Processamento de Linguagens (3º ano de Curso)

Trabalho Prático Nº1

Relatório de Desenvolvimento

Diogo Braga	João Silva	Ricardo Caçador
A82547	A82005	A81064

31 de Março de 2019

Resumo

Neste relatório é apresentada a resolução de um exercício referente ao TP1, que tem como objetivos a utilização de Expressões Regulares para descrição de padrões de frases, e a utilização do Flex para gerar filtros de texto em C. Outro objetivo será ainda desenvolver, a partir de ERs, sistemática e automaticamente Processadores de Linguagens Regulares, que filtrem ou transformem textos com base no conceito de regras de produção Condição-Ação.

Conteúdo

Capítulo 1

Introdução

Seguindo a fórmula *exe= (N_Alu% 7) + 1* e o número de aluno mais baixo presente no nosso grupo (81064), o enunciado correspondente é o **5 - WikiQuotes: provérbios**.

Este enunciado apresenta-nos um tema relacionado com várias citações de vários autores, vários provérbios de línguas diferentes, e ainda provérbios que são variações dos originais e alguns que são adulterações dos originais. Este tipo de material está armazenado num ficheiro *XML*, onde se encontra muitas vezes organizado e outras vezes não tão bem organizado.

Ao longo deste trabalho produzimos essencialmente 3 filtros cada um com um objetivo bem delimitado. O primeiro filtro atua sobre as citações contidas em páginas de autor, o segundo filtro retém todas as citações cujo título da página comece com "Provérbios"e, por último, o terceiro filtro procura um padrão de provérbios que estejam adulterados. Neste último caso optamos também por filtrar os provérbios que são variações. De referir ainda que ligamos os provérbios adulterados e os que são variações ao respetivo provérbio original. O pergunta 4 do enunciado, o grupo optou por não a fazer separadamente dos outros filtros pelos que cada filtro apresenta estatísticas sobre o seu trabalho.

Com este relatório pretendemos apresentar as nossas opções, algoritmos desenvolvidos e ainda estruturas utilizadas para a realização de cada filtro. Pretendemos também apoiar aquelas que foram as nossas soluções, com conhecimento obtido nas aulas teóricas, relembrando por exemplo o caso dos autómatos.

Para uma melhor visão do que irá ser abordado neste relatório deixamos uma breve descrição daquilo que foi feito. No segundo capítulo foi feita uma análise informal e uma especificação dos requisitos deste projeto. No terceiro capítulo foi realizado o desenho da conceção no qual estão envolvidos os algoritmos e estruturas de dados usados. No quarto capítulo mostramos alguns exemplos de implementações e vários resultados de testes realizados. Por último no capítulo 5 fazemos uma retrospetiva do trabalho realizado e concluímos.

Capítulo 2

Análise e Especificação

Analisando o problema como um todo o que podemos encontrar aquando da observação do ficheiro **ptwikiquote-20190301-pages-articles.xml**, são várias meticulosidades no que toca à organização das páginas, títulos, citações, provérbios, etc. Contudo o nosso objetivo central é filtrar somente o que achamos necessário para respeitar os requisitos impostos pelo enunciado e descritos nas subsecções seguintes.

Nas secções seguintes serão apresentados os objetivos de cada alínea do exercício e ainda as observações que foram feitas ao ficheiro que contém todo o *XML*, por forma a pensar que casos iríamos ter futuramente e começarmos a delinear uma arquitetura duma possível solução.

2.1 Análise e Especificação dos Requisitos

2.1.1 Lista de Citações

Na primeira alínea do exercício, era requerido um filtro que criasse uma lista de citações com o respetivo autor, somente se estas citações se encontrarem numa página de autor.

Ao proceder à análise do ficheiro *XML* reparamos que uma página seria de autor se contivesse o seguinte cenário dentro da mesma:

```
{{Autor
| Nome        = George W. Bush
| Foto        = George-W-Bush.jpeg
| Wikisource  =
| Wikipedia   = George W. Bush
| Wikicommons = George Walker Bush
| Gutenberg   =
| Cervantes   =
| DominioPu   =
| DomiPubli   =
| EbooksG     =
| Cor         = #c0c0c0
}}
```

Figura 2.1: Estrutura descritiva de um autor.

Obviamente existem páginas que contém informações mais volumosas sobre um autor e outras páginas que contém menos. Posto isto reparamos que o campo **Nome**, pode ou não estar preenchido. Para os casos em que não está preenchido, o grupo achou por bem não considerar que a página se trataria de um autor, ignorando todos as citações na página contidas. O campo **Nome** possuía ainda outro pormenor. Muitas vezes surge referenciado noutra língua como por exemplo espanhol (*Nombre*) e inglês (*Name*).

Em páginas de autor investigámos de que forma aparecem a maior parte das citações que estes fazem. Chegamos à conclusão que na maior parte dos casos as citações surgem da seguinte forma:

```
*&quot;Nossa eterna mensagem de [[esperança]] é que a aurora chegará.&quot;
::- ''Our eternal message of hope is that dawn will come.
:::- ''A Martin Luther King treasury<200e> - Página 182, Martin Luther King
```

Figura 2.2: Estrutura duma citação.

Conseguimos identificar que a maior parte das citações de autores se encontram no meio de **"** e apenas teríamos de filtrar o que estivesse entre duas marcas deste tipo. Mas com a análise de mais casos deparámo-nos com situações desta natureza:

```
* &quot;    Uma coisa &quot;é&quot; de uma maneira, mas &quot;não é&quot; de maneiras infinitas.    &quot;
::- &quot;  Gente que diz coisas    &quot;, Mixórdia de Temáticas   23-02-2012
```

Figura 2.3: Exemplo de quatro marcas dentro duma citação.

Contudo surgiram muitos mais casos durante esta análise podemos referir casos em que citações não terminam com a marca **"** mas em vez disso acabam com por exemplo **::-**, ou **:-**, e ainda ******.

