Universidade do Minho

Mestrado Integrado em Engenharia Informática

3º ANO, 2º Semestre, 2018/2019

Processamento de Linguagens

Trabalho Prático - Flex

Ana Pereira
A81712

Maria Dias
A81611

May 23, 2019

Índice

1 Introdução

No âmbito da Unidade Curricular de Processamento de Linguagens, foi proposto o desenvolvimento de filtros de texto em C utilizando a ferramenta `Flex`. Para tal, aplicamos os conhecimentos adquiridos nas aulas sobre gramática e expressões regulares.

Neste relatório é apresentado o enunciado do problema, bem como todas as escolhas efetuadas pelo grupo para a formulação da solução do problema proposto. De seguida, são expostos alguns exemplos de utilização dos filtros implementados.

De acordo com o critério de atribuição de enunciados, coube ao nosso grupo resolver o enunciado número 6, "Wiki Quotes - Traduções".

2 Especificação do Problema

2.1 Enunciado - Wiki Quotes: Traduções

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

1. Criar uma lista citações (com respectivo autor se a citação estiver contida numa página de autor).

2. Criar uma tabela de traduções que conseguir encontrar. Procure um padrão para traduções (é natural que algumas traduções não sejam detetáveis).

3. Criar umas estatísticas dos elementos encontrados.

2.2 Descrição do Problema

Como podemos verificar no enunciado acima, é necessário implementar filtros de texto de modo a obter as citações, identificando o autor da mesma, e posteriormente as traduções disponíveis para essas mesmas citações. A apresentação dos dados será efetuada em HTML para que se consiga visualizar mais facilmente o resultado.

2.2.1 Criar uma lista de citações

Nesta alínea, a preocupação passa por identificar o maior número possível de citações, tentando cobrir todos os casos existentes no ficheiro fornecido. Esta tarefa torna-se especialmente desafiante, tendo em conta que existe uma variedade de padrões que identificam citações. Para além disso, é preciso notar que alguns desses padrões não são exclusivos para citações, e por isso não será possível apanhar todas as citações presentes no ficheiro sem trazer com elas conteúdo considerado descabido no contexto deste problema.

2.2.2 Criar uma tabela de traduções

Este problema pode ser considerado um caso específico de citações, em que estas se seguem de uma tradução do seu significado para outra língua. Tal como na alínea anterior, existem várias formas de identificar traduções no ficheiro fornecido, sendo que o desafio é tentar apanhar o maior número de padrões possível sem trazer conteúdo despropositado para o ficheiro resultante.

2.2.3 Criar umas estatísticas dos elementos encontrados

Para resolver esta questão, basta adaptar as resoluções das alíneas anteriores para que se possa fazer uma recolha de estatísticas relacionadas com os problemas em causa, como por exemplo registar o número de traduções ou citações lidas e os respetivos autores.

3 Desenho e Implementação da Solução

3.1 Estruturas de Dados Utilizadas

De forma a mais facilmente recolher e organizar os dados dos textos fornecidos, escolhemos implementar uma estrutura de dados, tendo para isso recorrido à biblioteca *Glib*. Para organizar os vários autores e o número de citações que cada um possui, implementamos uma tabela *hash*.

```
hash = g_hash_table_new(g_str_hash, g_str_equal);
```

O uso da função ***foreach*** permite percorrer toda a tabela de autores e aplicar a cada um deles a função ***printAutor***, que trata de imprimir numa tabela o número de citações que cada autor possui.

```
g_hash_table_foreach(hash, printAutor, NULL);

gboolean printAutor(gpointer key, gpointer value, gpointer data){
    fprintf(file, "<tr>");
    fprintf(file, "<td>%s</td>\n",key);
    fprintf(file, "<td>%d</td>",*((int *)value));
    fprintf(file, "</tr>");
    return FALSE;
}
```

3.2 Citações

No caso das citações, é possível apanhar todas as citações com uma só expressão regular, tal como demonstramos na tabela que se segue.

