ISSN 2236-8884

RetiraAcentuação e PostCounter: Ferramentas de otimização e preparação de arquivos

Yasmmin C. Martins, Breno F. T. Azevedo, Hélvia P. P. Bastos

Núcleo de Informática na Educação (NIE)
Instituto Federal Fluminense Campus Campos-Centro (IFF) – Campos dos Goytacazes, RJ – Brasil
{helviabastos, brenoter}@gmail.com, nim_asay@hotmail.com

Abstract: *This paper describes tools created to optimize text mining activities. Their development aimed at preparing files to be analyzed by the software "Presente!" [Kambara-Silva 2011]. In spite of being structurally simple, their application allows for more flexibility in text mining, as well as other applications.*

Resumo: O trabalho descreve duas ferramentas criadas para otimizar atividades de mineração de textos. Ambos os recursos foram desenvolvidos para preparar arquivos a serem analisados pelo software "Presente!" [Kambara-Silva 2011]. Embora sejam estruturalmente simples, sua aplicação otimiza e flexibiliza atividades de mineração de textos, além de permitir outras aplicações.

1. Introdução

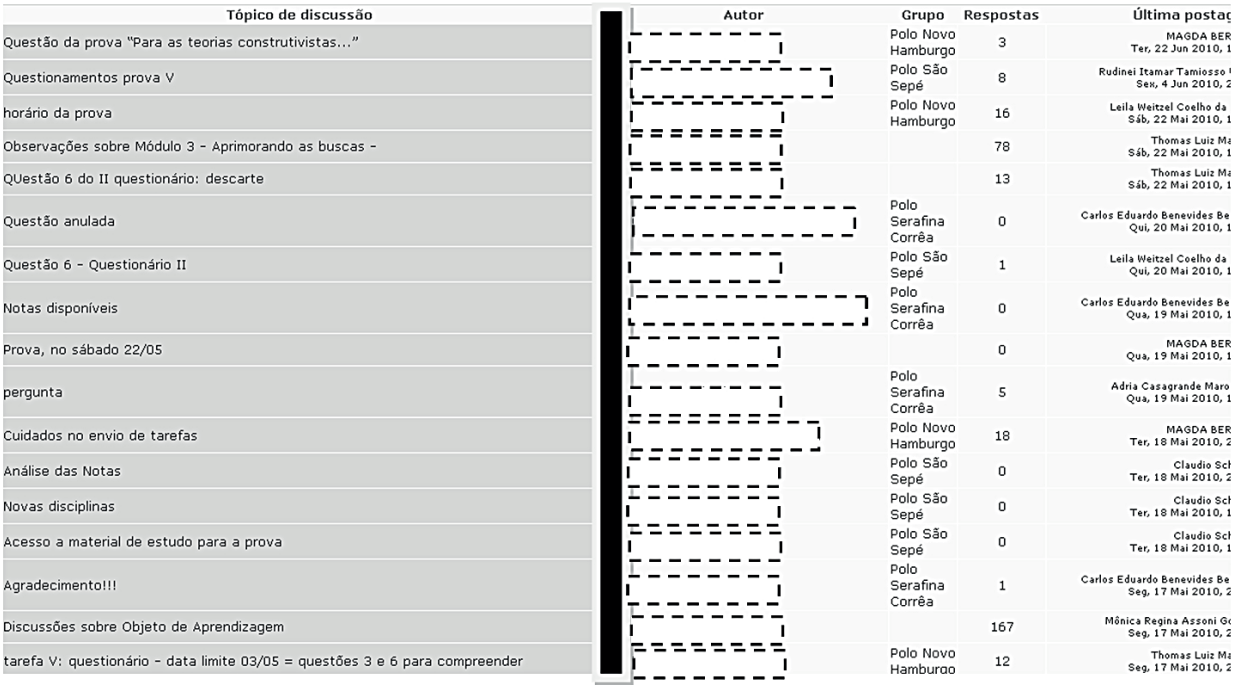
As ferramentas foram desenvolvidas para otimizar a preparação e verificação dos arquivos a serem processados e analisados pelo programa *Presente!* [Kambara-Silva 2011], assim como tornar mais rápidas determinadas atividades necessárias na organização dos arquivos de entrada.

O PostCounter pode ser usado para calcular o total de tópicos e postagens de um fórum de discussão. A ferramenta também fornece valores estatísticos que podem ser úteis na apresentação dos resultados finais.O RetiraAcentuação pode ser usado para qualquer arquivo de texto, com extensões como HTML, XML, TXT, DOC etc. O objetivo deste software é remover a acentuação das palavras existentes nos arquivos.