Estes casos e muitos mais serão tratados nos capítulos seguintes.

2.1.2 Lista de Provérbios

Nesta segunda alínea do exercício, era requerido um filtro que criasse uma lista de provérbios (citações contidas em páginas cujo título começa por "Provérbios").

Procedendo à análise do ficheiro *XML* foi possível entender a forma como se organiza a informação nas páginas que referem provérbios:

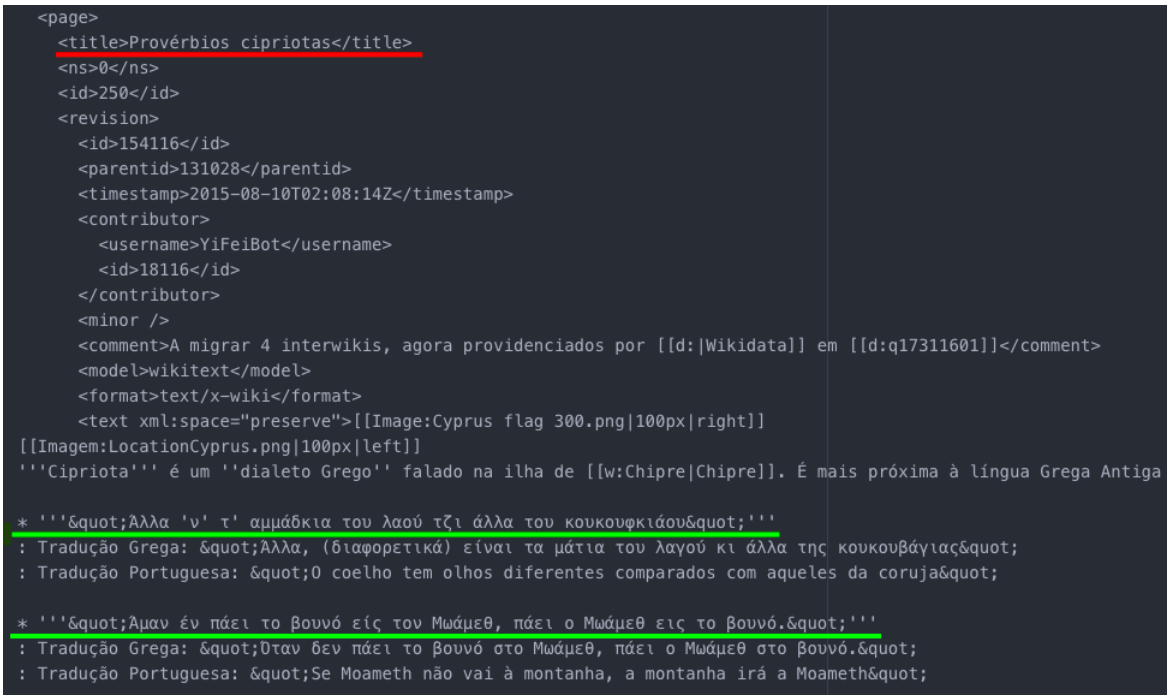


Figura 2.4: Exemplo de uma página de provérbios.

Neste caso, está exemplificada a página na qual o título é **Provérbios cipriotas**, visto que este se encontra entre as marcas <title> e </title>. É, portanto, desta forma que são identificadas as páginas que temos interesse em aceder. Tal facto aparece sublinhado a vermelho na figura.

Aquando da análise do ficheiro foram encontrados casos em que o título continha a palavra "Provérbio", mas não era inicializado por esta. É importante, de facto, ter em conta este caso pois tal é requirido no enunciado do exercício.

Filtrando os títulos das páginas, procuramos saber quais as marcas que indicam o início de um provérbio. Neste caso, sublinhado a verde, é possível concluir que a marca de início é * '''" , enquanto a marca de fim é "'''.

Realizada a análise das páginas de provérbios, foi possível concluir que existem inúmeras formas de identificar o início e o fim dum provérbio. Tal acontece porque o ficheiro contém muita variedade de idiomas e, consequentemente, diferentes ideias das diferentes pessoas que decidem quais as marcas a usar nas citações.

Alguns exemplos de marcas de início de provérbios presentes no ficheiro são:

- *"
- *"''
- **''"
- :'''

Por outro lado, alguns exemplos de marcas de final de provérbios presentes no ficheiro são:

- "
- ""
- ""
- "'

A ideia inicial foi generalizar as expressões regulares para qualquer página de provérbios, mas devido ao facto de existirem inúmeras marcas diversificadas foi necessário criar alguns contextos específicos para cada país, de forma a que não houvessem conflitos entre marcas e fosse possível criar um output perceptível.

Nos capítulos seguintes serão abordados pormenorizadamente estes casos.

2.1.3 Lista de Provérbios Adulterados

Nesta terceira alínea do exercício, era requerida a listagem dos provérbios "adulterados"e o seu original. Era suposto procurar um padrão que identificasse estes provérbios.

Ao proceder à análise do ficheiro *XML* foi possível concluir que os provérbios adulterados e alternativos seriam apresentados da seguinte forma:

```
* &amp;quot;A cavalo dado não se olham os dentes.&amp;quot; (Brasil)
** '''Alternativos:'''
*** &amp;quot;A cavalo dado não se olha o dente.&amp;quot; (Portugal)
** '''Adulteração:'''
*** &amp;quot;A cavalo dado não se olham os dentes para não levar mordida.&amp;quot;
*** &amp;quot;A jantar dado, não se olha o molho.&amp;quot; (Portugal)
* &amp;quot;A cavalo roedor, cabresto curto.&amp;quot; (Portugal)
* &amp;quot;A chave do almoço é um bocado de pão e, a da zaragata é uma palavra.&amp;quot; (Portugal)
```

Figura 2.5: Exemplo de uma página com provérbios adulterados e alternativos.