Expressão regular	Padrão
$\wedge([*][\ ]*"\"")$	$*\"$ $*\sqcup\"$ $*\sqcup\sqcup\"$ $*\sqcup\sqcup\sqcup\"$

De forma a registar os autores das citações, e sabendo que, em cada página, um autor pode ser identificado pela existência de uma tag ***<text>*** seguida do excerto {{***Autor***, iniciamos o estado `AUTOR` sempre que encontramos estes elementos. Para efetivamente registar o seu nome, procuramos pelo elemento ***Wikipedia***, que existe em todas as páginas de autor e que contém o nome do mesmo. Escolhemos identificar o nome do autor através deste elemento ao invés do elemento ***Nome***, uma vez que este último se encontra muitas vezes sem conteúdo, ao contrário do primeiro, no qual isso acontece raras vezes.

Todo este processo encontra-se refletido no excerto de *FLEX* que se segue.

```
^([ ]*"<page>")                          { nome=""; BEGIN PAGE; }
<PAGE>"<text".*"{{Autor"                 { BEGIN AUTOR; }
<AUTOR>"|"[ ]*("Wikipedia")[ ]*"="[ ]*   { BEGIN NOME; }
<NOME>[^("|"|\n)]*                       { nome = strdup(yytext);
                                           BEGIN INITIAL; }
^([*][ ]*"&quot;")                       { BEGIN QUOTE;}
<QUOTE>[^\n]*                            { fprintf(file,"<p><b>%s</b>",nome);
                                           fprintf(file," %s</p>\n",yytext);
                                           BEGIN INITIAL; }
<AUTOR>^("[[Categoria")                  { BEGIN PAGE; }
<PAGE>^([*][ ]*"&quot;")                 { BEGIN QUOTE; }
<*>(.|\n)                                {;}
```

Como podemos verificar no excerto anterior foi necessário criar vários estados para retirar todas as citações do ficheiro recebido. Sendo assim, está representado na figura seguinte todos esses estados e a alternância entre os mesmos.

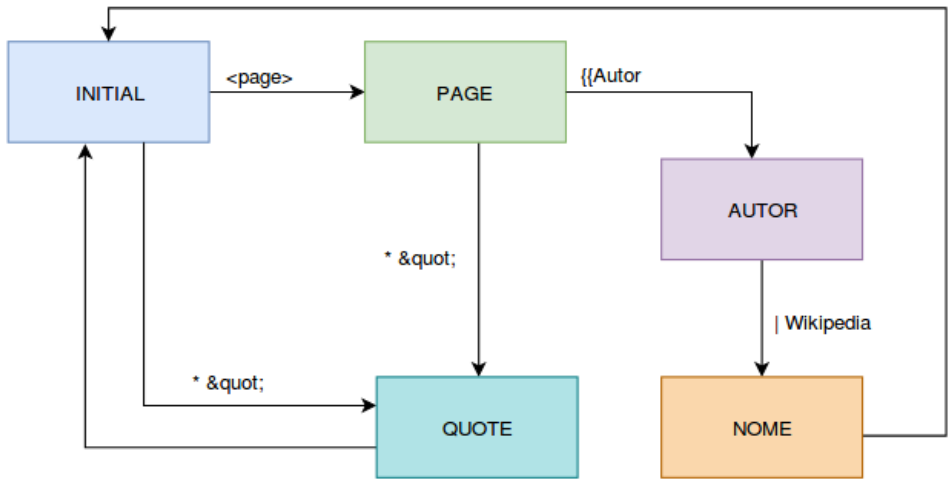


Figura 1: Ciclo de estados possíves para uma citação

Para que o conteúdo do texto resultante seja escrito no ficheiro de output, criou-se um ficheiro html **quotes.html**, e foi nesse mesmo ficheiro que se realizaram as escritas do texto, recorrendo à função *fprintf*.

```
int main(){
    file = fopen("quotes.html","w");
    fprintf(file, "<HTML> <BODY> <meta charset='UTF-8'/>");
```

```
    yylex();
    fprintf(file,"</BODY> </HTML>");
    fclose(file);
    return 0;
}
```

3.3 Traduções

Neste caso, os padrões encontrados já diferem bastante. Segue uma lista das diferentes formas de traduções encontradas.

