X Congresso Fluminense de Iniciação Científica e Tecnológica
23º Encontro de Iniciação Científica da UENF
15º Circuito de Iniciação Científica do IFFluminense
11º Jornada de Iniciação Científica da UFF
III Congresso Fluminense de Pós-Graduação
18ª Mostra de Pós-Graduação da UENF
3ª Mostra de Pós-Graduação do IFFluminense
3ª Mostra de Pós-Graduação da UFF
Ciência no nosso dia-a-dia
25 a 29 de junho de 2018

Ferramenta de Classificação de Documentos baseada em Mineração de Texto

Sávio Siqueira Cardoso de Souza, Leandro da Silva Foly

No cenário mundial da grande rede, já há algum tempo busca-se o desenvolvimento de tecnologias de busca e recuperação eficiente das informações, pois são grandes as dificuldades de encontrar informações relevantes em meio a este espaço não estruturado. O objetivo deste trabalho é apresentar uma proposta de Ferramenta de Classificação de Documentos baseada em recursos de Mineração de Textos. A ferramenta permite que o usuário cadastre alguns domínios do saber, bem como termos específicos ao mesmo, e indique um local contendo documentos textuais diversos. Em seguida, o protótipo faz a leitura de tais documentos, que serão categorizados de forma inteligente em cada domínio registrado previamente. A ferramenta, que utiliza recursos de Mineração de Texto, executa uma fase de pré-processamento, contendo passos como a Tokenização do texto, filtragem de *Stopwords* (palavras irrelevantes), e *Stemming* (extração de radicais dos termos), para que depois, o texto possa ser classificado na fase de processamento, onde são agrupados aos *tokens* radicalizados todas as palavras derivadas destes radicais, calculada a frequência de cada termo encontrado e comparados com os termos previamente cadastrados. O presente trabalho então apresenta uma experimentação do protótipo com diversos documentos em formato PDF obtidos no site *Google Acadêmico*, e utilizando-se como exemplo três domínios diferentes: “Educação”, “Computação” e “Saúde”. Dentre os resultados percebidos, nota-se que a ferramenta atingiu a eficácia por conseguir distinguir com sucesso os domínios de todos os documentos analisados através da arquitetura de mineração proposta. Adicionalmente, através da análise dos resultados, percebe-se que determinados arquivos obtiveram probabilidade de pertencerem a duas áreas distintas, dentre as cadastradas, por conta da própria síntese interdisciplinar do texto, como visto em documentos que versam sobre “técnicas pedagógicas para se lecionar computação”, por exemplo. Tais resultados contribuem para se discutir novas estratégias de classificação dos documentos, levando-se em consideração as interseções existentes entre os diversos domínios do saber.

Palavras-chave: Mineração de Texto, Classificação, Sistemas Inteligentes.

Instituição de fomento: IFFluminense