A seleção das páginas a filtrar é realizada tal como na subsecção 2.1.2, isto é, selecionamos as páginas cujo título começa por "Provérbios". Tendo em conta este requisito, o filtro a seguir é também aplicado ao que se encontra entre as marcas <title> e </title> duma página.

Filtrando os títulos das páginas, procuramos depois saber quais as marcas que indicam um provérbio adulterado e um provérbio alternativo. Importante referir que decidimos tomar como diferentes estas qualificações.

Segundo a nossa análise, a indicação de provérbios adulterados/alternativos acontece, em regra geral, da seguinte forma:

1. É apresentado um provérbio inicializado com a marca * " e finalizado com a marca " ;
   (a) Caso a linha seguinte seja inicializada da mesma forma, o provérbio não contém adulterações;
2. É apresentada a marca ** "'Alternativos:"', ou ** "'Adulteração:"' ou ** "'Adulterados:"', que mostra que a seguir vão ser listados esses provérbios;
   (a) Caso seja ** "'Alternativos:"', é inicializada a fase dos provérbios alternativos;
   (b) Caso seja ** "'Adulteração:"' ou ** "'Adulterados:"', é inicializada a fase dos provérbios adulterados;

3. É apresentado um provérbio inicializado com a marca *** " e finalizado com a marca " , que é o alternativo ou o adulterado da citação apresentada no passo 1.

Na figura 2.5, é possível visualizar um caso destes:

- Provérbio original que contém adulterações ⇒ "A cavalo dado não se olham os dentes."
- Provérbio alternativo ⇒ "A cavalo dado não se olha o dente."
- Provérbio adulterado 1 ⇒ "A cavalo dado não se olham os dentes para não levar mordida."
- Provérbio adulterado 2 ⇒ "A jantar dado, não se olha o molho."
- Provérbio sem adulterações ⇒ "A cavalo roedor, cabresto curto"
- Provérbio indefenido (depende do que aparecer na linha seguinte) ⇒ "A chave do almoço é um bocado de pão e, a da zaragata é uma palavra."

Existe também a possibilidade de um provérbio ter só adulterados ou só alternativos. Todos estes casos serão abordados nos capítulos seguintes.

2.1.4 Estatísticas de cada filtro

Nesta quarta alínea do exercício, era requerida a apresentação de estatísticas referentes ao que foi filtrado nas alíneas anteriores.

Importante referir que o grupo tomou a decisão de apresentar tais resultados individualmente nas questões ao invés de tudo nesta parte. Esta decisão assim aconteceu porque achamos mais benéfico saber as estatísticas dos resultados em cada questão, visto ser mais fácil a sua análise.

Capítulo 3

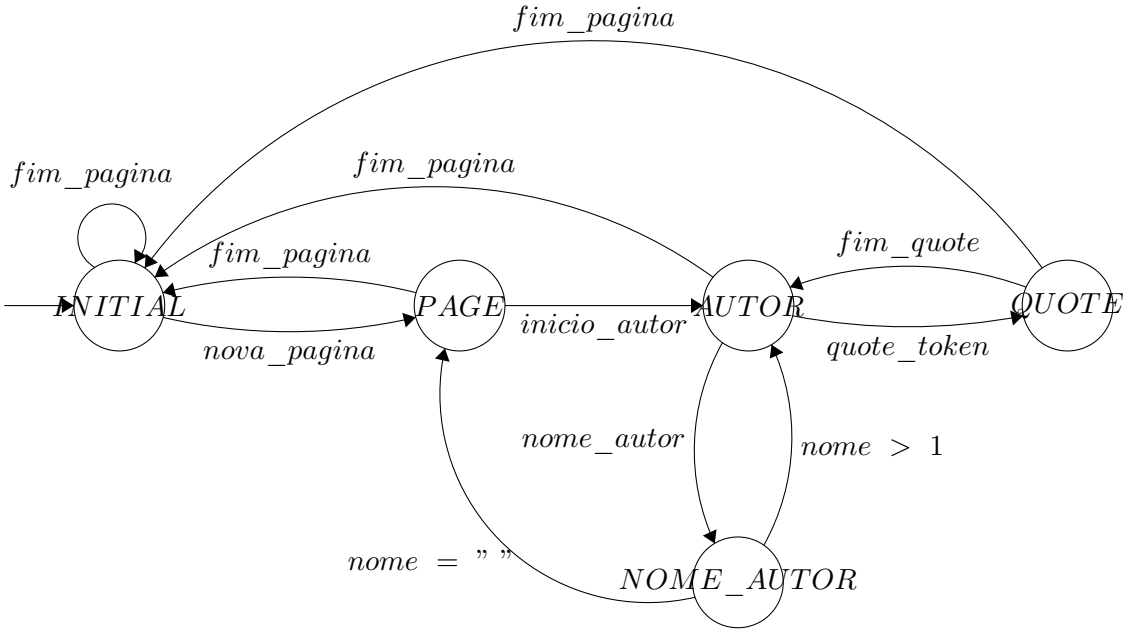
Conceção/desenho da Resolução

Neste capítulo baseando-nos nas análises feitas ao ficheiro *XML*, construímos 3 autómatos que achamos que seriam necessários para a implementação de cada um dos 3 filtros.

Os autómatos criados são uma representação de alto nível dos algoritmos implementados e serviram para criar uma base para a fase de implementação e ter uma boa noção dos contextos a utilizar. Não têm por isso, todo o detalhe que deviam uma vez que existem casos muito peculiares de implementação que não devem devem nem podem (por motivos de espaço e compreensão) ser colocados nos autómatos.

3.1 Algoritmos

3.1.1 Lista de Citações



Nota:

- O *fim_ quote* no nosso filtro é representado em muitos casos, decidimos colocar tudo como sendo *fim_ quote*.