- :Tradução Literal:
- : Tradução Literal:
- ::Tradução Literal:
- ** Tradução Literal:
- :- ''Tradução Literal:
- ::- Tradução:
- : Tradução:
- : Tradução Grega:
- : Tradução Portuguesa:
- : <u>Tradução:</u>
- : <u>Literalmente:</u>
- :- Tradução:
- ::- ''Tradução:
- :- ''Tradução:
- ** Tradução:
- **Tradução:
- **''Tradução'':
- ** ''Tradução'':
- **Original:

Para apanhar uma tradução, convém também filtrar o texto ao qual a mesma se refere. Assim sendo, e como podemos ver no excerto de `FLEX` que se segue, na maioria dos casos é apanhada uma linha começada por um asterisco e, opcionalmente, pelo identificador `&quot;` bem com um certo número de pelícas (') e, de seguida, é apanhada a linha correspondente à tradução (correspondente ao pedaço de expressão regular que se segue a \n). As expressões regulares que se seguem condensam quase todos os tipos de traduções enunciados anteriormente. Para os padrões ** Tradução Literal: e : Tradução Portuguesa: não foi possível capturar as traduções definidas por estes dado que se encontrava texto entre a citação e a respetiva tradução.

```
%%
[*].*\n[:]*[-]*[ ]*[']*(?i:("Tradução"|"Tradução literal"))[:]?[^\n]*
[*]" &quot;".*\n"**"[ ]?"Tradução:"[^\n]*
[*][ ]*"&quot;".*\n[*]*"Original"[ ]*[:]?[^\n]*
[*][ ]?.*\n[*]{2}[']{2}"Tradução''"[:][^\n]*
[*][ ]*[']{3}.*\n[*]{2}[ ]*[']{2}"Tradução"[^\n]*
[*]" &quot;".*"&quot;"\n": "[A-Za-z][^\n]*
[:].*[']{3}.*\n[:]+"Tradução literal"[^\n]*
[*].*\n[:][ ]"&lt;u&gt;"("Tradução"|"Literalmente")[:]"&lt;/u&gt;"[^\n]*
<*>(.|\n)
%%
```

Figura 2: Especificação `Flex` para obter traduções

3.4 Estatísticas

Neste ponto, o objetivo era agregar os dois problemas anteriores num só, definindo estatísticas para os mesmos. A abordagem que decidimos tomar foi a de criar uma página html que servisse de índice. Nesta página teríamos hiperligações para uma página contendo todas as traduções apanhadas, uma página com todas as citações apanhadas, e uma página contendo uma tabela de autores, com links para outras páginas (uma para cada autor), contendo as suas citações. Para além disso, apresentaríamos também no final deste índice algumas estatísticas que conseguimos calcular, tal como o número de autores e o número de citações filtradas.

```
%%
^([ ]*"<page>")                          { if(page) { fclose(page); page = null; } nome=""; BEGIN PAGE; }
<PAGE>"<text".*"{{Autor"                 { BEGIN AUTOR; len++; }
<AUTOR>"|"[ ]*("Wikipedia")[ ]*"="[ ]*   { BEGIN NOME; }
<NOME>[^("|"|\n)]*                       { nome = strdup(yytext); auxnome = strdup(nome); removeSpaces(auxnome);
                                           sprintf(filename,"autores/%s.html",auxnome);
                                           page = fopen(filename, "w+");
                                           fprintf(page,"<html><head><meta charset='UTF-8'/>\n </head><body><ul>\n");
                                           fprintf(page,"<h1 align=\"center\"> %s </h1>\n<hr>", nome);
                                           BEGIN INITIAL; }
^([*][ ]*"&quot;")                       { insertAutor(nome); BEGIN QUOTE; }
<QUOTE>[^\n]*                            { if(page) insertQuote(page,yytext); ncits++; BEGIN INITIAL; }
<PAGE>^([*][ ]*"&quot;")                 { BEGIN QUOTE; }
<*>(.|\n)                                { ; }
%%
```

