



# Processamento de Linguagens

MIEI - 3º ANO - 2º SEMESTRE  
UNIVERSIDADE DO MINHO

## TRABALHO PRÁTICO Nº1 FLEX



Pedro Freitas  
A80975



Francisco Freitas  
A81580

30 de Março de 2019

# *Conteúdo*

<b>1</b>	<b>Contextualização</b>	<b>2</b>
<b>2</b>	<b>Wiki Quotes: autores</b>	<b>3</b>
2.1	Enunciado . . . . .	3
2.2	Descrição do problema . . . . .	3
2.2.1	1) . . . . .	3
2.2.2	2) . . . . .	4
2.2.3	3) . . . . .	4
2.3	Decisões e implementação . . . . .	4
2.3.1	1) . . . . .	4
2.3.2	2) . . . . .	6
2.3.3	3) . . . . .	9
2.4	Como obter os resultados . . . . .	10
2.5	Resultados Obtidos . . . . .	11
2.5.1	1) . . . . .	11
2.5.2	2) . . . . .	12
2.5.3	3) . . . . .	13
<b>3</b>	<b>Apreciação Crítica</b>	<b>15</b>

# 1. *Contextualização*

Este relatório é o resultado do primeiro trabalho prático da Unidade Curricular Processamento de Linguagens. Este trabalho consistia em aplicar filtros em FLEX de forma a obter a informação que nos fosse útil. Esta filtração de texto é aplicada sobre o ficheiro XML passado como input ao executável.

De todos os enunciados possíveis vamos trabalhar sobre o enunciado **Wiki Quotes - autores**.

Ao longo do relatório vamos explicar o que nos foi pedido assim como as decisões e abordagens ao enunciado para podermos obter o resultado final pretendido.

## 2. *Wiki Quotes: autores*

### 2.1 Enunciado

Dado o ficheiro ptwikiquote-20190301-pages-articles.xml.bz2 com inúmeras páginas wiki de citações, pretendemos:

1. Criar uma lista citações (com respectivo autor se a citação estiver contida numa página de autor)
2. Criar HTML correspondente aos autores encontrados:
  - Será considerado página de autor se tiver "Autor...."
  - criar um ficheiro HTML (post-6243.html) por cada Autor (incluindo a metainformação que conseguir encontrar, biografia resumida)
  - juntar um link para o ficheiro com texto nome do autor em cada Categoria encontrada (ver categorias no final de cada *page* - exemplo: [[Categoria:Pessoas]])
3. Criar umas estatísticas dos elementos encontrados.

### 2.2 Descrição do problema

Tal como foi referido anteriormente todo este trabalho se baseia em analisar a estrutura do ficheiro passado como input de forma a encontrar padrões identificativos dos diferentes tipos de dados e assim produzir o resultado final com toda a informação que consideramos útil.

No nosso caso temos de pegar num ficheiro xml que contém inúmeras páginas wiki de citações. Neste ficheiro iremos, por exemplo, fazer procuras de **citações**, assim como os **autores** das mesmas e as **categorias** sobre as quais os autores atuam. Toda esta informação será organizada e armazenada em ficheiros *HTML*.

#### 2.2.1 1)

Neste primeiro exercício teríamos de percorrer o ficheiro *XML* e sempre que seria encontrado uma citação escrever no ficheiro *HTML*. Se esta citação tivesse dentro de uma página de autor teríamos também de colocar o respetivo autor antes da citação.

### 2.2.2 2)

Neste segundo exercício temos de percorrer outra vez o ficheiro *HTML* mas desta vez à procura de autores e criar a sua respetiva página com os seus dados encontrados no ficheiro. Da mesma forma terá de ser criada uma página *HTML* sempre que encontramos uma categoria. Dentro de cada página de categoria teremos um link para todas as páginas de autor que se enquadram nela.

### 2.2.3 3)

Neste exercício voltamos a percorrer o ficheiro *XML* mas neste caso para obtermos estatísticas que consideramos relevantes.

## 2.3 Decisões e implementação

Para resolvermos todos os problemas que nos foram apresentados temos de tomar algumas considerações gerais. Assim, depois de uma análise ao ficheiro *XML* podemos ver que cada página inicia sempre que nos é apresentado : `<page>`. Desta forma decidimos que sempre que encontramos essa tag entrarmos num novo estado `<PAGE>`.

