

# Universidade do Minho

## Trabalho Prático I

Mestrado Integrado em Engenharia Informática

Processamento de Linguagens.

2º Semestre — 2018/2019

a77523 - Alexandre Martins  
a74814 - João Bernardo Freitas  
a74806 - João Amorim

30 de Março 2019  
Braga

## Resumo

Este projeto tem como objetivo a interação dos alunos com a ferramenta *Flex* permitindo o aumento da capacidade destes relativamente à escrita de expressões regulares para filtrar e transformar textos. Este relatório pretende sumarizar os esforços efetuados para alcançar o objetivo proposto utilizando a ferramenta *Flex*.

## Contents

<b>1</b>	<b>Introdução</b>	<b>2</b>
1.1	Estutura do Relatório . . . . .	2
<b>2</b>	<b>Problema Proposto</b>	<b>3</b>
2.1	Descrição do problema . . . . .	3
<b>3</b>	<b>Solução</b>	<b>4</b>
3.1	Resposta às questões do enunciado . . . . .	4
3.2	Opcional . . . . .	10
3.3	Autômato . . . . .	11
3.3.1	TP1 . . . . .	11
3.3.2	OPCIONAL . . . . .	11
<b>4</b>	<b>Conclusão</b>	<b>12</b>
<b>5</b>	<b>Código</b>	<b>13</b>
5.1	TP1 . . . . .	13
5.2	Opcional . . . . .	15
5.3	MakeFile . . . . .	16
5.4	Comando . . . . .	16

# 1 Introdução

## 1.1 Estrutura do Relatório

O enunciado foi atribuído de acordo com o menor número de aluno de todos os elementos do grupo. Como tal o relatório debruçar-se-à sobre o enunciado número cinco **Wiki Quotes: provérbios**.

Começaremos por introduzir o problema proposto, passaremos para a concepção da solução e, finalmente, iremos realizar uma análise crítica ao trabalho elaborado.

Na secção enunciada como **Problema Proposto** apresentaremos o problema retratando o formato sobre o qual iremos ter de extrair informação.

Na secção **Solução**, iremos abordar os métodos utilizados na resposta ao enunciado.

Finalmente, na secção **Conclusão** enumeraremos as dificuldades sentidas na resolução do enunciado, mencionando como foram ultrapassadas.

## 2 Problema Proposto

### 2.1 Descrição do problema

O nosso problema prático aborda **HTML**, *Hyper Text Markup Language*, que é uma linguagem de marcação utilizada na construção de páginas na Web. O nosso objectivo é utilizar a semântica peculiar que **HTML** proporciona para responder às questões que nos foram propostas.

Foi-nos então apresentado um enunciado com as seguintes questões:

- Criar uma lista citações com respectivo autor;
- Criar uma lista de provérbios;
- Criar uma listagem de provérbios "adulterados" e seu original;
- Apresentar estatísticas relativas ao trabalho efectuado.

Para a resolução destas questões foi nos fornecido, pelos docentes, o seguinte ficheiro:

*ptwikiquote-20190301-pages-articles.xml.bz2*.

## 3 Solução

### 3.1 Resposta às questões do enunciado

- Alínea 1)

Para a solução da *alínea 1)*, é necessário retirar todas as citações presentes no ficheiro *ptwikiquote-20190301-pages-articles.xml* com o seu respectivo autor, para tal, começamos por determinar como é que as citações estão identificadas. Após leitura, verificamos que as citações estão definidas, inicialmente, por:

---

```
*&quot;;  
- &quot;;  
* ''&quot;;  
&quot;;
```

---

E terminam com

---

```
&quot;;
```

---

Como tal, tivemos de desenvolver quatro expressões regulares de forma a extrair todas as citações e seus conteúdos.

---

```
^(\*)(&quot;);.+(&quot;);  
^(\-[ ])(&quot;);.+(&quot;);  
^(\*[ ]\'+)(&quot;);.+(&quot;);  
^(\*[ ])(&quot;);.+(&quot;);
```

---

Após extração das citações reparamos que era necessário a remoção dos caracteres correspondentes ao início e fim de uma citação, dependendo do tipo.