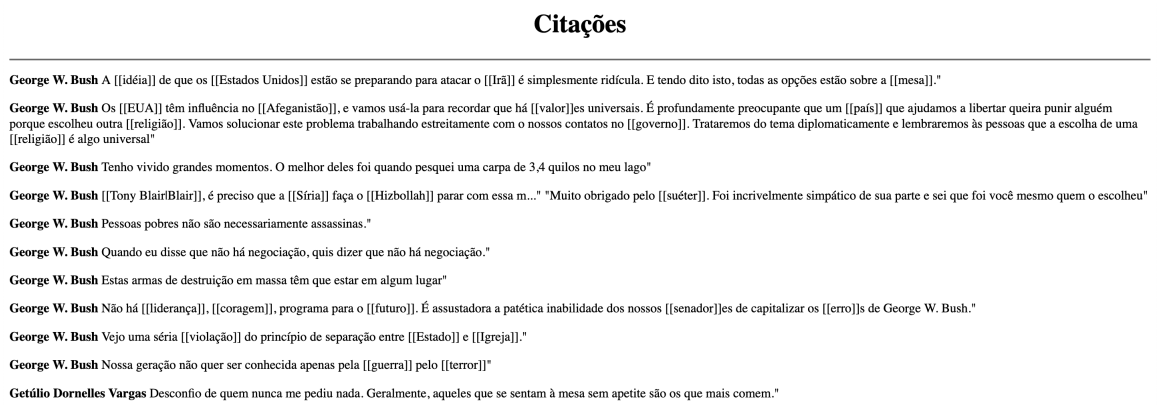
Figura 3: Especificação `Flex` para obter as estatísticas

4 Resultados Obtidos

4.1 Alínea a)

Depois de compilar os três programas, temos como resultado os ficheiros *quotes.html*, *traducoes.html*, *indice.html* e *autores.html*.

Como resposta à primeira alínea, surge então a página *quotes.html*, doaqual se mostra um excerto de seguida. Neste ficheiro, podemos observar todas as citações que foram filtradas do ficheiro original. Para as citações que contêm autor, o nome do mesmo aparece antes de cada tradução que lhe pertença.



Citações

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

George W. Bush [[Tony Blair|Blair]], é preciso que a [[Síria]] faça o [[Hizbollah]] parar com essa m..." "Muito obrigado pelo [[suéter]]. Foi incrivelmente simpático de sua parte e sei que foi você mesmo quem o escolheu"

George W. Bush Pessoas pobres não são necessariamente assassinas."

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

George W. Bush Estas armas de destruição em massa têm que estar em algum lugar"

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

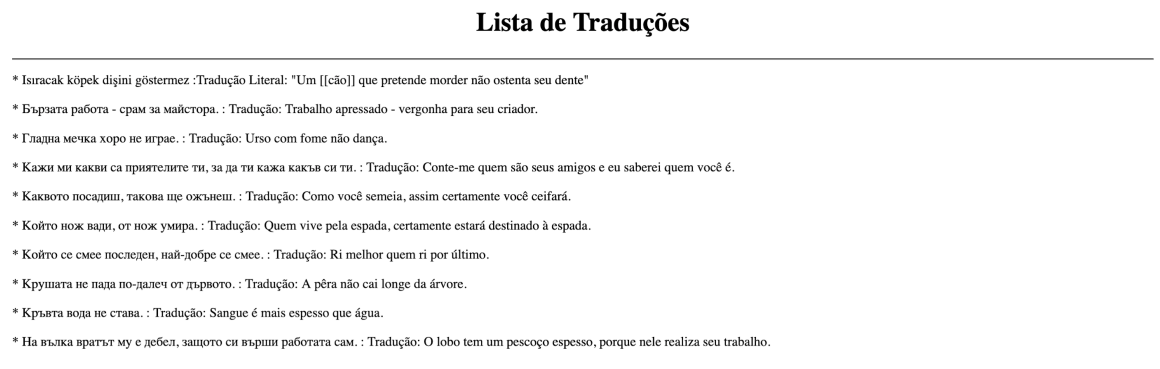
George W. Bush Nossa geração não quer ser conhecida apenas pela [[guerra]] pelo [[terror]]"