2. PostCounter

ISSN 2236-8884

Este software foi desenvolvido com o objetivo de realizar a contagem automática de tópicos em fóruns de discussão e a quantidade de postagens feitas nos mesmos. Essa função evita a ocorrência de erros e a necessidade de realizar tal operação repetidas vezes (caso a contagem fosse feita manualmente).



Tópico de discussão	Autor	Grupo	Respostas	Última postag
Questão da prova "Para as teorias construtivistas..."		Polo Novo Hamburgo	3	MAGDA BER Ter, 22 Jun 2010, 1
Questionamentos prova V		Polo São Sepé	8	Rudinei Itamar Tamiosso Sex, 4 Jun 2010, 2
horário da prova		Polo Novo Hamburgo	16	Leila Weitzel Coelho da Sáb, 22 Mai 2010, 1
Observações sobre Módulo 3 - Aprimorando as buscas -			78	Thomas Luiz Ma Sáb, 22 Mai 2010, 1
QUestão 6 do II questionário: descarte			13	Thomas Luiz Ma Sáb, 22 Mai 2010, 1
Questão anulada		Polo Serafina Corrêa	0	Carlos Eduardo Benevides Be Qui, 20 Mai 2010, 1
Questão 6 - Questionário II		Polo São Sepé	1	Leila Weitzel Coelho da Qui, 20 Mai 2010, 1
Notas disponíveis		Polo Serafina Corrêa	0	Carlos Eduardo Benevides Be Qua, 19 Mai 2010, 1
Prova, no sábado 22/05			0	MAGDA BER Qua, 19 Mai 2010, 1
pergunta		Polo Serafina Corrêa	5	Adria Casagrande Maro Qua, 19 Mai 2010, 1
Cuidados no envio de tarefas		Polo Novo Hamburgo	18	MAGDA BER Ter, 18 Mai 2010, 2
Análise das Notas		Polo São Sepé	0	Claudio Sch Ter, 18 Mai 2010, 2
Novas disciplinas		Polo São Sepé	0	Claudio Sch Ter, 18 Mai 2010, 1
Acesso a material de estudo para a prova		Polo São Sepé	0	Claudio Sch Ter, 18 Mai 2010, 1
Agradecimento!!!		Polo Serafina Corrêa	1	Carlos Eduardo Benevides Be Seg, 17 Mai 2010, 2
Discussões sobre Objeto de Aprendizagem			167	Mônica Regina Assoni Gc Seg, 17 Mai 2010, 2
tarefa V: questionário - data limite 03/05 = questões 3 e 6 para compreender		Polo Novo Hamburgo	12	Thomas Luiz Ma Seg, 17 Mai 2010, 2

Figura 1: Arquivo de entrada para o contador

A Figura 1 mostra uma parte da página web que o PostCounter utiliza, com vários tópicos e a quantidade de postagens representada pelos números na coluna "respostas". Com esse arquivo HTML, o programa faz uma varredura em todo o documento e armazena as linhas do HTML como

ISSN 2236-8884

texto codificado (no formato UTF-8). Em seguida, ele busca por determinadas tags que marcam cada repetição de blocos semelhantes com as informações requeridas, como se verifica nas colunas 1 e 4 da figura.

Com isso, após processar todos os arquivos HTML dentro de uma pasta comum, ele fornece as respostas na área de texto logo abaixo do botão "Contar postagens", organizando a ordem de acordo com a sequência dos arquivos que estavam no diretório escolhido.