Definições:

- espacos ⇒ [ ]*
- nova_pag ⇒ <page>
- fim_pag ⇒ </page>
- nome_linguas ⇒ (Nome|Nombre|Name)
- inicio_autor ⇒ {{(?i:Autor)
- fim_autor ⇒ }}
- nome_autor ⇒ |{espacos}{nome_linguas}{espacos}={espacos}
- quote_token ⇒ {espacos}"

Definição do autómato:

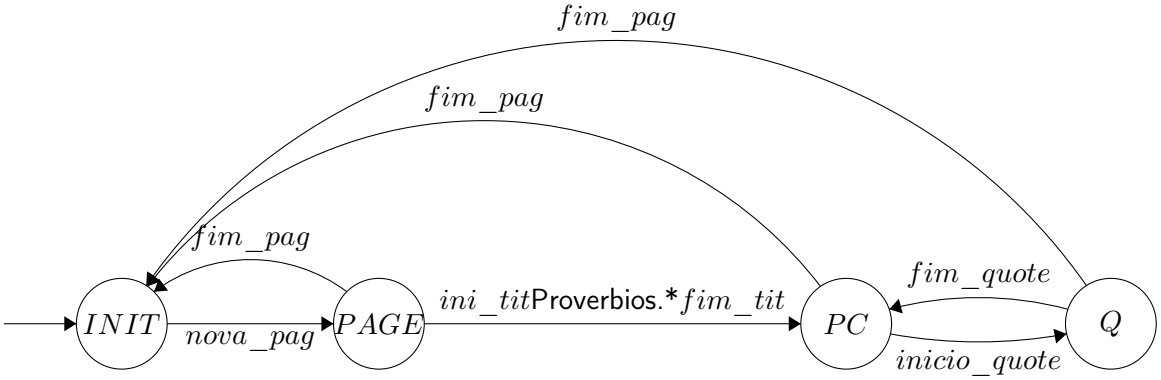
- Estados do autómato:
  - INITIAL
  - PAGE
  - AUTOR
  - NOME_AUTOR
  - QUOTE
- Estado inicial:
  - INITIAL
- Estados finais:
  - INITIAL
  - PAGE
  - AUTOR
  - NOME_AUTOR
  - QUOTE
- Funções de Transição:
  - INITIAL → $\{nova\_pagina\}$ → PAGE
  - PAGE → $\{inicio\_autor\}$ → AUTOR
  - PAGE → $\{fim\_pagina\}$ → INITIAL
  - AUTOR → $\{nome\_autor\}$ → NOME_AUTOR
  - AUTOR → $*\{quote\_token\}$ → QUOTE
  - AUTOR → $\{fim\_pagina\}$ → INITIAL
  - NOME_AUTOR → $[\wedge \setminus n] + \setminus n$ → PAGE ou AUTOR

- NOME_AUTOR → $[\wedge \setminus |] + \setminus |$ → PAGE ou AUTOR
- NOME_AUTOR → $\{fim\_pagina\}$ → INITIAL
- QUOTE → $\{fim\_quote\}$ → AUTOR
- QUOTE → $\{fim\_pagina\}$ → INITIAL

Nota:

- De referir que os estados terminais deste autómato podem ser qualquer um, uma vez que que um estado final é atingido quando um filtro atinge EOF (end-of-file).

3.1.2 Lista de Provérbios



Notas:

- O autômato aqui representado apenas apresenta a generalização de todos os casos particulares desenvolvidos para a resolução do enunciado do exercício.
- Devido a questões de perceptibilidade, as condições *inicio_quote* e *fim_quote* na representação deste autômato não são funções de transição, mas meramente indicam que nestas ações ocorrem o início e o fim de uma citação, respetivamente.

Definições:

- espaco ⇒ ( )
- espacos ⇒ [ ]*
- nova_pag ⇒ <page>
- fim_pag ⇒ </page>
- ini_tit ⇒ <title>
- fim_tit ⇒ </title>
- quote_token ⇒ {espacos}"{espacos}
- letras ⇒ [A-Za-z]+

- especiais ⇒ [

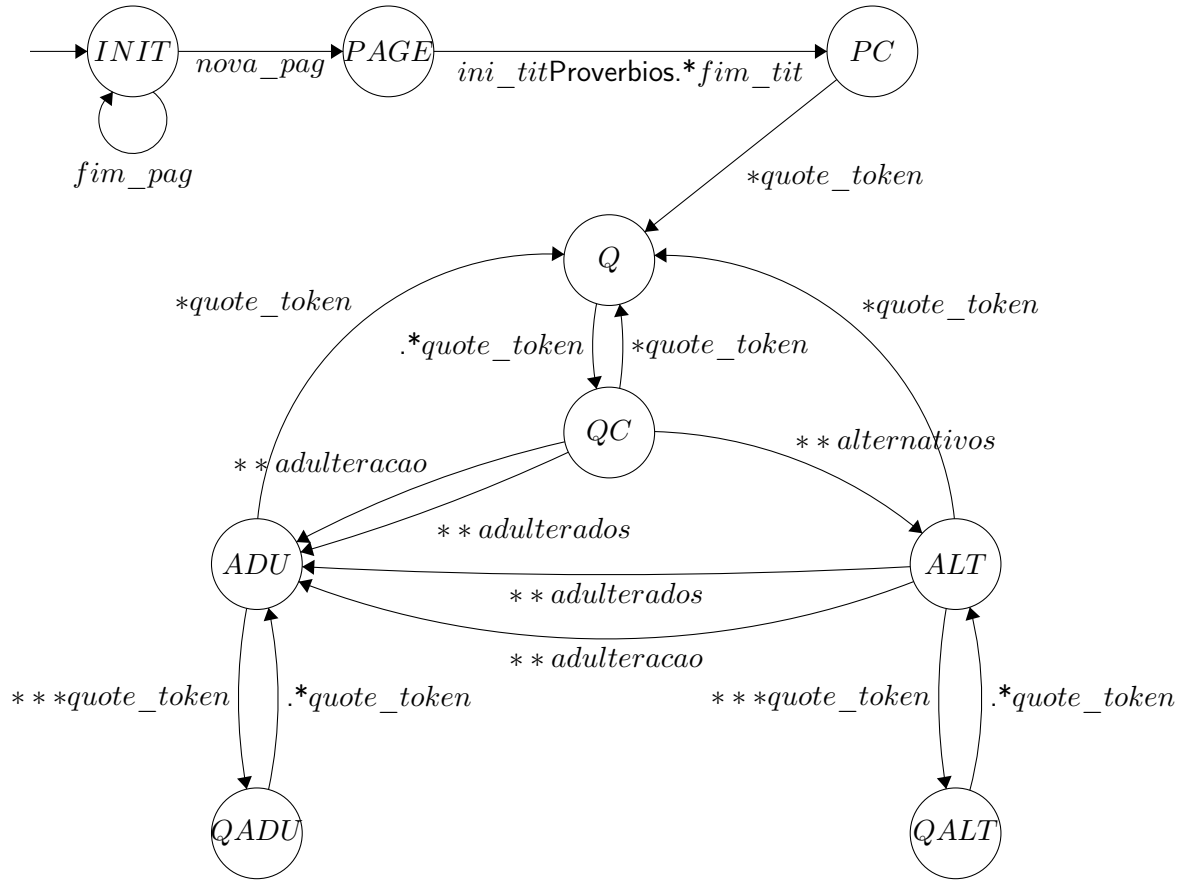
Definição do autómato:

