



USO DE PLN PARA SANEAMENTO DE CADASTRO

Mateus de Oliveira Arcanjo¹; Lucas Magno de Andrade Bastos²; Douglas Roberto de Matos Pimentel³

¹ Dom Rock. (mateus.arcanjo@domrock.ai).

² Dom Rock. (lucasmag261099@gmail.com).

³ Dom Rock. (douglas.pimentel@domrock.ai).

Resumo: O uso do processamento de linguagem natural é bastante promissor e amplo e proporciona diversos cenários de aplicação no mundo dos negócios. Um caso particular que pudemos estudar foi o contexto de uma empresa de manufatura que obteve um crescimento expressivo em seus negócios levando a diferentes momentos de implantação de processos e sistemas que gerou, ao longo do tempo, um cenário de múltiplas cópias de cadastros que são usadas por diferentes áreas da organização. Essa situação é bastante comum em ambientes de negócios que passam por essas transformações ao longo do seu ciclo de vida contrariando a lógica da administração que pressupõem que o uso de sistemas deveria promover ambientes de cadastros confiáveis e unificados. Nesse sentido, tomamos como cenário de estudo um caso particular de cadastro que é o de produtos que são necessários para compor as soluções dessa indústria e, portanto, referem-se a itens de compras junto à fornecedores. Com essa abordagem, encontramos o ambiente propício para aplicar tecnologia de processamento de linguagem natural (PLN) com objetivo de sanear e unificar as diferentes cópias de cadastros de itens de compra e fornecedores. Mais especificamente, a técnica selecionada para o estudo foi a aplicação de análises de similaridades a fim de identificar e reconciliar diferentes descrições de itens cadastrados em várias fontes distintas e criar condições de gerar uma base única de itens de forma automatizada. A metodologia aplicada foi: i) identificação das diferentes fontes de dados; ii) pré-processamento das fontes de dados para seleção de conjunto de itens coerentes com as principais etapas do processo de negócio de compras de itens e respectivos período; iii) carga das bases de dados; iv) aplicação de modelos de Embedding (doc2vec e Bert) e técnicas de similaridade (cosseno e distância de Levenshtein); v) avaliar os critérios de assertividade e; vi) selecionar a melhor abordagem. Os resultados alcançados demonstraram que o modelo doc2vec teve o melhor desempenho para identificação de similaridade de descrição de itens de diferentes bases de dados consideradas. O modelo doc2vec foi elaborado a partir de treinamento baseado em base selecionada para essa finalidade. Dessa forma, foi possível automatizar o processo de reconciliação de informações, reduzindo erros e inconsistências nos cadastros de itens de compra. As conclusões deste estudo enfatizam a importância do uso de processamento em linguagem natural, especialmente o modelo doc2vec, para o saneamento de cadastro em ambientes empresariais complexos, onde múltiplas descrições de um mesmo produto são comuns. A automatização proporcionada pelo estudo otimiza o tempo e os recursos envolvidos no processo de reconciliação de dados e, também, eleva a confiabilidade e precisão das informações cadastrais.

Palavras-chave: Processamento de Linguagem Natural (PLN); Saneamento de Cadastro; Manufatura; Similaridade; Processos de negócios.