Outra consideração geral a ter é quando estamos dentro do estado `<PAGE>` e encontramos a tag `>{{Autor "` entramos noutro novo estado, neste caso `<AUTOR>`.

Estas considerações serão aplicadas nos três exercícios assim sempre que referimos que entramos num novo estado, sabemos que encontramos a respetiva tag indicada em cima.

### 2.3.1 1)

Como foi referido na secção anterior este primeiro exercício baseia-se em criar uma página *HTML* com todas as citações presentes no ficheiro de input e os respetivos autores, caso esteja na sua página.

Assim o primeiro passo é entrar no estado `<PAGE>`. Depois de estarmos no novo estado temos duas hipóteses: ou entramos no estado `<AUTOR>` ou então encontramos uma tag que se encaixe na expressão regular:

$$[*][ ]*['"]*& \text{quot;}$$

e entramos em um novo estado: `<QUOTE>`, que nos indica que estamos na presença de uma citação. No primeiro caso em que entramos no estado do *Autor* também é possível entrar no estado *Quote* através da mesma expressão regular. Nesta situação é "levantada" uma flag que assinala que uma dada citação tem autor.

Já no estado *Quote* escrevemos a citação na página *HTML*. Caso a flag esteja levantada escrevemos o autor da citação antes da mesma.

## Hierarquia de estados

Na seguinte figura podemos ver a hierarquia e a passagem dos diferentes estados. De realçar que esta imagem **não representa** o ciclo de estados mas sim as interações entre eles.