- Estados do autómato:
  - INIT ⇒ INITIAL
  - PAGE
  - PC ⇒ PAGE_CONTEUDO
  - Q ⇒ QUOTE
- Estado inicial: INIT
- Estados finais:
  - INIT ⇒ INITIAL
  - PAGE
  - PC ⇒ PAGE_CONTEUDO
  - Q ⇒ QUOTE
- Funções de Transição:
  - INITIAL → $\{nova\_pag\}$ → PAGE
  - PAGE → $\{ini_tit\}$Proverbios.*$\{fim_tit\}$ → PAGE_CONTEUDO
  - PAGE_CONTEUDO → $inicio\_quote$ → QUOTE
  - QUOTE → .* fim_quote → PAGE_CONTEUDO
  - * → $\{fim\_pag\}$ → INITIAL

Nota:

- De referir que os estados terminais deste autómato podem ser qualquer um, uma vez que que um estado final é atingido quando um filtro atinge EOF (end-of-file).

3.1.3 Lista de Provérbios Adulterados



Nota:

- Todos os estados do autômato possuem uma condição *fim_pag* para o estado INIT. Neste autômato só está representado no próprio estado INIT.

Definições:

- espacos ⇒ [ ]*
- nova_pag ⇒ <page>
- fim_pag ⇒ </page>
- ini_tit ⇒ <title>
- fim_tit ⇒ </title>
- quote_token ⇒ {espacos}"
- adulteracao ⇒ {espacos}"'Adulteração:"'
- alternativos ⇒ {espacos}"'Alternativos:"'
- adulterados ⇒ {espacos}"'Adulterados:"'

Definição do autómato:

- Estados do autómato:
  - INIT ⇒ INITIAL
  - PAGE
  - PC ⇒ PAGE_CONTEUDO
  - Q ⇒ QUOTE
  - QC ⇒ QUOTE_CONT
  - ADU ⇒ ADULTERADOS
  - ALT ⇒ ALTERNATIVOS
  - QADU ⇒ QUOTES_ADULTERADOS
  - QALT ⇒ QUOTES_ALTERNATIVOS
- Estado inicial: INIT
- Estados finais:
  - INIT ⇒ INITIAL
  - PAGE
  - PC ⇒ PAGE_CONTEUDO
  - Q ⇒ QUOTE
  - QC ⇒ QUOTE_CONT
  - ADU ⇒ ADULTERADOS
  - ALT ⇒ ALTERNATIVOS
  - QADU ⇒ QUOTES_ADULTERADOS
  - QALT ⇒ QUOTES_ALTERNATIVOS
- Funções de Transição:
  - INITIAL → $\{nova\_pag\}$ → PAGE
  - PAGE → $\{ini_tit\}$Proverbios.*$\{fim_tit\}$ → PAGE_CONTEUDO
  - PAGE_CONTEUDO → $*\{quote\_token\}$ → QUOTE
  - QUOTE → .* $*\{quote\_token\}$ → QUOTE_CONT
  - QUOTE_CONT → $*\{quote\_token\}$ → QUOTE
  - QUOTE_CONT → $**\{adulterados\}$ → ADULTERADOS
  - QUOTE_CONT → $**\{adulteracao\}$ → ADULTERADOS
  - QUOTE_CONT → $**\{alternativos\}$ → ALTERNATIVOS
  - ADULTERADOS → $***\{quote\_token\}$ → QUOTES_ADULTERADOS
  - ADULTERADOS → $*\{quote\_token\}$ → QUOTE
  - ALTERNATIVOS → $***\{quote\_token\}$ → QUOTES_ALTERNATIVOS
  - ALTERNATIVOS → $*\{quote\_token\}$ → QUOTE
  - ALTERNATIVOS → $**\{adulterados\}$ → ADULTERADOS

- ALTERNATIVOS $\rightarrow * * \{adulteracao\} \rightarrow$ ADULTERADOS
- QUOTES_ADULTERADOS $\rightarrow$.* $*\{quote\_token\} \rightarrow$ ADULTERADOS
- QUOTES_ALTERNATIVOS $\rightarrow$.* $*\{quote\_token\} \rightarrow$ ALTERNATIVOS
- * $\rightarrow \{fim\_pag\} \rightarrow$ INITIAL

Nota:

- De referir que os estados terminais deste autómato podem ser qualquer um, uma vez que que um estado final é atingido quando um filtro atinge EOF (end-of-file).

