotulação, para o certificado em modo texto da Figura 2, é criada a estrutura em Python apresentada na Figura 3.

```
("CERTIFICADO Certificamos para os devidos fins que João da Silva Lemos participou do XI
Encontro Anual de Tecnologia da Informação (EATI), ocorrido entre 25/01/2021 e
28/01/2021, perfazendo um total de 16 horas. André Fiorin Coordenação do Curso de
Sistemas para Internet-IFFar Solange Pertile Coordenação do curso Sistemas de
Informação-UFSM Este certificado foi entregue a João da Silva Lemos e registrado à fl:
XX do livro respectivo número YY sob o número de registro ZZ. Chave de Verificação:
BBBB.BBBB.BBBB.BBBB.BBBB      Verificação:      www2.fw.iffarroupilha.edu.br/autenticacao
Frederico Westphalen, 28 de janeiro de 2021.", {"entities":[(50, 69,"NOME"),(84,
136,"TITULO"),(201, 209,"CARGA"),(153, 176, "PERIODO")]})
```

Figura 3. Exemplo de uma estrutura de rotulação utilizada no treinamento do spaCy

Após a especificação dos dados de treinamento, o procedimento de treino pode ser iniciado por um determinado número de épocas. Durante cada época, os dados de treinamento são embaralhados visando evitar padrões de aprendizado. Além do número de épocas é possível definir uma taxa de controle de *dropout* para evitar *overfitting*. Após o fim do treinamento, o modelo treinado é salvo em um arquivo e pode ser invocado para realização de testes.

4. Experimentos Computacionais e Resultados

Para validar a ferramenta proposta neste trabalho são utilizados certificados de atividades complementares de curso obtidos da base de dados do Sistema Integrado de Validação de Atividades Complementares (SIVAC), um sistema web desenvolvido em Django por SOUTHIER e DORNELES (2022), que é utilizado pelos cursos de Bacharelado em Ciência da Computação e Técnico em Informática Integrado do IFFar/FW. A base total de certificados do SIVAC possui em torno de 530 certificados dos mais variados tipos, de cursos, de eventos, palestras, de estágio, entre outros, sendo grande parte dos arquivos em formato PDF. Após aplicar as restrições consideradas na Seção 3.1, foram utilizados 430 certificados que atendiam as delimitações deste estudo e, a este conjunto chamaremos de *Corpus* de Dados do SIVAC (CDS). Esse conjunto de arquivos foi dividido em 2 partes: CDS80 contendo 80% dos certificados para a realização do treinamento e CDS20 contendo

20% dos certificados para teste. A divisão foi realizada de forma aleatória por meio da biblioteca random.

Os resultados dos experimentos foram obtidos em uma máquina com um processador Intel® Core™ i7-8565U 1.8GHz e 8GB de memória RAM, executando o Windows 11 Home Insider Preview Single Language como sistema operacional. A implementação da ferramenta foi realizada em Python 3.11.3, utilizando a biblioteca spaCy 3.6.1, no IDE PyCharm 2023.1 (Community Edition) e foi utilizada a versão 3.03 do utilitário Pdftotext.

O primeiro experimento realizado consistiu no treinamento do modelo de REN usando o método apresentado na Seção 3.3. Por meio de testes *ad-hoc* foram identificados parâmetros para o treinamento definindo o número de épocas=200 e uma taxa de dropout=0,5. Com estes parâmetros, foi utilizado aproximadamente 67 minutos de tempo computacional para a conclusão do treinamento e geração do modelo com o conjunto CDS80.

O segundo experimento teve como objetivo avaliar o desempenho do modelo produzido no experimento anterior utilizando o conjunto de dados CDS20. Os resultados deste experimento são apresentados na Tabela 1. Essa tabela apresenta as métricas de avaliação informando a *precisão*, *recall* e *F-score* para cada uma das quatro entidades de interesse, bem como uma média das métricas considerando todas as entidades. A métrica de *precisão*, mede a proporção de ocorrências positivas identificadas corretamente pelo modelo em relação ao número total de ocorrências que o modelo previu como positivas, incluindo verdadeiros positivos e falsos positivos, em outras palavras, a métrica informa a quantidade de acerto do modelo em relação aos resultados dos dados de teste informados. A métrica *recall*, também conhecida como sensibilidade, mede a proporção de ocorrências positivas que foram previstas corretamente pelo modelo em relação ao total de ocorrências positivas dos dados, incluindo os verdadeiros positivos e falsos negativos. Por fim, a métrica *F-score* que combina o resultado das métricas de *precisão* e *recall* em um resultado único, calculando a média harmônica entre as 2 métricas, como resultado temos uma medida balanceada do desempenho do modelo. Todas as métricas estão dispostas em valores entre 0 e 1.