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

Figura 4: Excerto do ficheiro quotes.html

4.2 Alínea b)

O ficheiro *traducoes.html* é o resultado da execução do segundo programa que, por sua vez, responde à segunda alínea da questão que abordamos. Neste ficheiro, apresentamos as citações que se seguem de uma tradução. Apresenta-se uma tradução por linha. No final do ficheiro, surge a contagem total de traduções lidas.



Lista de Traduções

* Isıracak köpek dişini göstermez :Tradução Literal: "Um [[cão]] que pretende morder não ostenta seu dente"

* Бързата работа - срам за майстора. : Tradução: Trabalho apressado - vergonha para seu criador.

* Гладна мечка хоро не играе. : Tradução: Urso com fome não dança.

* Кажи ми какви са приятелите ти, за да ти кажа какъв си ти. : Tradução: Conte-me quem são seus amigos e eu saberei quem você é.

* Каквото посадиш, такова ще ожънеш. : Tradução: Como você semeia, assim certamente você ceifará.

* Който нож вади, от нож умира. : Tradução: Quem vive pela espada, certamente estará destinado à espada.

* Който се смее последен, най-добре се смее. : Tradução: Ri melhor quem ri por último.

* Крушата не пада по-далеч от дървото. : Tradução: A pêra não cai longe da árvore.

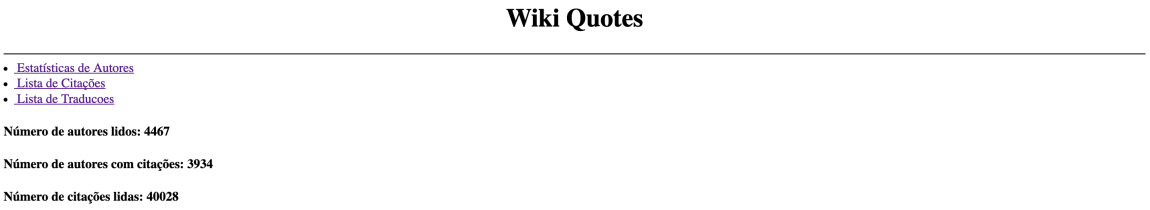
* Кръвта вода не става. : Tradução: Sangue é mais espesso que água.

* На вълка вратът му е дебел, защото си върши работата сам. : Tradução: O lobo tem um pescoço espesso, porque nele realiza seu trabalho.

Figura 5: Excerto do ficheiro traducoes.html

4.3 Alínea c)

Depois de executar os dois primeiros programas, podemos então executar o terceiro e demonstrar todas as estatísticas apanhadas, bem como proporcionar uma visão geral do problema, uma vez que agregamos numa página *indice.html* as respostas às três alíneas da questão em causa.