---

```
^(\*)(&quot;);.+(&quot;); {yytext[yytextlen-6]='\0'; quote =  
    strdup(yytext+7);printf("%s \n ", quote);}  
^(\-[ ])(&quot;);.+(&quot;); {yytext[yytextlen-6]='\0'; quote =  
    strdup(yytext+8);printf("%s \n ", quote);}  
^(\*[ ]\'+)(&quot;);.+(&quot;); {yytext[yytextlen-6]='\0'; quote =  
    strdup(yytext+10);printf("%s \n ", quote);}  
^(\*[ ])(&quot;);.+(&quot;); {yytext[yytextlen-6]='\0'; quote =  
    strdup(yytext+8);printf("%s \n ", quote);}
```

---

Tendo em conta os requisitos das próximas alíneas e a falta de autores das citações decidimos definir várias **Start Conditions** para facilitar a extração do nome dos autores e provérbios.

Para esta alínea necessitávamos de apenas duas **start conditions**, *META* e *QUOTE*.

*META* é a **start condition** inicial, que começa sempre que é encontrado o identificador de início de página, sendo que o objetivo desta é indicar se a página é para ignorar, para citações ou para provérbios:

---

```
\<page\>
```

---

Definimos então a entrada na *META* como:

---

```
\<page\> {BEGIN META;}
```

---

De seguida, para extrair apenas citações, que definir que estas estão contidas em páginas cujo o título inclui pelo menos duas palavras começadas por maiúscula.

Ficamos então com a seguinte definição:

---

```
<META>\<title\>([A-Z][a-z ]+){2,}\</title\>
{yytext[yy leng-8]='\0';autor=strdup(yytext+7);BEGIN QUOTE;}
```

---

Para ignorar tudo o que não era pertinente para a resolução do enunciado decidimos que sempre que é encontrado o identificador de final de página deve voltar à **start condition** inicial, ignorando todos os caracteres que não fazem parte das citações.

Para tal definimos as seguintes expressões regulares globais:

---

```
<*>\</page\> {BEGIN INITIAL;}
```

---

```
<*>(.|\n) {;}
```

---



- Alínea 2)

Para a solução da *alínea 2)* era necessário retirar todos os provérbios, para isso, decidimos expandir as **start conditions** de forma a extrair os provérbios. De acordo com o enunciado só devíamos incluir provérbios que estavam contidos em páginas cujo título começava por "Provérbios".

Para isso definimos a start condition *PROVERBIO*

---

```
<META>\<title>(Proverbio).*\</title> {BEGIN PROVERBIO;}
```

---

Para a extracção dos provérbios tivemos que, mais uma vez, determinar como é que os provérbios estão identificados. Após analisar novamente o documento, verificamos que todos os provérbios estão definidos de forma semelhante às citações, isto é:

---

```
*&quot;;  
* ''&quot;;  
* &quot;;
```

---

E tal como nas citações terminavam por:

---

```
&quot;;
```

---

Logo desenvolvemos as seguintes expressões regulares:

---

```
<PROVERBIO>~(\*)(&quot;).*(&quot;) {yytext[yy leng-6]='\0'; quote =  
    strdup(yytext+7);printf("PROVERBIO:%s \n ", quote);}  
<PROVERBIO>~(\*[ ]\')+(&quot;).*(&quot;) {yytext[yy leng-6]='\0';  
    quote = strdup(yytext+10);printf("PROVERBIO:%s \n ", quote);}  
<PROVERBIO>~(\*[ ])(&quot;).*(&quot;) {yytext[yy leng-6]='\0'; quote  
    = strdup(yytext+8);printf("PROVERBIO:%s \n ", quote);}
```

---



- **Alínea 3)**

Para a resolução da *alínea 3)* é necessário extrair todos os provérbios adulterados, para isso, foi necessário a criação de uma última **start condition** à qual chamamos *ADULTERADO*.

Após leitura do ficheiro reparamos que todos os provérbios adulterados eram precedidos pela seguinte linha:

---

```
** ''' Adulterao :'''
```

---

Sendo que o início destes estava definido por:

---

```
*** &quot;;
```

---

E, tal como nas ultimas duas alíneas, acabava por:

---

```
(&quot;;)
```

---

Para além disso reparamos que a lista de adulterados acabava sempre quando no início de uma linha era:

---

```
* &quot;;
```

---

Que indica um provérbio novo.