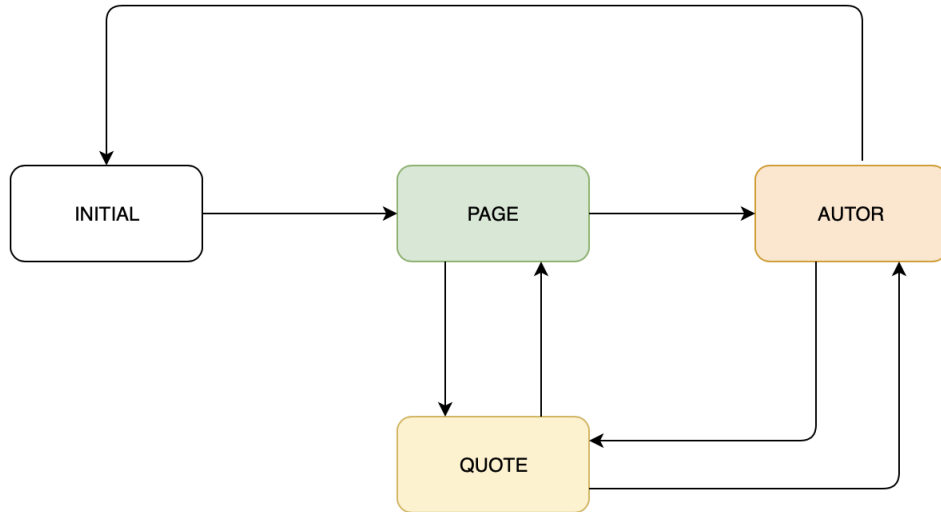


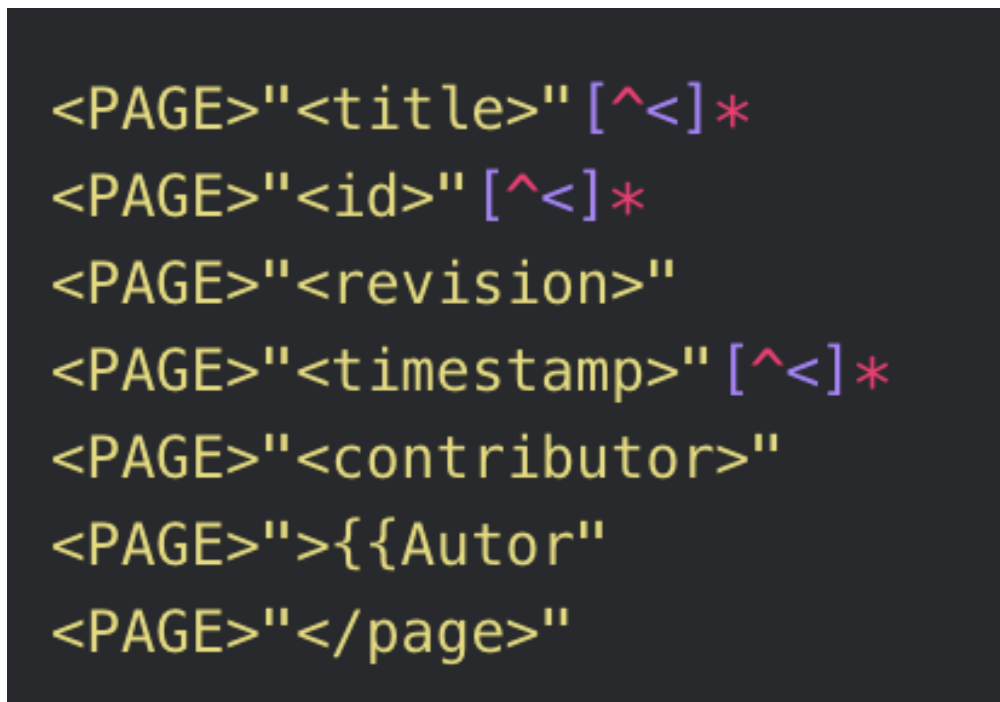
Figura 2.1: Hierarquia de estados

### 2.3.2 2)

Para este segundo problema temos de seguir um raciocínio idêntico ao do exercício anterior no que toca a entrar nos estados *Page* e *Autor*.

Para isto nós optamos por dividir o raciocínio em dois. Primeiro para tratar de criar páginas dos autores com as suas respetivas informações e depois tratar das categorias e respetivos autores que se encaixam na mesma.

Mais uma vez o primeiro passo é entrar no estado *Page* sempre que a expressão já referida nos aparece. Dentro desse estado vão aparecer-nos bastante informação, informação esta importante para a organização dos dados. Na seguinte imagem podemos ver que expressões regulares dentro do estado *Page* que nos permitiam aceder e guardar informação útil:



```
<PAGE>"<title>" [^<]*
<PAGE>"<id>" [^<]*
<PAGE>"<revision>"
<PAGE>"<timestamp>" [^<]*
<PAGE>"<contributor>"
<PAGE>">{{Autor"
<PAGE>"</page>"
```

Figura 2.2: Expressões Regulares do estado *Page*

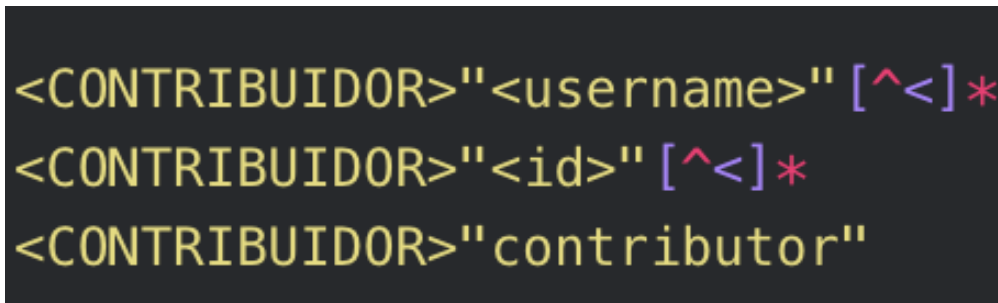
A primeira expressão regular permite-nos guardar o título da página e a segunda o ID. Com a terceira podemos criar um novo estado *Revision* que serve apenas para evitar obter informação inútil que contém as mesmas tags de informações úteis. Já a quarta expressão permite-nos criar outro novo estado para obter informações sobre o contribuidor da página(iremos especificar mais à frente). A quinta além de entrarmos no estado do *Autor* permite-nos criar a página *HTML* correspondente e imprimir os dados obtidos anteriormente assim como os dados que vão ser obtidos em outros estados. Quanto ao último, este permite sair do estado *Page*.

O estado *Revision* foi criado porque o ficheiro após esta tag encontrava um *id*, que neste caso, era uma informação inútil e repetida. Assim foi necessário a criação deste estado para ignorar esse valor. Dito isto sempre que tínhamos a seguinte expressão regular:

"<id>"

teríamos de voltar para o estado anterior, ou seja, *Page*.