Capítulo 4

Codificação e Testes

Para a presente secção de codificação e testes baseámo-nos nos capítulos anteriores, pois ambos são extremamente importantes para uma boa implementação desta fase. O capítulo de análise teve a sua relevância pois permitiu-nos ter já uma ideia de que expressões regulares utilizar para filtrar o que desejávamos e o capítulo da conceção da resolução foi relevante pois definimos os autómatos necessários e os contextos precisos para coordenar o processo de filtragem para cada requisito.

4.1 Estruturas de Dados

4.1.1 Lista de Citações

Quando começamos a desenvolver os primeiros filtros para testar com o ficheiro de input, deparamo-nos com a situação de existir mais que uma página para um mesmo autor. Seria portanto ideal que armazenássemos todas as citações referentes a um autor numa **Tabela de Hash** e íamos inserindo a esta as citações que aparecessem em diferentes páginas referentes a um mesmo autor.

Adotanto esta abordagem temos que filtrar tudo o que é necessário e só no fim despejar para o ecrã, de forma organizada, tudo aquilo que armazenamos na estrutura.

Desta forma criamos uma estrutura **TodasCitações**, que contém uma **GHashTable** cuja chave é o nome do autor e o valor é uma outra estrutura **Autor** que contém um nome e uma lista de citações. Estas estruturas são apresentadas nas seguintes imagens.

```
struct TodasCitacoes{
    GHashTable* autor_quote;
};
```

Figura 4.1: Estrutura TodasCitacoes.

```
struct Autor{
   char* nome;
   GList* quotes;
};
```

Figura 4.2: Estrutura Autor.

4.2 Alternativas, Decisões e Problemas de Implementação

4.2.1 Lista de Citações

Ao implementar esta questão começou-se por reparar que nem todas as páginas de autor estão de facto identificadas por um autor. Desta forma foi necessário ignorar todas as páginas de autor que não fossem identificadas por um autor que tivesse um nome com tamanho superior a 1.

O maior problema neste filtro foi identificar todas as formas como uma citação acabava e gerir esta questão. Encontrámos ao todo 10 formas diferentes de término duma citação e alguns dos casos são muito parecidos, de facto são especializações de casos mais gerais, pelo que a tarefa de encontrar o término duma citação ficou dificultada.

A propriedade *longest match* do **flex**, foi outro entrave à implementação uma vez que como tínhamos casos que eram especialização de outros, muitas vezes o caso que era o *longest match* não era o que era pretendido que fosse apanhado. Esta é a razão pela qual aparecem **"** no final de algumas citações.

Muita vezes fomo-nos deparando com citações que continham **"** não no final mas sim no meio das citações. Decidimos considerar que estes "tokens"fariam parte da citação uma vez que apanhámos a citação toda até uma situação de término, ignorando o conteúdo de cada citação.

Para a realização deste filtro utilizámos as estruturas descritas acima em conjunto com algumas funções auxiliares que no permitiam inserir citações, autores, e ainda imprimir todas as citações da estrutura existente.

4.2.2 Lista de Provérbios

Tendo em conta a grande variedade de marcas para representar provérbios em diferentes países, a resolução desta questão foi condicionada por tal. Desta forma, são muitos os estados criados para as mesmas partes de diferentes páginas do ficheiro.

Após a análise explicada na secção 2.1.2 foi possível concluir quais seriam os casos mais críticos. São exemplo disso os provérbios russos, que têm como marca inicial apenas *. Devido à elevada ocorrência desta marca em todo o ficheiro, foi necessário limitar a filtragem com esta marca apenas à parte do ficheiro que pretendiamos, e tal foi possível através do conceito de Estado.

Relembrando o autómato apresentado na secção 3.1.2, existem quatro fases importantes no algoritmo.

Na fase inicial vamos percorrer as páginas do ficheiro. Na segunda fase vamos interpretar o título da página e, tendo em conta o resultado obtido, o estado seguinte vai ser diferente. Existe o caso *default* que apanha tudo o que aparece em frente à palavra *Provérbios* mas, devido à grande variedade de marcas, tivemos aqui que forçar alguns processos a seguir caminhos particulares, sendo nesses casos a condição o título inteiro, como por exemplo *Provérbios russos*.

A partir deste momento o estado passa a seguir um caminho particular, que é o *CONTEUDO_RUS*. Esta é, portanto, a terceira fase. Depois de estar no conteúdo russo e serem atingidas as marcas de início de provérbio, o estado passa a ser o *QUOTE_RUS*, que vai ler a citação e, de seguida, imprimi-la no ecrã com a formatação desejada. Esta quarta fase regressa depois ao conteúdo da página de provérbios russos para continuar a sua filtragem.

Após a análise total das páginas, e de forma a não ter contextos em exagero, tentamos unificar num só contexto os países que teriam as mesmas marcas, ou marcas que fossem compatíveis. São exemplos deste facto os provérbios turcos, búlgaros, chineses e holandeses, que são compatíveis e deram origem aos estados CONTEUDO_TURQ_BUL_CHI_HOL e QUOTE_TURQ_BUL_CHI_HOL.

4.2.3 Lista de Provérbios Adulterados

A abordagem a esta questão é diferente das anteriores. Ao invês de possuir muitos casos particulares, a resolução desta questão possui antes um maior encadeamento de estados, como foi possível concluir na secção 3.1.3. Tal acontece porque estamos interessados em saber provérbios que estejam relacionados com outros provérbios.

A fase inicial de filtrar os títulos funciona tal e qual como na questão 2. Depois de estar no conteúdo duma página, o algoritmo é diferente.

Primeiro lemos o provérbio e guardamo-lo em memória, fazendo uso da função *strdup*. Se a linha seguinte for uma das marcas que indicam que vão ser listados os provérbios alternativos ou adulterados, imprimi-mos no ecrã o provérbio que tinha sido guardado. Caso contrário, continuamos a filtragem normalmente.