Tabela 1. Resultados do modelo

	NOME	**TITULO**	**CARGA**	**PERIODO**	**Média**
Precisão	1.0	0.93750	1.0	0.95238	0.97247
Recall	1.0	0.87209	0.97674	0.93023	0.94477
F-score	1.0	0.90361	0.98824	0.94118	0.95826

Com base nos resultados apresentados na tabela, é possível realizar uma análise do desempenho do modelo para as diferentes entidades. Em relação à entidade NOME, o modelo obteve excelentes resultados com as métricas de *precisão, recall* e *F-score* com valor 1.0, indicando que o modelo foi capaz de identificar corretamente todas as ocorrências desta entidade. Para a entidade TITULO, os resultados também são bastante positivos, com a *precisão* de 0.93750, *recall* de 0.87209 e *F-score* de 0.90361, que sugere uma boa capacidade do modelo em identificar títulos, mas com uma pequena margem de erro. No caso da entidade CARGA, é apresentada uma *precisão* perfeita com valor 1.0, que indica um acerto para todas as predições positivas, mas um *recall* ligeiramente inferior de 0.97674, o que resultou em um *F-score* de 0.98824. Para a entidade PERIODO, o modelo também obteve um desempenho sólido, com a *precisão* de 0.95238, *recall* de 0.93023 e *F-score* de 0.94118. Analisando o resultado geral do modelo para todas as entidades, a média alcançou a *precisão* de 0.97247, *recall* de 0.94477 e *F-score* de 0.95826.

5. Considerações Finais

Neste trabalho foi apresentada uma proposta de ferramenta para a realizar o reconhecimento de entidades nomeadas de certificados de atividades complementares de curso considerando quatro entidades: nome, título, carga horária e período. Foi apresentada uma implementação de um modelo REN treinado com a biblioteca spaCy, cujo desempenho foi avaliado por meio de um *corpus* de dados composto de 430 certificados provenientes dos cursos do eixo de informação e comunicação do IFFar/FW. Os resultados computacionais indicam que o modelo REN implementado apresentou um bom desempenho geral na tarefa de identificação das entidades de interesse. Em especial, o modelo se destacou na extração das entidades de nomes e de carga horária atingindo os valores máximos das métricas avaliadas. Em relação a extração de títulos dos certificados, mesmo sendo uma informação menos estruturada, o modelo apresentou um bom desempenho, comparável com o obtido na extração de períodos.

Com base nesses resultados, conclui-se que o modelo proposto é promissor para auxiliar na melhoria de sistemas informatizados específicos para validação e gerenciamento de atividades complementares de curso como o SIVAC, ou ainda ser integrado em sistemas acadêmicos pré-existentes em instituições de ensino públicas e privadas.

Este trabalho apresenta ainda as seguintes sugestões de trabalhos futuros: (i) avaliar o conjunto desempenho do modelo treinado utilizando validação cruzada; (ii) extração de novas entidades nomeadas em outros tipos de certificados; (iii) avaliação do modelo proposto com dados de testes coletados de outras instituições públicas.

Referências

ALLES, V. J. Construção de um Corpus para Extrair Entidades Nomeadas do Diário Oficial da União Utilizando Aprendizado Supervisionado. Dissertação de Mestrado em Engenharia Elétrica, Publicação 714/2018, Departamento de Engenharia Elétrica, Faculdade de Tecnologia Universidade de Brasília. Brasília, DF. dez. 2018.

FONSECA, E. V; CHIELE, G. C; VIEIRA, R; VANIN, A. A. Reconhecimento de Entidades Nomeadas para o Português Usando o OpenNLP. PUCRS. Porto Alegre, RS, Anais do ENIAC 2015, 2015.

SILVA, A. V. e. Um modelo de classificação para o Reconhecimento de Entidades Nomeadas. Dissertação de Mestrado. Faculdade de Filosofia, Letras e Ciências Humanas. USP. São Paulo, SP, dez. 2020.

SOUTHIER, P. H; DORNELES, A. P. Desenvolvimento de uma Plataforma Web em Django para Gerenciamento de Atividades Complementares de Curso. Instituto Federal de Educação, Ciência e Tecnologia Farroupilha (IFFar), Frederico Westphalen, RS, Brasil. Ano 11, n. 1. Anais do XIII Encontro Anual de Tecnologia da Informação - EATI. nov. 2022.

SPACY. spaCy. Disponível em: <https://spacy.io/>. Acesso em: 07 out. 2023.

POPPLER. Poppler. Disponível em: <https://poppler.freedesktop.org/>. Acesso em: 07 out. 2023.