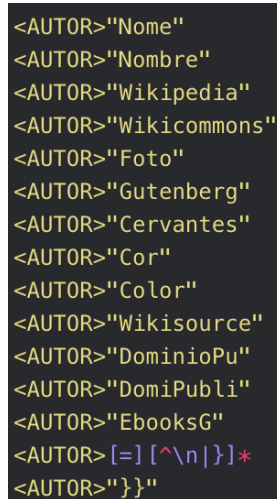
Outro estado deste exercício é o estado *Contribuidor*. Dentro deste estado iremos procurar informação como o **id** do contribuidor da página assim como o username do mesmo. Existe ainda uma expressão regular que permite voltar ao estado *Page*. Todas essas expressões regulares podem ser vistas, respetivamente, na imagem seguinte:



```
<CONTRIBUIDOR>\"username>\" [^<]*
<CONTRIBUIDOR>\"id>\" [^<]*
<CONTRIBUIDOR>\"contributor\"
```

Figura 2.3: Expressões Regulares do estado *Contribuidor*

Já no estado *Autor* temos uma maior quantidade de informação útil a obter, assim como uma expressão para entrar num estado novo: *Bio*. Neste estado podemos obter toda a informação a colocar na página, como a metainformação e biografia resumida. Todas as expressões regulares para encontrar essa informação estão na seguinte imagem, sendo que é a última que permite entrar no estado *Bio*.



```
<AUTOR>\"Nome\"
<AUTOR>\"Nombre\"
<AUTOR>\"Wikipedia\"
<AUTOR>\"Wikicommons\"
<AUTOR>\"Foto\"
<AUTOR>\"Gutenberg\"
<AUTOR>\"Cervantes\"
<AUTOR>\"Cor\"
<AUTOR>\"Color\"
<AUTOR>\"Wikisource\"
<AUTOR>\"DominioPu\"
<AUTOR>\"DomiPubli\"
<AUTOR>\"EbooksG\"
<AUTOR>\"[=] [^\n|]*\"
<AUTOR>\"}\"
```

Figura 2.4: Expressões Regulares do estado *Autor*



Agora no estado *Bio* queremos retirar toda a informação escrita nela. Por isso até encontrar a tag que indica que a biografia acabou ( "`\*`" ) vamos escrever tudo o que está lá e só depois entrar no estado *Cat*. Neste processo de retirar informação da Bio vamos encontrar parênteses retos e hifens a mais, por isso sempre que encontramos esses caracteres vamos retirá-los. As expressões regulares que nos permitem realizar toda essa seleção são: Esta última expressão é o que permite

```
<BIO>[\*]  
<BIO>[[  
<BIO>]]  
<BIO>[-]  
<BIO>( . |\n)
```

Figura 2.5: Expressões Regulares do estado *Bio*

escrever no ficheiro HTML a biografia.

No estado *Cat* procuramos a categoria correspondente para podermos guardar na árvore de categorias caso não exista ou então acrescentar o nome do autor dessa categoria. Na seguinte imagem vamos as duas expressões regulares usadas:

```
<CAT>"[" [Cc]"ategoria:" [^]]*  
<CAT>"</text>"
```

Figura 2.6: Expressões Regulares do estado *Cat*

## Hierarquia de estados

Na seguinte figura podemos ver a hierarquia e a passagem dos diferentes estados. De realçar que esta imagem **não representa** o ciclo de estados mas sim as interações entre eles.

