Avaliação de Funções de Similaridade Textual para Identificação de Duplicatas

Higor Moreira[1], Ronaldo Serpa da Rosa[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 tecnológica 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 determinadas informações, uso de acrônimos e abreviações. Nesse contexto, está sendo desenvolvido um projeto de pesquisa que 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 bibliotecas Textdistance e Strsim e selecionadas aquelas que retornam um escore normalizado. A etapa de seleção de dados visa obter uma amostra de especificações de produtos do sistema da Cooperativa e outra das notas fiscais eletrônicas emitidas pelos fornecedores. Nessa etapa, foi desenvolvido um extrator utilizando a linguagem de programação Java e a biblioteca w3c.dom para coletar das notas fiscais eletrônicas apenas as informações requeridas para o projeto, uma vez que há informações confidenciais. Atualmente, o projeto encontra-se na etapa de rotulação de dados, que visa identificar manualmente pares de registros, onde cada par contém duas especificações do mesmo produto, uma do sistema da Cooperativa e outra de uma nota fiscal de fornecedor. Esses pares serão utilizados na etapa de avaliação. A etapa de limpeza de dados consistirá em suavizar ruídos, identificar valores discrepantes e inconsistências nas notas fiscais. A etapa de avaliação consistirá em avaliar as funções de similaridade textual para identificar especificações de produtos oriundas de sistemas diferentes que se referem ao mesmo produto. Os pares de registros serão comparados através de diferentes funções de similaridade, se o escore for maior que um determinado limiar, o par será considerado duplicata. Como resultados parciais, foram obtidas 133 mil especificações de produtos do sistema da Cooperativa e 876 notas fiscais eletrônicas de fornecedores contendo apenas os dados necessários pelo projeto, uma vez que foram pré-processadas pelo extrator desenvolvido. O projeto está sendo executado conforme o cronograma previsto. Após a rotulação dos dados, serão avaliadas 18 funções de similaridade da biblioteca Textdistance e 4 da Strsim.

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

CNPq