Processamento Estruturado de Documentos

LMCC & LESI, Universidade do Minho

Ano lectivo 2000/2001

Ficha Teórico-Prática Nº3

José Carlos Ramalho

18 de Outubro de 2000

1 Transformação de Documentos

1.1 *Partindo do documento HTML "Publicações do jcr"disponível no site da disciplina, que está escrito em SGML (HTML é SGML) faça as alterações necessárias para o converter para XML (sugestão: utilize o sx):*

- *Faça o download do bib.html, do html.dtd,e do docbook.dcl.*
- *Acrescente a declaração do tipo de documento à página HTML.*
- *Aplique o s2x à página e analise o resultado.*
- *Utilize o switch* $-biso-8859-1$ *com o s2x, analise os resultados.*
- *Concatene a declaração do docbook à cabeça da página e experimente de novo.*
- *Se o resultado ainda não fôr satisfatório desligue as 3 variáveis de ambiente do XML.*
- *Valide o resultado - bib.xml - com o nsgmls ou gerando uma script com o XML::DT para o novo documento.*

2 Text Mining

2.1 *Utilize a script desenvolvida na última aula para calcular as estatísticas de anotação do índice da gaveta das cartas que utilizou na última aula.*

Solução

2.2 *Use a script atrib − id para acrescentar um atributo ident a cada elemento. Grave o resultado num novo ficheiro de nome cartas − id.xml. Verifique se o novo ficheiro é um documento XML. Caso contrário altere a script para que assim seja.*

2.3 *Crie uma nova script de análise que para cada etiqueta vai indicar a lista de identificadores associados a essa etiqueta. Produza uma tabela HTML com os resultados da análise:*

```
<TABLE BORDER=1>
  <TR><TH>Elemento<TH>Lista de identificadores
  <TR><TD>...     <TD> ...
  ...
</TABLE>
```

Teste a script nos vários documentos.

2.4 *Crie uma script que dado um documento XML cria um índice de palavras em que a cada palavra, associa o path na árvore documental e o valor do atributo ident.*

```
#!/usr/bin/perl
use XML::DT ;
my $filename = shift;


%handler=(
     '-outputenc' => 'ISO-8859-1',
     '-default'   => sub{ for $pal (split( /\s+/, $c ))
                          { push( @{$ind{$pal}}, [join('/',@dtcontext), $v{ident}] );
                          }; ""},
);
dt($filename,%handler);


for (sort keys %ind){
  print "$_\n";
  foreach (@{$ind{$_}})
    {
     print "    $_->[0] :: $_->[1]\n";
    }
}
```