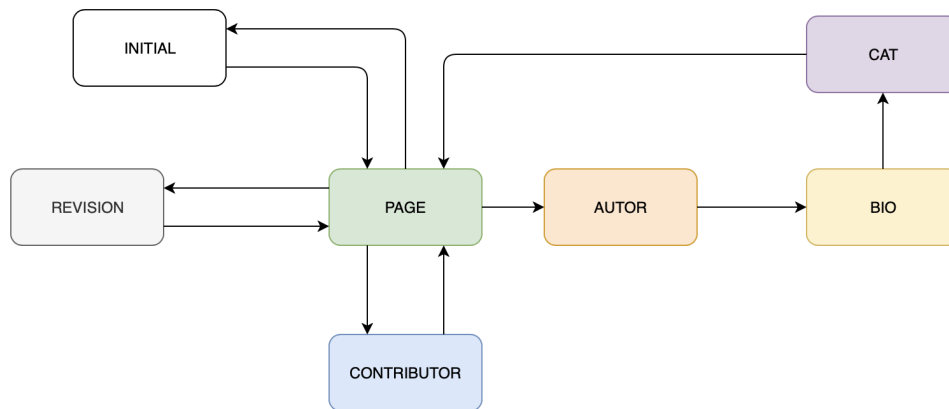


Figura 2.7: Hierarquia de estados

### 2.3.3 3)

Neste último exercício teríamos de obter estatísticas sobre o ficheiro *XML*. Para isto decidimos obter cinco informações: o número total de citações, o número de citações com autor, o número total de autores, o número de autores com citações e o número de citações de cada autor. No que toca a expressões regulares e transição de estados este exercício é idêntico ao primeiro.

Mais uma vez primeiro entramos no estado *Page*. Já dentro da *page* podemos ou guardar o nome do título da página, ou então entrar em um dos estados: *Autor* ou *Quote*. As expressões regulares encontram-se respetivamente:

```

<PAGE>\<title>[^<]*
<PAGE>">{{Autor"
<PAGE>[*] [ ]*[']*"&quot;;"

```

Figura 2.8: Expressões Regulares do estado *Page*

No estado *Autor* apenas temos duas hipóteses, ou encontramos a tag para entrarmos no estado *Quote* ou então encontramos outra tag que faz-nos voltar ao estado *Page*. As expressões regulares estão, respetivamente, apresentadas na seguinte imagem:

```

<AUTOR>[*] [ ]*[']*"&quot;;"
<AUTOR>"[" [Cc]"<ategoria:"

```

Figura 2.9: Expressões Regulares do estado *Autor*

Por último no estado *Quote* ou inserimos o autor numa árvore com todos os autores com quotes ou então voltamos para o estado *Page*. Destas duas ações apenas podem ser executadas uma delas, sendo o critério de decisão uma flag que é levantada caso no estado *Page* encontre a tag para entrar no estado *Autor*. A expressão regular é:

The image shows a regular expression `<QUOTE>[^&\n]*` on a dark background. The characters are color-coded: `<` is yellow, `QUOTE` is green, `[` is blue, `^` is purple, `&` is purple, `\` is purple, `n` is purple, `]` is blue, and `*` is red.

Figura 2.10: Expressões Regulares do estado *Quote*

## Hierarquia de estados

Na seguinte figura podemos ver a hierarquia e a passagem dos diferentes estados. De realçar que esta imagem **não representa** o ciclo de estados mas sim as interações entre eles.

