

Universidade de Pernambuco

Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Tese de Doutorado

Área: Inteligência Computacional

Título: Extração de Entidades e Relacionamentos em Documentos Tabulares e de Contexto Longo

Orientador – Byron Leite Dantas Bezerra (byron.leite@upe.br)

Coorientador – Cleber Zanchettin (cz@cin.ufpe.br)

Descrição

Em face da recente transformação digital, as empresas estão cada vez mais digitalizando seus processos e diminuindo custos operacionais. Como exemplo, podemos citar o uso de tecnologias de OCR para converter em texto editável os documentos digitalizados [1]. Com isso, a empresa pode diminuir custos com a digitação de informações contidas nestes documentos.

No exemplo citado anteriormente, um objetivo possível seria extrair do documento múltiplos dados e relações entre estes dados. Por exemplo, na imagem do cupom fiscal ao lado, podemos identificar o estabelecimento, com seu nome (Peixaria Sul), endereço e telefone. Podemos identificar informações da venda, data e hora, que poderiam ser agrupados em outra entidade. Além destas, podemos também identificar os itens comprados, sendo que cada item tem uma descrição, valor unitário, quantidade e valor total do item.

Nos últimos anos, o desenvolvimento de algoritmos ponta a ponta para detecção e reconhecimento de texto a partir de imagens tem evoluído significativamente devido a utilização de *Deep Learning* [2]. Da mesma forma, a utilização de modelos de Inteligência Artificial em Processamento de Linguagem Natural (Natural Language Processing) [3] possibilitou avanço na extração de informações a partir de texto estruturado. Neste sentido, o recente desenvolvimento dos modelos Transformer, BERT e seus sucessores viabilizou a extração de dados de interesse em documentos [4].

Nesse sentido, a extração de entidades e relações entre estas é um problema relevante e sobre o qual muitas pesquisas continuam sendo desenvolvidas [5]. Esse problema pode ser mais complexo em documentos mais longos e composto de tabelas que apresentem relações estruturais em múltiplas páginas. É neste contexto que reside o projeto de doutorado aqui proposto. A proposta envolve uma equipe multidisciplinar e faz parte do projeto de pesquisa e inovação "**Algoritmos e Modelos de Inteligência Artificial e Visão Computacional para Processamento Inteligente de Documentos**" fomentado pelo CNPQ, e em parceria com a empresa Di2Win (www.di2win.com). Para conhecer mais sobre o orientador e seus temas de pesquisa, convido a assistir a entrevista [aqui](#).



*** CUPOM PARA SIMPLES CONFERENCIA ***
*** NÃO É DOCUMENTO FISCAL ***

PEIXARIA SUL
AV. ENGENHEIRO DOMINGOS FERREIRA 500
BAIRRO: BOA VIAGEM
FONE: (81)3328-3101

Venda: 01/12 - Cupom: - Op: CUMING
Data: 01/10/2022 16:26
Vendedor:

| Descrição do item | valor | Qtde. | Total |
|--------------------------|------------|-------|---------------|
| CAMARAO INTERMED FRESCO | 39,90 | 4,20 | 169,40 |
| DESCASQUE | 1,00 | 8 | 8,00 |
| MARISCO | 24,90 | 0,574 | 14,29 |
| FILE SURUBIM | 39,90 | 0,614 | 24,50 |
| FILE SURUBIM | 39,90 | 0,336 | 13,41 |
| FILE SURUBIM | 39,90 | 0,622 | 24,80 |
| FILE SURUBIM | 39,90 | 0,530 | 21,15 |
| FILE SURUBIM | 39,90 | 0,574 | 22,90 |
| FILE SURUBIM | 39,90 | 0,514 | 20,51 |
| Sub total..... | | | 309,98 |
| Desconto nos itens: 0,00 | | | 0,00 |
| Desconto (-)..... | | | 0,00 |
| Acrescimos (+)..... | | | 0,00 |
| TOTAL PEDIDO | R\$ | | 309,98 |
| Qtde. de itens: 9 | | | |
| CARTAO CREDITO | | | 309,98 |

Referências

- Bezerra, Byron L. D.; Zanchettin, Cleber; Toselli, Alejandro H.; Pirlo, Giuseppe (Eds.). Handwritten: Recognition, Development and Analysis. New York: Nova Science Publishers, 2017.
- LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015). <https://doi.org/10.1038/nature14539>.
- RAMOS NETO, Antônio dos Santos; FELIX, João Paulo; SANTOS, Wylliams; BEZERRA, Byron L. D.; RODRIGUES, Cleyton Mário de Oliveira. CROSSAGE: A cross-attentional graph and Transformer architecture for skill and knowledge recognition in job descriptions. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 16. , 2025, Fortaleza/CE. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025. p. 361-373. DOI: <https://doi.org/10.5753/stil.2025.37838>.
- Minh-Tien Nguyen, Dung Tien Le, Linh Le. Transformers-based information extraction with limited data for domain-specific business documents. Engineering Applications of Artificial Intelligence, v. 97, 2021.
- Wang, Hailin, et al. "Deep neural network-based relation extraction: an overview." Neural Computing and Applications (2022): 1-21.