XI Congresso Fluminense de Iniciação Científica e Tecnológica
IV Congresso Fluminense de Pós-Graduação
A Ciência e os caminhos do desenvolvimento

Técnicas de Mineração de Texto aplicadas à classificação de documentos digitais

Samer das Chagas Macário, Leandro da Silva Foly

Mineração de Textos pode ser entendida como uma forma de extrair informações relevantes em bases de dados não estruturadas ou semiestruturadas. Surgiu a partir da necessidade de se descobrir automaticamente informações em textos, permitindo recuperar informações, extrair dados, resumir documentos, descobrir padrões, dentre outras análises possíveis de se realizar em documentos de texto. Pode ser utilizada com muitos propósitos, por exemplo identificar documentos similares e buscar dados relevantes dentro do documento. Na Mineração de textos, utilizam-se algumas técnicas, das quais destacam-se: Associação; Sumarização; Clusterização; Classificação/Categorização. Este trabalho objetiva acrescentar tais técnicas ao protótipo de software classificador de documentos já desenvolvido, sendo então adotado uma metodologia de pesquisa das técnicas citadas e implementação de um algoritmo a ser incluído no software. A pesquisa tem como foco principal a técnica de Classificação / Categorização, utilizada para classificar um conjunto de documentos em uma ou mais categorias predefinidas. Existem duas maneiras distintas para modelar um categorizador de textos: Orientada a documento, na qual os documentos se tornam disponíveis em diferentes momentos no tempo. E Orientada a categoria, quando uma nova categoria pode ser adicionada a um conjunto existente depois que uma série de documentos já tenham sido classificados em um determinado conjunto. Segundo (ALPAYDIN, 2004), as decisões de categorização tomadas pelos classificadores não podem ser as mesmas. As técnicas estudadas usam como base alguns algoritmos consagrados. Dentre eles, citam-se o K-Nearest Neighbors, o AdaBoost e o Naïve Bayes. O K-Nearest Neighbors (KNN) é um algoritmo de aprendizagem supervisionado, pertencente a um grupo de técnicas denominado de Instance Based Learning, e é considerado um dos melhores métodos para a classificação de texto. Assim, esse algoritmo foi escolhido para a implementação de um protótipo, que encontra- se em fase de experimentação, e pretende-se obter resultados a tempo de serem apresentados na versão final deste documento junto às discussões, mas já pode-se concluir, baseado nas pesquisas, que a técnica certamente trará resultados positivos ao software já construído.

Palavras-chave: Mineração de Texto, Classificação de Dados, Algoritmos de Classificação

Instituição de fomento: IFFluminense