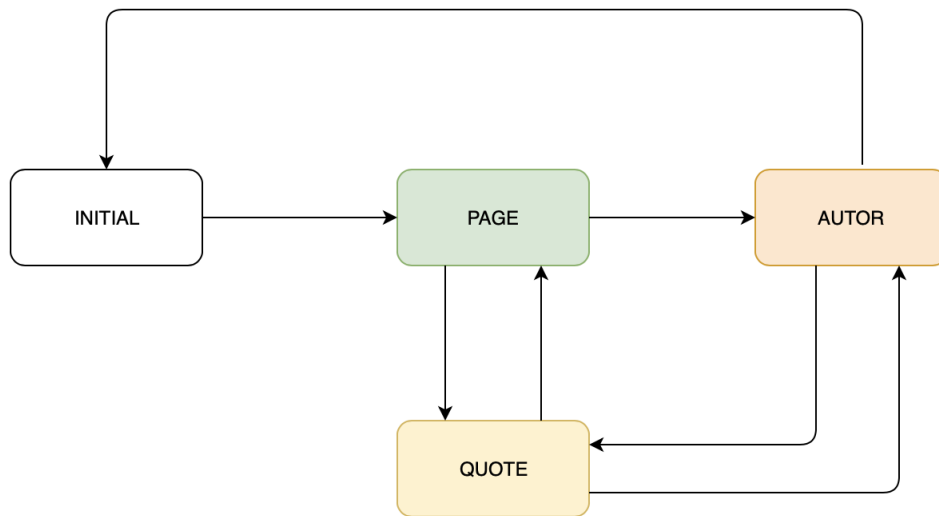


Figura 2.11: Hierarquia de estados

## 2.4 Como obter os resultados

De forma a podermos testar o trabalho final e ver os resultados de toda a implementação realizada teremos de abrir o terminal e ir para a diretoria onde está o projeto: `YOUR_PATH/PL/TP1$`. Dentro dessa diretoria veremos que temos 3 ficheiros de extensão `.l` e uma Makefile. Para correr também será necessário o ficheiro de extensão XML que podemos encontrar em: <http://natura.di.uminho.pt/~jj/pl-19/wiki/ptwikiquote/>

Para testar qualquer um dos exercícios teremos de fazer 2 passos. Primeiro fazemos **make exY**, onde Y é o exercício que pretendemos testar e apenas toma os valores 1,2 ou 3. Depois de fazermos make vemos que existem agora mais dois ficheiros, um resultado do flex (`lex.yy.c`) e o outro o executável do exercício: **exY**. Para agora executar o exercício em si basta fazer:

```
./exY < YOUR_PATH/ptwikiquote-20190301-pages-articles.xml
```

onde o `YOUR_PATH` é o caminho para a diretoria onde está o ficheiro `XML`.

Depois de executar os programas vão ser criados ficheiros *HTML*. Se quiser apagar os ficheiros resultantes de um exercício e o executável basta fazer **make clean-exY**. Para eliminar todos os executáveis e todos os ficheiros resultantes: **make clean** .

## 2.5 Resultados Obtidos

Nesta secção poderemos ver os resultados obtidos depois de implementada a nossa solução.

### 2.5.1 1)

Para o primeiro exercício o output será o ficheiro *citacoes.html* que terá todas as citações do ficheiro passado como input. Assim poderemos ter as citações com o respetivo autor ou então as citações sem autor. Nas seguintes figuras podemos ver os dois casos:

Não há luar como em Janeiro nem lenha como a de azinho, e não há filho de padre que não chame ao pai padrinho.  
A verdade ensina o caminho, mas a mentira confunde toda a gente  
A vida é louca, nela eu estou de passagem  
A abelha não leva chumbo.  
A açorda faz a velha gorda e a menina formosa.  
A água de Janeiro vale dinheiro.  
A água é tão útil às plantas como o alimento aos animais.  
A água salobra, na terra seca, é doce.  
A ambição, assim como a cólera, é muito má conselheira.  
A ambição cerra o coração.  
A amenidade no semblante, anuncia a bondade do coração.  
A amizade não se adquire, senão pela amizade.  
A apressada pergunta, vagarosa resposta.  
A aversão é para o coração, o que a prevenção é para o espírito.  
A beleza não se põe na mesa, mas eu não como no chão.  
A boa fé é uma moeda, que quase não tem curso no comércio da vida.  
A boa ventura de uns, ajuda aos outros.  
À boca da barra, se perde o navio.  
A boca do ambicioso só se fecha com terra de sepultura.  
A boca não admite fiador.  
À boda do ferreiro, cada um com o seu dinheiro.  
À boda e a baptizado, não vás sem ser convidado.

Figura 2.12: Excerto do ficheiro HTML de quotes sem autor

George Walker Bush Our enemies are innovative and resourceful, and so are we. They never stop thinking about new ways to harm our country and our people, and neither do we.

George Walker Bush A [[idéia]] de que os [[Estados Unidos]] estão se preparando para atacar o [[Irã]] é simplesmente ridícula. E tendo dito isto, todas as opções estão sobre a [[mesa]].

George Walker Bush Os [[EUA]] têm influência no [[Afeganistão]], e vamos usá-la para recordar que há [[valor]]es universais. É profundamente preocupante que um [[país]] que ajudamos a libertar queira punir alguém porque escolheu outra [[religião]]. Vamos solucionar este problema trabalhando estreitamente com o nossos contatos no [[governo]]. Trataremos do tema diplomaticamente e lembraremos às pessoas que a escolha de uma [[religião]] é algo universal

George Walker Bush Tenho vivido grandes momentos. O melhor deles foi quando pesquei uma carpa de 3,4 quilos no meu lago

George Walker Bush [[Tony Blair|Blair]], é preciso que a [[Síria]] faça o [[Hizbollah]] parar com essa m...

