Recuperação inteligente da informação Relatório atividade E-1.9

André Pacheco

Doutorado em Ciência da Computação
Programa de pós-graduação em informática
Universidade Federal do Espírito Santo

Conteúdo

1 Introdução

A atividade E-1.9 tem como objetivo a identificação dos tipos de texto existentes na base *repositorio-ufes*. Neste repositório existem diversos trabalhos realizados na universidade que não estão separados por tipos dos mesmos. Com isso foi desenvolvido um algoritmo para que esses arquivos fosse classificados. Por fim, foi gerado uma saída HTML linkando todos os arquivos de acordo com suas respectivas classes.

2 Classificação dos documentos

A base de dados em questão possui 4723 arquivos no formato .pdf de trabalhos produzidos na Universidade Federal do Espírito Santo. Esses trabalhos se dividem em quatro classes: dissertações de mestrado, teses de doutorado, livros e artigos. Para identificar os padrões dos arquivos foi utilizado um modelo probabilístico "gerenciado"seguindo a mesma ideia da atividade E-1.8. Esse gerenciamento foi feito baseado em observações des palavras mais comuns contidas nos arquivos da base e quantidade de caracteres contidas na amostra observada.

Como os arquivos estão em PDF, foi utilizado a biblioteca PDFMiner [1] no ambiente python. Essa biblioteca é capaz de ler os arquivos e retornar o conteúdo contido nos mesmos. Foram carregas as primeiras quatro paǵinas do documento para tomar a decisão em qual classe o mesmo pertence. Foi tomada essa decisão

[1] https://pypi.python.org/pypi/pdfminer/

pelos seguintes motivos: primeiramente a base possui cerca de 14 GB de documentos, se carregado todas as páginas, o tempo de excução seria extremamente alto; Além disso, utilizando as 4 primeiras páginas é possível classificar com confiança os documentos, tendo em vista que eles iniciam de maneira diferente.

Para classificar dissertações de mestrado são observados uma sequência de palavras chaves como as palavras: tese, dissertação, mestre, doutor, doutorado, mestrado, orientador, dentre outras. Com isso os caracteres contidas na amostra eram comparados com os valores e de acordo com a quantidade de palavras chaves encontradas é gerado uma probabilidade do documento pertencer a classe doutorado ou mestrado. No caso de livros, palavras chaves que possuem apenas nesse tipo de arquivo são: ISBN, EDUFES, editor, diagramação, dentre outras. E para artigos, DOI, palavras-chaves e publicado. Com isso, é utilizado a mesma estratégia para classificar. Além disso, o número de caracteres é importante principalmente na identificação de artigos. Como são utilizados apenas 4 páginas, no caso da dissertação, tese e livros, são recuperados informações, quase sempre, de capa, contra capa e resumo. No caso do artigo, como é um documento com menos páginas, essas quatro páginas já possui o texto do mesmo. Com isso, caso o número de caracteres extrapolem 1000 (número obtido por meio de observações), a probabilidade desse documento ser um artigo se potencializa.

Por fim, foi gerado a página HTML como ilutrado na Figura 1. Como os arquivos não são rotulados, não é possível obter um número exato de acurácia. Porém, em uma inpeção por amostragem, essa abordagem funciona bem sendo capaz de classificar praticamente todos os arquivos encontrados de maneira correta. O únicos problemas de classificação encontrados foram arquivos que não estão completos, por exemplo, um texto que inicia sem título ou capa e nenhuma identificação de autor. Nesses casos, nem um ser humano seria capaz de dizer com certeza a classe na qual o mesmo seria colocado. Além disso, outro problema encontrado é no modo de permissão do PDF. Alguns deles estão protegidos contra leitura. Nesse caso, o código não é capaz de alterar essa permissão, o que acarreta no não carregamento do mesmo e a não classificação desse arquivo.