Caso a marca atingível seja a dos provérbios adulterados, vai ser iniciada na linha seguinte a listagem destes mesmos. Estes provérbios vão ser impressos e marcados como adulterados no output do filtro. A listagem termina quando a marca deixar de ser ***.

Tal processo acontece também com os provérbios alternativos, sendo que estes têm apenas uma diferença nos estados ao qual se dirigem. Após análise do ficheiro foi possível concluir que os provérbios adulterados, caso existam, sucedem sempre os provérbios alternativos. Desta forma é necessário, depois de listar os provérbios alternativos, verificar também os adulterados do provérbio original, caso surja a marca destes.

4.3 Testes realizados e Resultados

4.3.1 Lista de Citações

Aplicando o filtro criado para a primeira questão com o ficheiro de input disponibilizado, podem-se verificar resultados com um formato muito parecido ao da seguinte imagem.

Foram apenas imprimidas citações de autores que existam e no final é apresentada uma estatística da filtragem, que contempla o número de páginas total, o número de páginas de autor, o número de autores e o número de citações.

```
Author: Martin Luther King Junior

    - Quote: A antiga [[lei]] do [[olho por olho, dente por dente|olho por olho]] acaba por deixar todo mundo [[cego]].
    - Quote: Mesmo as noites completamente sem [[estrela]]s, podem anunciar a aurora de uma grande realização.
    - Quote: Um líder autêntico, em vez de buscar consenso, molda-o.
    - Quote: Nós não estaremos satisfeitos até que a [[justiça]] corra como [[água]] e a retidão como um caudaloso [[rio
    - Quote: A greve, no fundo, é a linguagem dos que não são ouvidos.
    - Quote: O bom vizinho olha além das circunstâncias externas e distingue aquelas qualidades intrínsecas que fazem de
    - Quote: Quase sempre minorias criativas e dedicadas tornam o mundo melhor.
    - Quote: Nossa eterna mensagem de [[esperança]] é que a aurora chegará.
    - Quote: Se um homem não descobriu algo por que morrer, ele não está preparado para viver.
    - Quote: O ser humano deve desenvolver, para todos os seus conflitos, um método que rejeite a vingança, a agressão e
    - Quote: A Verdadeira [[paz]] somente não é a ausência de tensão, é a presença de [[justiça]].
    - Quote: Nós temos que combinar a dureza da [[serpente]] com a suavidade da [[pomba]], uma [[mente]] dura e um [[cor
    - Quote: Nada no mundo é mais perigoso que a [[ignorância]] sincera e a [[estupidez]] conscienciosa.
    - Quote: O Amor é a única força capaz de transformar um inimigo num amigo.
    - Quote: Eu tenho o sonho de ver um dia meus 4 filhos vivendo numa nação em que não sejam julgados pela cor de sua p
    - Quote: Pessoas oprimidas não podem permanecer oprimidas para sempre.
    - Quote: Agradeço ao Deus todo-poderoso, nós somos livres afinal.
    - Quote: A verdadeira medida de um homem não é como ele se comporta em momentos de conforto e conveniência, mas como
```

Figura 4.3: Lista de citações de Martin Luther King Junior.

4.3.2 Lista de Provérbios

Tendo em conta a imensa variedade de marcas de início de provérbios, nesta fase os testes tiveram uma enorme importância.

O resultado do filtro está organizado por Título e Citações, sendo que no final é apresentada uma estatística da filtragem, que contempla o número de páginas total, o número de páginas com título começado por 'Provérbios', e o número de citações recolhidas no filtro.

Seguem-se alguns exemplos de provérbios obtidos da filtragem.

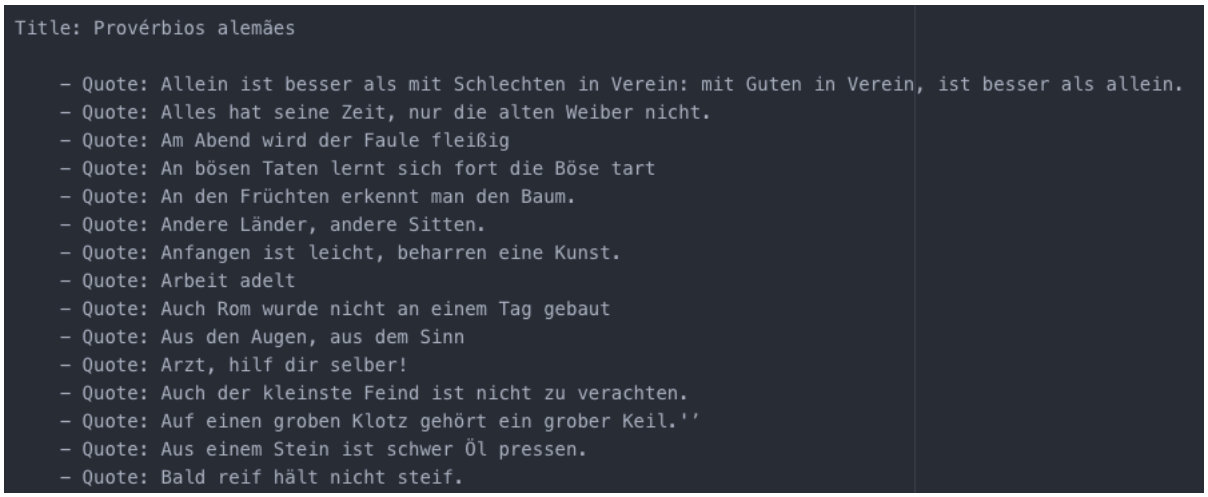


Figura 4.4: Parte do resultado obtido para os provérbios alemães.

