AVALIAÇÃO DA EFICÁCIA DE FUNÇÕES DE SIMILARIDADE PARA IMPORTAÇÃO DE NOTAS FISCAIS ELETRÔNICAS

Higor Moreira[1], Iago Mocelin da Silva[1], Edimar Manica[1*]
*Orientador(a)

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS) - *Campus* Ibirubá. Ibirubá, RS

Uma Cooperativa de Ibirubá/RS verificou a possibilidade do IFRS – Campus Ibirubá desenvolver uma solução para minimizar o esforço manual necessário para a importação das notas fiscais eletrônicas de fornecedores para seu sistema interno. Esse esforço é necessário porque o código e a descrição do produto no cadastro dos fornecedores são diferentes das especificações de produtos presentes no software utilizado pela empresa. Dentre as diferenças encontram-se variação na representação do conteúdo, omissão de informações, uso de acrônimos e abreviações. Nesse contexto, este trabalho tem como objetivo identificar as funções de similaridade textual mais eficazes para encontrar especificações de produtos oriundas de sistemas diferentes que se referem ao mesmo produto. A metodologia adotada é composta pelas seguintes etapas: revisão bibliográfica, seleção de dados, rotulação dos dados, limpeza de dados e avaliação. Na etapa de revisão bibliográfica, foram analisadas as funções de similaridade disponíveis nas bibliotecas Textdistance e Strsim e selecionadas aquelas que retornam um escore normalizado. Na etapa de seleção de dados, foi obtida uma amostra de especificações de produtos do sistema da Cooperativa e outra das notas fiscais eletrônicas emitidas pelos fornecedores. Na etapa de rotulação de dados, foi criado um gabarito, onde para cada especificação de produto contida em uma nota fiscal, foi identificada manualmente qual a especificação de produto era equivalente no sistema da Cooperativa. A etapa de limpeza de dados consistiu em suavizar ruídos, identificar valores discrepantes e inconsistências nas notas fiscais. Na etapa de avaliação, foram testadas 11 funções de similaridade textual para verificar a eficácia em identificar as especificações de produtos oriundas de sistemas diferentes que se referem ao mesmo produto. A eficácia foi calculada por meio das seguintes métricas: precisão, revocação e F1. Também, foi avaliada a relação de custo-benefício em aplicar a técnica de blocagem por NCM (Nomenclatura Comum do Mercosul). Como resultados foi possível elaborar uma base de dados com 5.788 especificações de produtos distintos extraídas das notas fiscais reais, associadas com especificação de produto equivalente no sistema interno da Cooperativa. Além disso, a abordagem aplicando a blocagem por NCM apresentou a melhor relação de custo-benefício (eficácia X tempo de processamento). Entre as 11 funções de similaridade testadas, as que apresentaram os melhores resultados de eficácia foram Cosine, Jaccard e Sorensen pela capacidade de identificar uma especificação de produto mesmo que essa esteja com a ordem das palavras trocadas. Os resultados obtidos foram utilizados para implementar um deduplicador semiautomático.

Palavras-chave: Nota fiscal eletrônica. Similaridade. Duplicatas.

CNPq FAPERGS