Definimos então que os provérbios adulterados estão dentro de páginas de provérbios logo criamos a seguinte expressão regular:

---

```
<PROVERBIO>^(\\*[ ])(\\'{3}Adulterao\\: '\\{3}) {BEGIN ADULTERADO;}
```

---

Para extrair os adulterados e voltar á **start condition** *PROVERBIO* criamos duas expressões regulares:

---

```
<ADULTERADO>^(\\*{3}[ ])(&quot;;).*(&quot;;) {yytext[yy leng-6]='\\0';
    quote = strdup(yytext+10);printf("ADULTERADO:%s \\n ", quote);}
<ADULTERADO>^(\\*[ ])(&quot;;) {BEGIN PROVERBIO;}
```

---

- Alínea 4)

Para esta alínea decidimos que era necessário a adição de vários contadores, nomeadamente, o número de páginas, citações, provérbios, provérbios adulterados e autores lidos.

---

```
int a=0; //autores
int q=0; //quotes
int p=0; //proverbios
int ad=0; //adulterados
int pa=0; //paginas
```

---

Alterando as linhas acima de forma a incrementar estes contadores temos:

---

```
\<page\> {pa++;BEGIN META;}

<META>\<title>([A-Z][a-z ]+){2,}\</title>
    {yytext[yytext-8]='\0';autor=strdup(yytext+7);a++;BEGIN QUOTE;}
<META>\<title>(Provrbio).*\</title> {BEGIN PROVERBIO;}

<QUOTE>^(\*)(&quot;).+(&quot;); {yytext[yytext-6]='\0'; quote =
    strdup(yytext+7);q++;printf("%s:%s \n ", autor, quote);}
<QUOTE>^(\-[ ])(&quot;).+(&quot;); {yytext[yytext-6]='\0'; quote =
    strdup(yytext+8);q++;printf("%s:%s \n ", autor, quote);}
<QUOTE>^(\*[ ]\'+)(&quot;).+(&quot;); {yytext[yytext-6]='\0'; quote =
    strdup(yytext+10);q++;printf("%s:%s \n ", autor, quote);}
<QUOTE>^(\*[ ])(&quot;).+(&quot;); {yytext[yytext-6]='\0'; quote =
    strdup(yytext+8);q++;printf("%s:%s \n ", autor, quote);}

<PROVERBIO>^(\*)(&quot;).*(&quot;); {yytext[yytext-6]='\0'; quote =
    strdup(yytext+7);p++;printf("PROVERBIO:%s \n ", quote);}
<PROVERBIO>^(\*[ ]\'+)(&quot;).*(&quot;); {yytext[yytext-6]='\0'; quote =
    strdup(yytext+10);p++;printf("PROVERBIO:%s \n ", quote);}
<PROVERBIO>^(\*[ ])(&quot;).*(&quot;); {yytext[yytext-6]='\0'; quote =
    strdup(yytext+8);p++;printf("PROVERBIO:%s \n ", quote);}
<PROVERBIO>^(\*[ ])(\{3\}Adulterao\:\{3\}) {BEGIN ADULTERADO;}

<ADULTERADO>^(\*[ ])(&quot;).*(&quot;); {yytext[yytext-6]='\0'; quote
    = strdup(yytext+10);ad++;printf("ADULTERADO:%s \n ", quote);}
<ADULTERADO>^(\*[ ])(&quot;); {BEGIN PROVERBIO;}
```

---

---

```

int main(){
    printf("Inicio da Filtragem\n");
    yylex();
    printf("Numero total de autores processados:%d \n",a);
    printf("Numero total de citacoes processadas:%d \n",q);
    printf("Numero total de proverbios processados:%d \n",p);
    printf("Numero total de proverbios adulterados processados:%d \n",ad);
    printf("Numero total de paginas processadas:%d \n",pa);
    printf("\nFim da Filtragem\n");
    return 0;
}

```

---

### 3.2 Opcional

Após a filtragem do ficheiro *ptwikiquote-20190301- pages-articles.xml* reparamos que existiam links e parênteses retos que estavam a mais, para os remover decidimos criar um segundo filtro que tem como único objetivo "limpar" o output do filtro original.

A utilização deste segundo filtro é puramente opcional, sendo que, este fica definido por quatro expressões regulares.

Expressão regular que remove os links.

