# Aprimoramento da extração de dados de portarias institucionais

Iago Ivanir Dalmolin[1], Edimar Manica[1*]
*Orientador(a)

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

Os Institutos e as Universidades Federais emitem documentos, denominados portarias, onde constam os afastamentos, as progressões e outras informações sobre seus servidores. Esses documentos são disponibilizados nos sites oficiais, organizados de diferentes formas e muitas vezes sem a existência de uma ferramenta adequada para realizar buscas ou aplicar filtros, acarretando um demasiado gasto de tempo para os servidores e demais interessados encontrarem uma portaria. Esse projeto de pesquisa visa otimizar e avaliar o buscador Touba, que coleta as portarias dos diversos repositórios oficiais e as disponibiliza para buscas por palavras-chave. A primeira versão do buscador foi desenvolvida em um projeto anterior. A ferramenta Touba coleta, nos sites institucionais, os arquivos PDF que publicam as portarias. Após, esses documentos são semiestruturados em arquivos XML, onde são identificados seu número, data de publicação e conteúdo. Na sequência, é realizada a disponibilização desses dados para buscas através de uma interface gráfica online. Nesse projeto, estão sendo aprimoradas as etapas do buscador a fim de melhorar sua eficácia. Até o momento foi aprimorada a identificação de número e data de publicação. Inicialmente, foram selecionadas cinco bases de dados: UFRGS, IFRS, Campus Ibirubá atual, Campus Ibirubá antigo e SIPPAGWeb. Em seguida, foram escolhidas aleatoriamente 60 portarias distintas para cada base de dados, sendo 30 para treino e 30 para teste. Após, foi criado um gabarito contendo para cada portaria o nome da base, o nome do arquivo, o link do documento, o número da portaria e a data da portaria. Para a base SIPPAGWeb, o gabarito foi criado automaticamente uma vez que o site disponibiliza os dados em JSON. Para as demais bases de dados, foi necessário fazer a anotação manual. Após, aplicou-se a ferramenta Touba para realizar a identificação do número e da data de publicação das portarias. Em seguida, novas otimizações foram realizadas a partir dos erros identificados. Esse processo se repetiu até uma melhoria significativa nos resultados. Por fim, a ferramenta Touba foi aplicada sobre as portarias de teste e a revocação, precisão e F1 foram calculadas para essas portarias. A hipótese era que as melhorias seriam genéricas o suficiente para contemplar portarias não utilizadas na otimização. Essa hipótese foi confirmada uma vez que a otimização da identificação do número obteve um ganho de F1 de 10%, 13%, 11%, 11%, 0% para as respectivas bases de dados UFRGS, IFRS, Campus Ibirubá atual, Campus Ibirubá antigo e SIPPAGWeb. Enquanto que a otimização da identificação da data obteve um ganho de F1 de mais de 100%, 8%, 25%, 31%, 0% para as respectivas bases de dados UFRGS, IFRS, Campus Ibirubá atual, Campus Ibirubá antigo e SIPPAGWeb. O presente trabalho foi realizado com apoio do IFRS.

Palavras-chaves: Portarias. Extração de Dados. Buscador.


CNPq FAPERGS