George Walker Bush Pessoas pobres não são necessariamente assassinas.

George Walker Bush Quando eu disse que não há negociação, quis dizer que não há negociação.

George Walker Bush Estas armas de destruição em massa têm que estar em algum lugar

George Walker Bush Não há [[liderança]], [[coragem]], programa para o [[futuro]]. É assustadora a patética inabilidade dos nossos [[senador]]es de capitalizar os [[erro]]s de George W. Bush.

George Walker Bush Vejo uma séria [[violação]] do princípio de separação entre [[Estado]] e [[Igreja]].

George Walker Bush Nossa geração não quer ser conhecida apenas pela [[guerra]] pelo [[terror]]

Getúlio Dornelles Vargas Desconfio de quem nunca me pediu nada. Geralmente, aqueles que se sentam à mesa sem apetite são os que mais comem.

Getúlio Dornelles Vargas Nada receio. Serenamente dou o primeiro passo no caminho da [[eternidade]] e saio da [[vida]] para entrar na [[história]].

Getúlio Dornelles Vargas A constituição é como as virgens. Foi feita para ser violada.

Getúlio Dornelles Vargas Não tenho inimigo de quem não possa me aproximar nem amigo de quem não possa me distanciar.

Getúlio Dornelles Vargas Vargas, esse garú vai muito longe!

Ludwig Mies van der Rohe [[Deus]] está nos detalhes

Ludwig Mies van der Rohe Menos é mais

São Vicente de Paulo Amemos a Deus, meus irmãos, amemos a Deus, mas que seja à custa de nossos braços, que seja com o suor de nosso rosto

São Vicente de Paulo A perfeição não consiste na multiplicidade das coisas feitas, mas no fato de serem bem feitas

Martin Luther King Junior A antiga [[lei]] do [[olho por olho, dente por dente|olho por olho]] acaba por deixar todo mundo [[cego]].

Figura 2.13: Excerto do ficheiro HTML de quotes com autor

NOTA: Os parênteses retos não foram retirados das quotes pois achamos, por bem, deixar na íntegra a citação presente no ficheiro *XML*.

## 2.5.2 2)

Este exercício terá vários ficheiros *HTML* como output. Terá um ficheiro para cada **autor**, outro para cada **categoria** e outro que funciona como **índice de categorias**. Na seguinte imagem podemos ver um excerto do *HTML* resultante do índice:

- [Pessoas.html](#)
- [Pessoas mortas.html](#)
- [Pessoas vivas.html](#)
- [Pessoas vivas|Jo Soares.html](#)
- [Pessoas vivas|Pessoas.html](#)
- [Pessoas |Pessoas .html](#)
- [Pessoas \\_vivas.html](#)
- [Pessoas|Aécio Neves.html](#)
- [Pessoas|Alexandre VI.html](#)
- [Pessoas|Alvares de Azevedo.html](#)
- [Pessoas|Alvaro Cunhal.html](#)
- [Pessoas|Alvaro Siza Vieira.html](#)
- [Pessoas|Amácio Mazzaropi.html](#)
- [Pessoas|Americo Vespucio.html](#)
- [Pessoas|Angela Ro Ro.html](#)
- [Pessoas|Camara Cascudo.html](#)
- [Pessoas|Cassia Eller.html](#)
- [Pessoas|Cesar Lattes.html](#)
- [Pessoas|Cicero.html](#)
- [Pessoas|Clara Tiezzi.html](#)
- [Pessoas|Claudia Jimenez.html](#)
- [Pessoas|Claudia Ohana.html](#)
- [Pessoas|Claudio Hummes.html](#)
- [Pessoas|Cleo Pires.html](#)
- [Pessoas|Eca de Queiroz.html](#)
- [Pessoas|Edouard Pailleron.html](#)
- [Pessoas|Eduardo Souto de Moura.html](#)
- [Pessoas|Emile Augier.html](#)
- [Pessoas|Emile Durkheim.html](#)

