

Ciência da Computação / Sistemas de Informação

## **Criação de um Conjunto de Dados para Reconhecimento de Entidades Nomeadas sobre Patentes de Mobilidade Inteligente**

Pedro Fonseca Rodrigues de Sousa - 7º módulo de Ciência da Computação, UFLA, Bolsista PIBIC/CNPq

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

André Grutzmann - Coorientador DCC, UFLA

Cláudio Manoel dos Reis Júnior - 8º módulo de Ciência da Computação, UFLA, Coautor

### **Resumo**

O trabalho apresenta a criação de um conjunto de dados para a tarefa de Reconhecimento de Entidades Nomeadas (NER) em textos relacionados a patentes de mobilidade inteligente. NER é uma tarefa fundamental na área de Processamento de Linguagem Natural (PLN) que consiste em identificar e classificar entidades nomeadas em textos, tais como nomes de pessoas, organizações e locais. O conjunto de dados criado é específico para o domínio de patentes de mobilidade inteligente, o que pode ser útil para aplicações como análise de patentes, monitoramento de tendências tecnológicas e identificação de oportunidades de inovação. Para criar o conjunto de dados, foi realizada uma coleta de dados, por meio de Web Scraping, no Google Patentes, onde foram coletadas 144 patentes relacionadas ao tema de mobilidade inteligente. As patentes coletadas foram rotuladas manualmente por dois rotuladores, que identificaram e classificaram as entidades nomeadas nos textos dentre as classes: funcionalidade, hardware, software, comunicação e pessoa. No entanto, existiram discordâncias entre os rotuladores, que foram resolvidas após discussões e revisão do conjunto de dados. Após as revisões, 26 patentes foram excluídas do conjunto de dados devido a inconsistências ou erros, restando um total de 118 patentes válidas. O objetivo é criar um dataset de alta qualidade que possa ser utilizado para treinar modelos de NER e melhorar a eficácia do reconhecimento de entidades nomeadas em textos relacionados a patentes de mobilidade inteligente. Como trabalho futuro, o conjunto de dados será avaliado por meio de modelos de aprendizagem de máquina, sendo testados com diferentes algoritmos e métricas de avaliação. O conjunto de dados também será disponibilizado para a comunidade científica, para que possam ser utilizados em estudos e projetos relacionados à mobilidade inteligente.

Palavras-Chave: Extração de Informação, Conjunto de Dados, Mobilidade Inteligente.

Instituição de Fomento: CNPq

Link do pitch: <https://www.youtube.com/watch?v=EdFZhqOt6AU>