---

```
(<ref>).*(</ref>) {}
```

---

Expressão regular que remove os parênteses rectos.

---

```
(\[) {}
(\]) {}
```

---

Expressão regular que remove &quot; presentes no meio de uma citação.

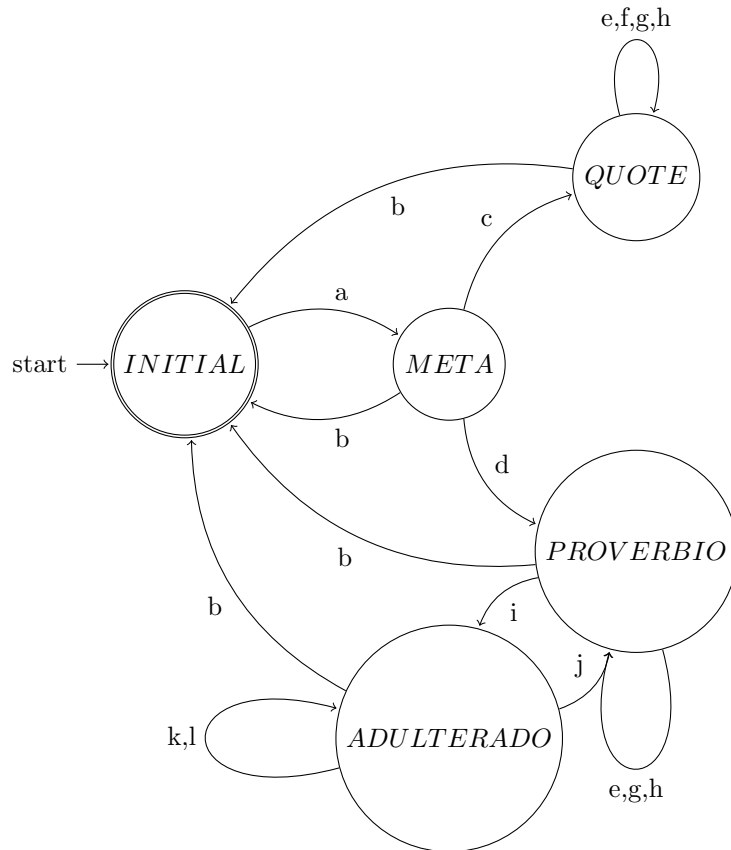
---

```
(&quot;[.;:]) {}
```

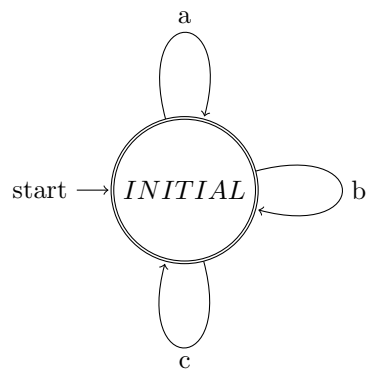
---

### 3.3 Autómato

#### 3.3.1 TP1



#### 3.3.2 OPCIONAL



## 4 Conclusão

Visto que o objectivo deste trabalho era aumentar contacto com a ferramenta *Flex*, as is comoma capacidade de escrever *Expressões Regulares* e desenvolver *Processadores de Linguagens Regulares* podemos afirmar que foi um sucesso.

Porém, no decorrer da resolução deste trabalho, foram surgindo alguns desafios que foram ultrapassados, nomeadamente: encontrar todos os identificadores das citações/provérbios, como separar as citações dos provérbios, encontrar o padrão dos provérbios adulterados e definir as respetivas expressões regulares sendo que o maior obstáculo foi descobrir como efetuar a "limpeza" das citações.

Em suma, o trabalho foi bastante benéfico para todos os elementos, sendo que aprofundou o entendimento das expressões regulares e pressupostos da ferramenta *Flex*.

## 5 Código

### 5.1 TP1

Código correspondente à solução das quatro alíneas.