Wiki Quotes

- Estatísticas de Autores
- Lista de Citações
- Lista de Traducoes

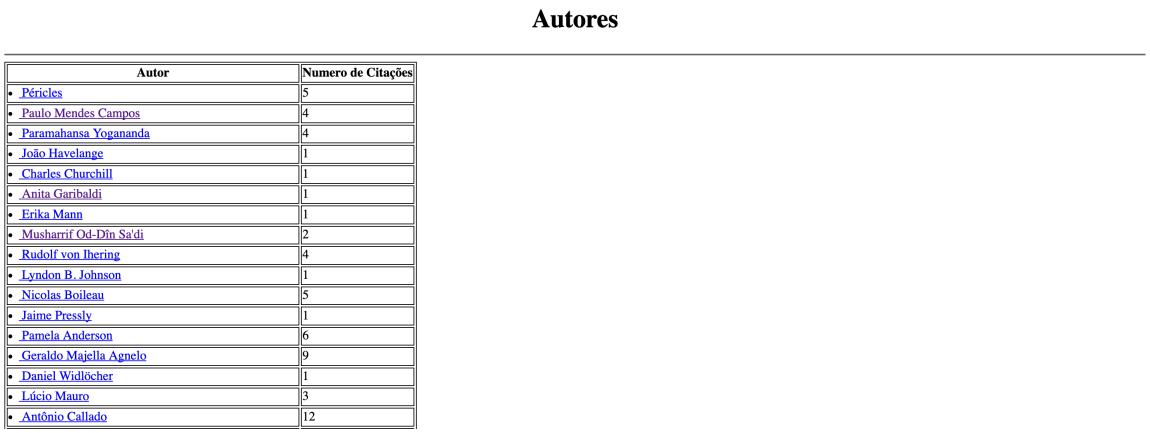
Número de autores lidos: 4467

Número de autores com citações: 3934

Número de citações lidas: 40028

Figura 6: Ficheiro indice.html

A imagem que se segue ilustra o resultado de aceder ao link "Estatísticas de Autores" da página índice.

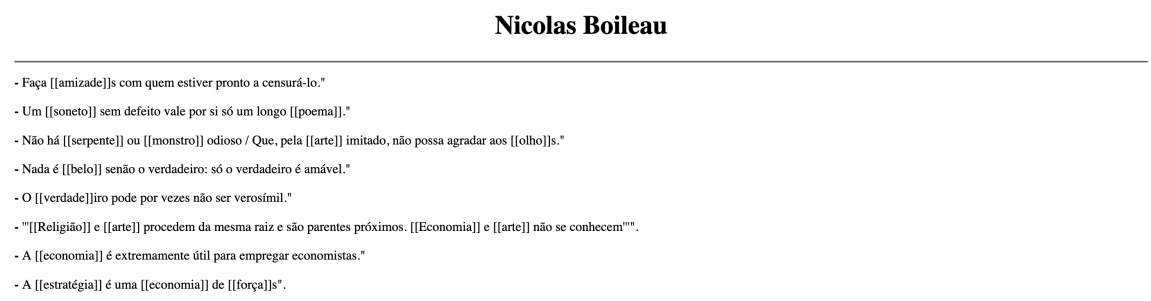


Autores

Autor	Numero de Citações
Péricles	5
Paulo Mendes Campos	4
Paramahansa Yogananda	4
João Havelange	1
Charles Churchill	1
Anita Garibaldi	1
Erika Mann	1
Musharrif Od-Dîn Sa'di	2
Rudolf von Ihering	4
Lyndon B. Johnson	1
Nicolas Boileau	5
Jaime Pressly	1
Pamela Anderson	6
Geraldo Majella Agnelo	9
Daniel Widlöcher	1
Lúcio Mauro	3
Antônio Callado	12

Figura 7: Excerto do ficheiro autores.html

Esta próxima imagem é resultado de clicar no link relativo ao autor *Nicolas Boileau* presente na tabela de autores da página *autores.html*.



Nicolas Boileau

- Faça [[amizade]]s com quem estiver pronto a censurá-lo."

- Um [[soneto]] sem defeito vale por si só um longo [[poema]]."

- Não há [[serpente]] ou [[monstro]] odioso / Que, pela [[arte]] imitado, não possa agradar aos [[olho]]s."

- Nada é [[belo]] senão o verdadeiro: só o verdadeiro é amável."

- O [[verdade]]iro pode por vezes não ser verosímil."

- "'[[Religião]] e [[arte]] procedem da mesma raiz e são parentes próximos. [[Economia]] e [[arte]] não se conhecem"".