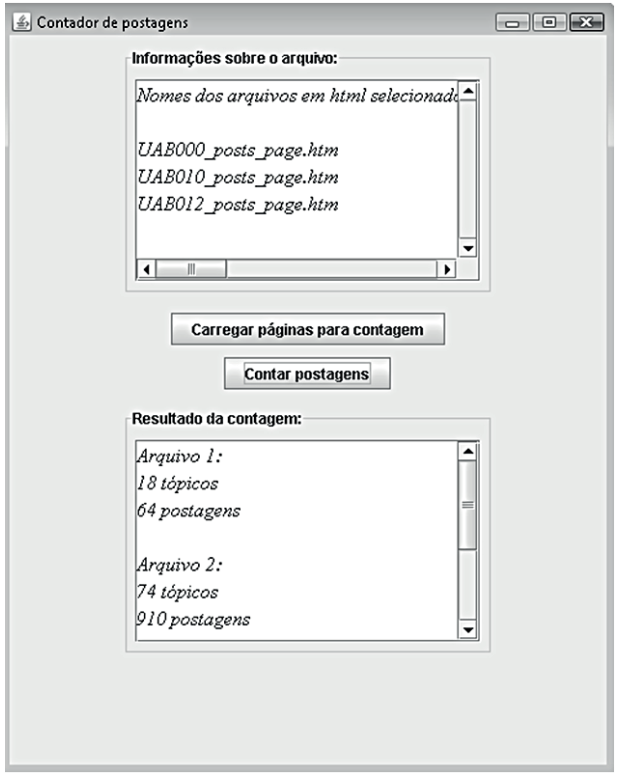


Figura 2: Ilustração do software PostCounter com os arquivos de entrada e os resultados dos mesmos

ISSN 2236-8884

Na figura acima, pode-se verificar os resultados apresentados pelo software. Um dos arquivos de fórum possuía mais de 900 postagens. O índice de erro que poderia haver caso a contagem fosse realizada manualmente poderia ser alto.

3. RetiraAcentuação

Este software foi desenvolvido para remover a acentuação das palavras existentes em um ou vários arquivos. No contexto deste trabalho, o software foi desenvolvido para analisar postagens de fóruns de discussão.

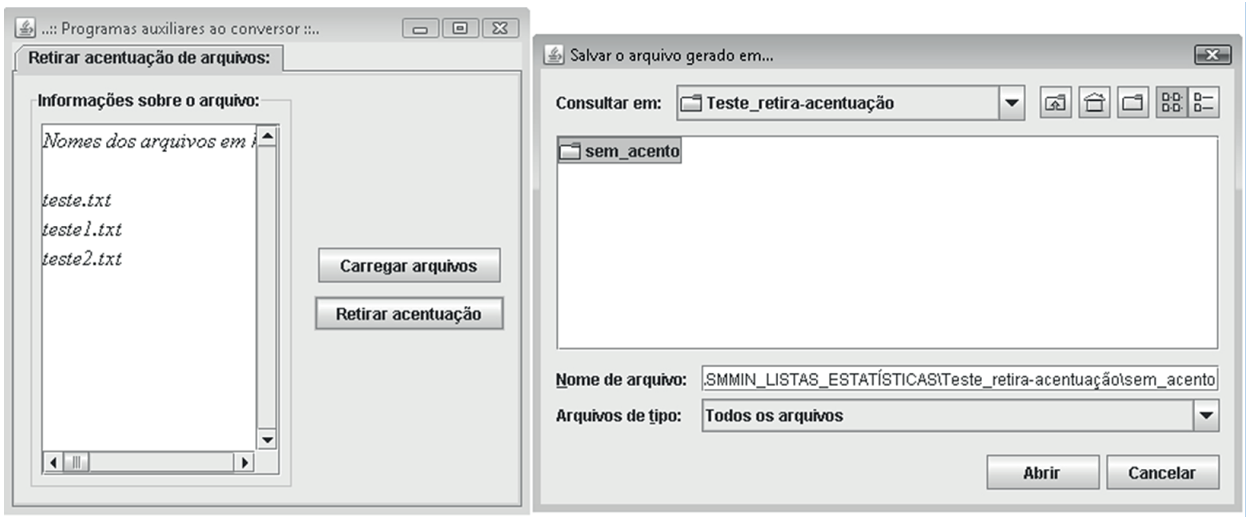


Figura 3: Ilustração do software de retirada de acentuação

A Figura 3 apresenta a interface do software. É possível escolher os diretórios tanto do arquivo de entrada quanto de saída. Após o processamento dos dados de entrada, um arquivo de saída será gravado com o resultado da operação.

ISSN 2236-8884

4. Conclusão

Os programas descritos neste trabalho são relativamente simples em sua aparência e funcionamento. Entretanto, ajudaram a otimizar e tornar mais rápidas algumas etapas da preparação dos arquivos de entrada da ferramenta "Analisador" do software "Presente!".

Referências

Kambara-Silva, J.K. (2011) "Automatização do processo de aquisição de presença social em fóruns e chats". Trabalho de Conclusão de Curso. Instituto de Informática. Universidade Federal do Rio Grande do Sul. Porto Alegre.