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

Proposta de Tese de Doutorado

Área: Inteligência Computacional

Título: Extração de Dados Relacionados em Documentos

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

Coorientador – Cleber Zanchetin (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.


*** CUPOM PARA SIMPLES CONFERENCIA ***
*** NAO E DOCUMENTO FISCAL ***

PEIXARIA SUL
AV ENGENHEIRO DOMINGOS FERREIRA 500
BAIRRO: BOA VIAGEM
FONE: (81)3326-3101

Ven.: - Cupom: - Op: CUMMING
Data: 05/10/2022 16:36
Vendedor:

Descricao do item	Valor	Qtde.	Total
CAMARAO INTERMED FRESCO	39,90	4,20	160,40
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	24,82
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
Desconto (-).....: 0,00
Acrescimos (+).....: 0,00

TOTAL PEDIDO R$ 309,98

Qtde. de itens: 8

CARTAO CREDITO 309,98

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].

Por outro lado, um problema ainda em aberto é justamente encontrar as relações existentes entre os campos lidos [5]. Esse problema pode ser mais complexo em documentos mais longos, como um contrato. 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 "***Soluções de Reconhecimento de Escrita e Processamento de Imagens para DPO's***" fomentado pela FACEPE e CNPQ, e em parceria com a empresa Di2Win (www.di2win.com). Em função disso, o aluno poderá receber bolsa, a depender do currículo do discente e dedicação ao projeto.

Referências

1. Byron L. D. Bezerra, Cleber Zanchettin, Alejandro H. Toselli, and Giuseppe Pirlo (Eds.). Handwritten: Recognition, Development and Analysis. New York: Nova Science Publishers, 2017.
2. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015). https://doi.org/10.1038/nature14539.
3. Neto, A.F.d.S.; Bezerra, B.L.D.; Toselli, A.H. Towards the Natural Language Processing as Spelling Correction for Offline Handwritten Text Recognition Systems. *Appl. Sci.* **2020**, *10*, 7711. https://doi.org/10.3390/app10217711.
4. 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.
5. Wang, Hailin, et al. "Deep neural network-based relation extraction: an overview." Neural Computing and Applications (2022): 1-21.