Figura 2.14: Excerto do índice de categoria em HTML

Cada um destes pontos contém uma hiperligação para outro ficheiro *HTML* que contém os nomes dos autores que atuam sobre essa categoria. Por exemplo, se seleccionarmos a categoria *Pessoas* ( seleccionar a página *Pessoas.html* teremos uma página cheia de nomes das pessoas que, segundo o ficheiro, contenham uma tag que identificam o autor nessa categoria. Assim temos um excerto desse ficheiro *Pessoas.html*:

- [Ewan McGregor](#)
- [Gerald Thomas Sievers](#)
- [Roberto Jefferson](#)
- [Roberto Campos](#)
- [Nick Mason](#)
- [Freddie Mercury](#)
- [Roger Meddows-Taylor](#)
- [John Deacon](#)
- [Otto Lara Resende](#)
- [Fernando Collor de Mello](#)
- [Ferdinand Foch](#)
- [Mano Brown](#)

Figura 2.15: Excerto do ficheiro Pessoas.html

Agora se carregarmos numa pessoa seremos redirecionados para a página dessa mesma pessoa. Por exemplo se carregarmos no nome Freddie Mercury será aberta a página correspondente como poderemos ver na seguinte imagem:

#### post-653.html

```

Id: 2913
timestamp: 2015-05-27T11:59:33Z
Contribuidor :
Username: Chico
UserId: 3
Nome: Freddie Mercury
Foto: Freddy Mercury statue in Montreux.jpg
Wikisource:
Wikipedia: Freddie Mercury
Wikicommons:
Gutenberg:
Cervantes:
DominioPu:
DomPubli:
EbooksG:
Cor: #c0c0c0

w:Freddie Mercury|'Freddie Mercury' pseudónimo de "Farookh Bomi Bulsara" (5 de setembro de 1946, Zanzibar, na w:TanzâniaTanzânia 24 de novembro de 1991, Londres, w:GrãBretanhaGrãBretanha) foi o
vocalista e líder da banda de rock britânica "Queen".

```

Figura 2.16: Página HTML sobre Freddie Mercury

### 2.5.3 3)

Neste último exercício temos como resultado o ficheiro *estatisticas.html* . Neste ficheiro iremos obter primeiro uma tabela com todos os autores que têm

quotes e a respetiva contagem de quotes de cada um, como foi anteriormente dito. Assim na seguinte imagem temos uma excerto dessa tabela:

Autor	Numero de citações
50 Cent	3
A Fantástica Fábrica de Cadáver	11
A286	5
Aaliyah	6
Aang	39
Aaron Burns	1
Abade de Jazente	1
Abdourahman Waberi	1
Abel Bonnard	3
Abel Hermant	1
Abel Salazar	1
Abgar Renault	1
Abigail Adams	2
Abraham J. Heschel	1
Abraham Lincoln	23
Abraham Maslow	2
Abraham Nicolas Amelot de la Houssaye	1
Abu Bakr al-Baghdadi	4
Abílio Diniz	3
Acelino Freitas	1
Adalgisa Colombo	1

Figura 2.17: Excerto da tabela Autores-Número de quotes

Depois da tabela ainda teremos o número de quotes com autor e total, assim como o número total de autores e autores com quotes:

**Existem 41752 quotes das quais 21527 têm autor**

**Existem no total 4463 autores dos quais 3973 têm quotes**

Figura 2.18: Excerto HTML com alguams estatísticas

### 3. *Apreciação Crítica*

Durante a realização deste primeiro trabalho prático desta unidade curricular várias foram as etapas realizadas de forma a obtermos um resultado final fidedigno e consistente.

Apesar de ter sido um trabalho relativamente simples no que toca a identificar padrões e análise de dados, este foi muito importante para podermos pôr em prática e melhorar a capacidade de escrever Expressões Regulares de forma a obtermos a informação considerada útil para a resolução de um problema. Para tal também foi necessário entendermos o processo de criação assim como a sintaxe de **FLEX**.

Tendo em conta as metas deste trabalho e os objetivos propostos no enunciado, achamos que obtivemos um resultado bastante positivo e satisfatório. Estamos também bastante satisfeitos e motivados para os desafios que no futuro esta Unidade Curricular irá nos apresentar pois pudemos aperfeiçoar os conhecimentos antes obtidos ao longo da realização deste trabalho.