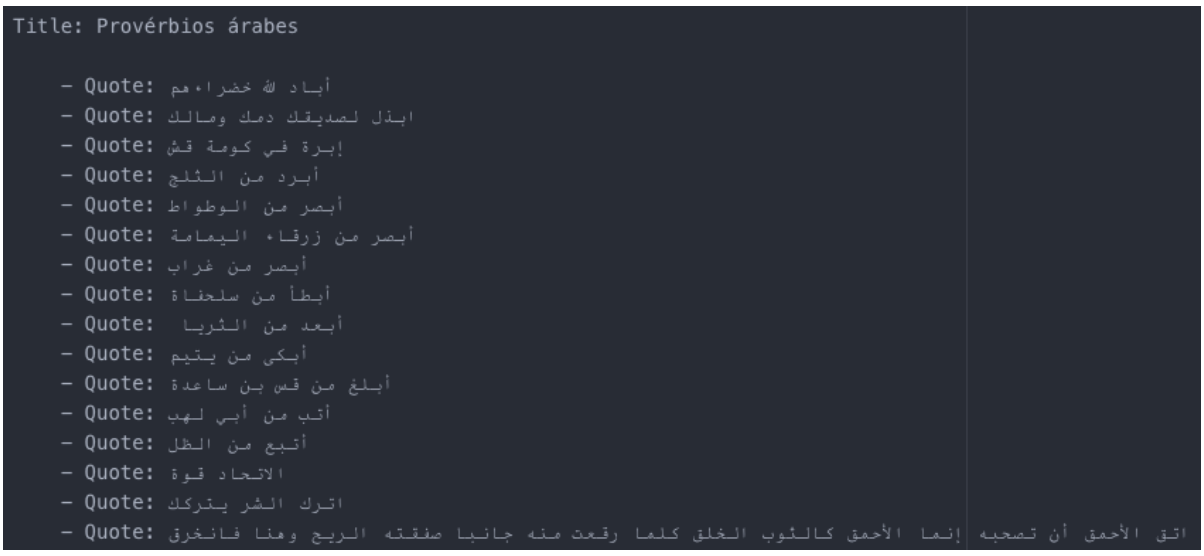


Figura 4.5: Parte do resultado obtido para os provérbios árabes.

4.3.3 Lista de Provérbios Adulterados

O resultado do filtro está organizado por Título e Citações (com os seus adulterados), sendo que no final é apresentada uma estatística da filtragem, que contempla o número de páginas total, o número de páginas com título começado por 'Provérbios', o número de adulterações, o número de alternativos, e o número de citações recolhidas no filtro.

Seguem-se alguns exemplos de provérbios obtidos da filtragem.

```
Title: Provérbios em Português

    - Quote: A cabeça não se fez só para usar chapéu.
        - Alternativo: A cabeça não serve só para usar boné.
        - Adulterado: A cabeça não serve só para criar piolhos.
        - Adulterado: A cabeça não serve só para separar as orelhas.
    - Quote: A carne é fraca.
        - Adulterado: A carne é fraca mas o molho é muito bom.
        - Adulterado: A carne é fraca mas o pecado não é vitamina.
    - Quote: A cavalo dado não se olham os dentes.
        - Alternativo: A cavalo dado não se olha o dente.
        - Adulterado: A cavalo dado não se olham os dentes para não levar mordida.
        - Adulterado: A jantar dado, não se olha o molho.
    - Quote: A fome é o melhor tempero.
        - Alternativo: A fome é o melhor condimento.
        - Alternativo: A fome é a melhor amiga do cozinheiro.
    - Quote: A galinha da minha vizinha é mais gorda que a minha
        - Alternativo: A galinha da minha vizinha é sempre melhor que a minha
    - Quote: A ignorância é a mãe de todos os erros.
    - Quote: A ignorância é a mãe do atrevimento.
        - Adulterado: A ignorância é a mãe de todas as doenças mas é um repousante.
    - Quote: A ignorância é mãe de todos os (vícios|doenças).
        - Alternativo: A ociosidade é mãe de todos os (vícios|doenças).
        - Alternativo: A pregiça é mãe de todos os (vícios|doenças).
    - Quote: À mulher de César não lhe basta ser séria.
        - Alternativo: A mulher de César tem que ser, não basta parecer.
    - Quote: A necessidade é mestra de engenhos.
        - Alternativo: A fome aguça o engenho.
        - Alternativo: A necessidade aguça o engenho.
```

Figura 4.6: Parte do resultado obtido para os provérbios adulterados.

4.3.4 Estatísticas de cada filtro

As seguintes imagens representam os resultados de cada filtro realizado, que aparecem no final de cada filtragem.

```
Número de páginas: 13852
Número de páginas de autor: 4512
Número de autores: 3737
Número de citações: 18846
```

Figura 4.7: Estatísticas relativas ao primeiro filtro.

```
Número de páginas: 13852
Número de páginas com título começado por 'Provérbios': 67
Número de citações: 4106
```

Figura 4.8: Estatísticas relativas ao segundo filtro.

```
Número de páginas: 13852
Número de páginas com título começado por 'Provérbios': 67
Número de adulterações: 109
Número de alternativos: 190
Número de citações com adulterações ou alternativos: 191
```

Figura 4.9: Estatísticas relativas ao terceiro filtro.

Capítulo 5

Conclusão

Tendo em conta os requisitos deste projeto, e o trabalho realizado pelo grupo, achamos que os objetivos fundamentais foram atingidos, sendo estes a capacidade de criar padrões com uso de **Expressões Regulares**, o entendimento da utilização da ferramenta **flex** para a criação destes padrões, a capacidade de analisar ficheiros de entrada e criar algoritmos de resolução recorrendo a **autómatos**.

Ao longo da realização deste projeto o grupo encontrou várias dificuldades, estando estas relacionadas maioritariamente com a quantidade de formas diferentes em que apareciam as citações e os provérbios a serem filtrados. Entendemos portanto, que esta foi a maior dificuldade pois foi necessário a agilização dos membros do grupo para encontrarem soluções para a filtragem de certos padrões.

Em jeito de conclusão o grupo acha que todas as alíneas requisitadas no enunciado foram terminadas com sucesso, e os objetivos principais deste projeto foram atingidos.