---

```
%{
#include <stdio.h>
#include <string.h>

char* autor="";
char* quote="";
int a=0;//autores
int q=0;//quotes
int p=0;//proverbios
int ad=0;//adulterados
int pa=0;//paginas

}%

%x META
%x QUOTE
%x PROVERBIO
%x ADULTERADO

%%
\<page\> {pa++;BEGIN META;}

<META>\<title>([A-Z][a-z ]+){2,}\</title>
{yytext[yyval-8]='\0';autor=strdup(yytext+7);a++;BEGIN QUOTE;}
<META>\<title>(Provrbio).*\</title> {BEGIN PROVERBIO;}

<QUOTE>^(\*)(&quot;).+(&quot;){yytext[yyval-6]='\0'; quote =
    strdup(yytext+7);q++;printf("%s:%s \n ", autor, quote);}
<QUOTE>^(\[-\ ])(&quot;).+(&quot;){yytext[yyval-6]='\0'; quote =
    strdup(yytext+8);q++;printf("%s:%s \n ", autor, quote);}
<QUOTE>^(\*[ \ ]'+)(&quot;).+(&quot;){yytext[yyval-6]='\0'; quote =
    strdup(yytext+10);q++;printf("%s:%s \n ", autor, quote);}
<QUOTE>^(\*[ \ ])(&quot;).+(&quot;){yytext[yyval-6]='\0'; quote =
    strdup(yytext+8);q++;printf("%s:%s \n ", autor, quote);}

<PROVERBIO>^(\*)(&quot;).*(&quot;){yytext[yyval-6]='\0'; quote =
    strdup(yytext+7);p++;printf("PROVERBIO:%s \n ", quote);}
<PROVERBIO>^(\*[ \ ]'+)(&quot;).*(&quot;){yytext[yyval-6]='\0'; quote =
    strdup(yytext+10);p++;printf("PROVERBIO:%s \n ", quote);}
<PROVERBIO>^(\*[ \ ])(&quot;).*(&quot;){yytext[yyval-6]='\0'; quote =
    strdup(yytext+8);p++;printf("PROVERBIO:%s \n ", quote);}
<PROVERBIO>^(\*[ \ ])(\{3\}Adulterao:\{3\}) {BEGIN ADULTERADO;}
```

```

<ADULTERADO>^(\*[3][ ])(&quot;).*(&quot;) {yytext[yyteng-6]='0'; quote
    = strdup(yytext+10);ad++;printf("ADULTERADO:%s \n ", quote);}
<ADULTERADO>^(\*[ ])(&quot;) {BEGIN PROVERBIO;}

<*>\</page\> {BEGIN INITIAL;}
<*>(.|\n) {;}
%%

int yywrap(){
    return 1;
}

int main(){
    printf("Inicio da Filtragem\n");
    yylex();
    printf("Nmero total de autores processados:%d \n",a);
    printf("Nmero total de citaes processadas:%d \n",q);
    printf("Nmero total de provrbios processados:%d \n",p);
    printf("Nmero total de provrbios adulterados processados:%d \n",ad);
    printf("Nmero total de pginas processadas:%d \n",pa);
    printf("\nFim da Filtragem\n");
    return 0;
}

```

---

## 5.2 Opcional

Código correspondente ao filtro opcional.

---

```
%{
#include <stdio.h>
#include <string.h>

%}

%%

(&lt;ref&gt;).*(&lt;\/ref&gt;) {}
(\[]          {}
(\])          {}
(&quot;[. ;:])    {}
%%

int yywrap(){
    return 1;
}

int main(){
    yylex();
    return 0;
}
```

---



### 5.3 MakeFile

Para agilizar o processo de compilação de ambos os filtros criamos o seguinte MakeFile.

---

```
inicial:
    lex -o parte1.c TP1.1
    lex -o parte2.c TP1_2.1
    gcc -o tp1 parte1.c
    gcc -o limpa parte2.c
    rm parte1.c
    rm parte2.c
```

---

### 5.4 Comando

Para fazer com que os dois filtros sejam utilizados utilizamos o seguinte comando.

