02 e 03 de Outubro de 2019
Ibirubá - RS - Brasil

INSTITUTO FEDERAL
Rio Grande do Sul
Campus Ibirubá

8º MOEPEX
MOSTRA DE ENSINO, PESQUISA E EXTENSÃO

Avaliação da Eficácia de Funções de Similaridade para Deduplicação de Especificações de Produto

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

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

Analisando o caso corrente de uma cooperativa local, notou-se que esta possuía funcionários para gerenciamento das notas fiscais de venda dos seus fornecedores, em arquivos *.xml*. Nestas, os produtos recebiam nomes com *strings* diferentes e estes funcionários, manualmente, digitavam novamente o nome dos produtos e elaboravam uma nota fiscal de compra para a cooperativa, lançando-a em seu banco de dados. A problemática surgiu quando esta cooperativa utilizava um sistema que limitava a quantidade de caracteres em uma *string*, resultando em especificações de produtos com grande uso de acrônimos e abreviações. Desta forma, o objetivo deste projeto é o desenvolvimento de um sistema automático que, através de um inteligente uso de funções de similaridade combinadas, a depender do caso analisado, consiga realizar este trabalho de identificar as especificações dos produtos descritos nas notas fiscais eletrônicas dos fornecedores e compará-las com as cadastradas no banco de dados da referida cooperativa, criando um sistema de deduplicação. Deduplicação é a identificação e redução de itens duplicados que se referem à mesma entidade. Em suma, pretende-se identificar as funções mais eficazes para encontrar especificações de produtos registradas de maneiras desiguais em sistemas diferentes. Foram testadas funções de similaridade que trabalham através de pontuações, sendo capazes de identificar duas *strings* como o mesmo produto, pedir confirmação do usuário ou descartar possibilidades, levando em consideração a pontuação dada para a semelhança entre as *strings*. O presente projeto é dividido em uma metodologia de sete etapas: revisão bibliográfica, seleção, rotulação e limpeza de dados, definição da técnica de blocagem e avaliação. Como resultados parciais, o projeto, que encontra-se em fase de seleção de dados, analisou vinte e duas notas fiscais de diferentes companhias e oitocentas notas fiscais da mesma cooperativa solicitante do projeto, com um extrator de dados de arquivos *.xml* próprio.

Palavras-chave: *Funções de Similaridade. Deduplicação.*

Trabalho executado com recursos do Edital nº 24/2019/PIBIC-EM, da Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação