Ciência da Computação / Sistemas de Informação

# UFLA-FORMS: Uma Base de Dados de Formulários da UFLA para a Tarefa de Extração de Informação

Victor Gonçalves Lima - 7º módulo de Ciência da Computação, UFLA, bolsista PIBIC/UFLA

Denilson Alves Pereira - Orientador DCC, UFLA - Orientador(a)

## Resumo

A Inteligência de Documentos é uma subárea da Inteligência Artificial que utiliza tarefas de Processamento de Linguagem Natural (NLP), como Reconhecimento de Entidades Nomeadas (NER) e Extração de Relação (RE) para capturar, extrair e compreender, de forma automática, documentos estruturados. Para o avanço da área, é fundamental que haja disponível bases de dados que possibilitem os treinamentos de modelos inteligentes. Nesse sentido, este trabalho consiste na criação e disponibilização de uma base de dados em língua portuguesa de formulários acadêmicos da UFLA, com ênfase na extração de pares chave/valor. Para a criação da base de dados, denominada UFLA-FORMS, e o formato de rotulação, foi selecionada outra base de dados como referência, a FUNSD. Na construção da base de dados, foram extraídos e selecionados 200 formulários dos domínios da UFLA para preenchimento e rotulação. Por questões de proteção de dados, cada formulário foi preenchido com dados fictícios, como nomes, CPFs e endereços, tomando o cuidado necessário para que eles sejam consistentes. Utilizando uma ferramenta em desenvolvimento implementada junto ao trabalho, cada sentença das amostras é selecionada manualmente para que sejam extraídos seu texto e o seu posicionamento por OCR. Cada sentença é rotulada entre ?Other? (Outro), ?Header? (Cabeçalho), ?Question? (Pergunta) ou ?Answer? (Resposta) e, em seguida, seus pares chave-valor são unidos, tudo de acordo com a rotulação da base tomada como referência. Como resultado, cada amostra é composta por uma imagem bidimensional extraída do documento e um arquivo de texto, no formato JSON, contendo toda a rotulação relacionada a essa imagem. Utilizando o modelo LayoutXLM, baseado em Transformer, um treinamento na tarefa de NER realizado com a base através de cross-validation com 10 partições aleatórias (10% para teste), resultou experimentalmente um score overall F1 de 0,92. Conclui-se que a produção e disponibilidade dessa base de dados em língua portuguesa contribui para o avanço nas pesquisas em Inteligência Artificial para documentos estruturados no país, abrangendo características linguísticas e culturais, não encontradas em outras bases de origem estrangeira.

Palavras-Chave: Document Understanding, Information Extraction, Dataset.
Instituição de Fomento: UFLA

Link do pitch: https://youtu.be/5fsv2LsaNRY