---

```
./tp1 <ptwikiquote-20190301-pages-articles.xml | ./limpa
```

---

## Anexos

```
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:Á idéia de que os Estados Unidos estão se preparando para atacar o Irã é simplesmente ridícula. É 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 valores universais. É profundamente preocupante que um país que ajudamos a libertar queira punir alguém porque escolheu outra religião. Vamos a Junooner este problema trabalhando estreitamente com o nossos contatos no governo. Tentaremos 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 d'ela foi quando pespei uma campo 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.... Muito obrigado pelo sofrer. Foi incrivelmente simpático de sua parte e sei que foi você mesmo quem o escolheu.
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 senadores de capitalizar os erros 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.
Ludwig Mies van der Rohe:Deus está nos detalhes.
Ludwig Mies van der Rohe:Menos é mais.
Martin Luther King Junior:À antiga lei do olho por olho, dente por dente:olho por olho acaba por deixar todo mundo cego.
Martin Luther King Junior:Meso as noites completamente sem estrelas, podem anunciar a aurora de uma grande realização.
Martin Luther King Junior:Um líder autêntico, em vez de buscar consenso, molda-o.
Martin Luther King Junior:Nós não estaremos satisfeitos até que a justiça corra como água e a retidão como um caudaloso rio.
Martin Luther King Junior:À greve, no fundo, é a linguagem dos que não são ouvidos.
Martin Luther King Junior:Que sempre crianças criativas e dedicadas tornam o mundo melhor.
Martin Luther King Junior:Nossa eterna mensagem de esperança é que a aurora chegará.
Martin Luther King Junior:Se um homem não descobriu algo por que morrer, ele não está preparado para viver.
Martin Luther King Junior:ser humano deve desenvolver, para todos os seus conflitos, um método que reflete a vingança, a agressão e a retaliação. A base para esse tipo de método é o amor.
Martin Luther King Junior:Through violence you may murder a murderer but you can't murder murder. Through violence you may murder a liar but you can't establish truth. Through violence you may murder a hater, but you can't murder hate.
Darkness cannot put out darkness. Only light can do that.
Martin Luther King Junior:Verdadeira paz somente não é a ausência de tensão, é a presença de justiça.
Martin Luther King Junior:Nós temos que combater a dureza da serpente com a suavidade da pomba, uma mente dura e um coração tenro.
Martin Luther King Junior:Nada no mundo é mais perigoso que a ignorância silenciosa e a estúpidez conscienciosa.
Martin Luther King Junior:Azer é a única força capaz de transformar um inimigo num amigo.
Martin Luther King Junior:Eu tenho o sonho de ver um dia meus 4 filhos vivendo numa nação em que não sejam julgados pela cor de sua pele, mas sim pelo seu caráter.
Martin Luther King Junior:Pessoas oprimidas não podem permanecer oprimidas para sempre.
```

Figure 1: Exemplo do output de citações

```
PROVERBIO:Na primeira quem quer cai, na segunda cai quem quer
ADULTERADO:Na primeira quem quer cai; na segunda cai quem quer; na terceira quem é parvo
PROVERBIO:Nada é mais incompatível com o estudo, do que o sono e o cansaço
PROVERBIO:Nada é tão bem empregado, como aquilo que se dá aos que precisam
PROVERBIO:Nada há mais difícil em tudo, que o bem começar
PROVERBIO:Nada há mais importuno que os cumprimentos, quando são excessivos
PROVERBIO:Nada há tão contagioso, como o medo
PROVERBIO:Nada se dá com tanta liberalidade como os conselhos
PROVERBIO:Não alimentes burros a pão-de-ló
PROVERBIO:Não confies em flores que desabrocham em Março, nem em mulher que não tem vergonha
PROVERBIO:Não contes com o ovo no cu da galinha
PROVERBIO:Não coso morto nem vivo, coso isto que está descosido
PROVERBIO:Não coso vivo nem morto, coso aquilo que está roto
PROVERBIO:Não crie cão, quem não lhe sobeje pão
PROVERBIO:Não dá quem tem, dá quem quer bem
PROVERBIO:Não deixe escapar camarão pela rede
PROVERBIO:Não deixes para amanhã o que podes fazer hoje
ADULTERADO:Não deixe para amanhã o que você pode fazer depois de amanhã
ADULTERADO:Não deixes para amanhã o que podes beber hoje
ADULTERADO:Não faças hoje o que podes deixar para amanhã
PROVERBIO:Não destrua a árvore, para depois ter seu fruto
```

Figure 2: Exemplo do output de provérbios e provérbios adulterados

```
Número total de autores processados:4510
Número total de citações processadas:15928
Número total de provérbios processados:1868
Número total de provérbios adulterados processados:83
Número total de páginas processadas:13852
```

Figure 3: Estatísticas finais