- A [[economia]] é extremamente útil para empregar economistas."

- A [[estratégia]] é uma [[economia]] de [[força]]s".

Figura 8: Exemplo de um ficheiro de quotes de autor

5 Guia de Utilização

5.0.1 Compilação

Como podemos ver na makefile apresentada em baixo, para compilar os programas basta correr o comando `make`, ficando com três executáveis com os nomes *citacoes*, *traducoes* e *estatisticas*. Esses executáveis correspondem, respetivamente, às alíneas a), b) e c) da questão resolvida.

Para além de compilar o projeto, esta makefile trata também da limpeza dos ficheiros criados aquando da compilação e da execução dos mesmos. Para usar esta funcionalidade basta correr o comando `make clean`.

```
1   all: citacoes traducoes estatisticas
2       mkdir -p autores
3
4   citacoes: ex1.l
5       flex ex1.l
6       cc lex.yy.c -o citacoes `pkg-config --cflags --libs glib-2.0`
7
8   traducoes: ex2.l
9       flex ex2.l
10      cc lex.yy.c -o traducoes `pkg-config --cflags --libs glib-2.0`
11
12  estatisticas: ex3.l
13      flex ex3.l
14      cc lex.yy.c -o estatisticas `pkg-config --cflags --libs glib-2.0
15
16  clean:
17      rm -f lex.yy.c
18      rm -f citacoes
19      rm -f traducoes
20      rm -f estatisticas
21      rm -rf autores/
22      rm -f quotes.html
23      rm -f traducoes.html
24      rm -f autores.html
25      rm -f indice.html
```

Figura 9: Makefile do projeto

5.0.2 Execução

Depois de compilar, é preciso ter em atenção a ordem com que se executam os três programas disponíveis. Uma vez que a alínea c) usa os resultados da alínea a) e b) para criar as estatísticas, é preciso correr estes dois últimos programas em primeiro lugar, e só depois o programa *estatisticas*.

Segue um exemplo de execução dos programas disponíveis.

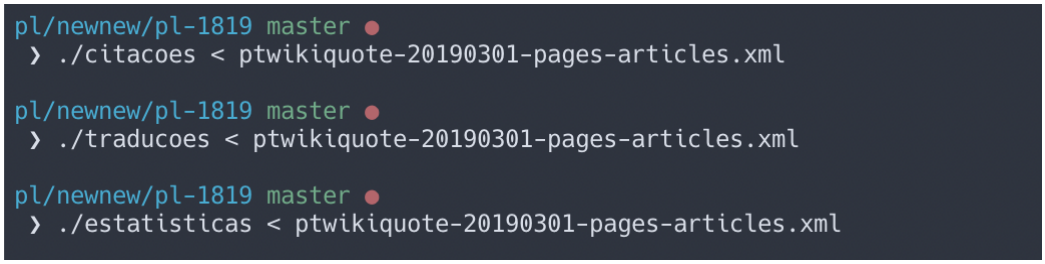


Figura 10: Exemplos de execução

6 Conclusão

Após ter sido descrito todo o processo de desenvolvimento deste trabalho, desde a descrição do problema em causa até à implementação da solução, resta agora apresentar uma breve conclusão sobre todo o processo.

A realização deste projeto trouxe a consolidação da matéria lecionada até ao momento relativa à ferramenta de filtragem de texto *Flex*. Esta ferramenta mostrou-se bastante útil no processamento e filtragem de textos, tornando bastante fácil a sua manipulação e a criação de diversos resultados a partir de um mesmo texto.

De uma forma geral, terminamos este trabalho prático confiantes de que o seu objetivo foi bem cumprido e que todas as alíneas do enunciado que nos foi atribuído foram corretamente respondidas, tendo sido um projeto no qual nos empenhamos e nos esforçamos para aprender e fazer algo além dos requisitos estabelecidos, como por exemplo a criação de páginas html para todas as alíneas para proporcionar uma leitura mais apelativa.