Trabalho Prático 1 de Processamento de Linguagens

André Vieira (A78322) Eduardo Rocha (A77048) Ricardo Neves (A78764)

6 de Maio de 2018

Resumo

Este documento apresenta o relatório do primeiro projeto de Proessamento de Linguagens, da Licenciatura em Engenharia Informática da Universidade do Minho.

Conteúdo

1	Introdução	1
2	Resumo	1
3	Exercício 1	2
4	Exercício 2	5
5	Conclusões	7

1 Introdução

Este trabalho prático foi realizado no ambito da Unidade Curricular de Processamento de Linguagens. Aqui, iremos dicutir toda a estratégia e linha de pensamento que o grupo tomou, de modo a cumprir os requisitos pedidos.

Uma vez que o menor número mecanográfico é igual a 77048, e ao dividir este mesmo número por 5, constatamos que o resto desta operação é igual a 3, que corresponde ao algarismo do enunciado pertencente. Deste modo, o grupo constatou que o enunciado atribuido foi o 2.3 - Processador / sincronizador de legendas. Este trabalho engloba, em geral, 2 exercícios distintos, que iremos mencionar mais à frente, em pormenor. O objetivo máximo do grupo em relação a este trabalho foi o de realizar os dois exercicios pedidos com a maior eficiência, ganhando, assim, experiência e conhecimento em relação à ferramenta AWK, que irá ser importante no seguimento desta Unidade Curricular. Esta ferramenta permite manipular o texto de ficheiros, pesquisar e obter apenas certas palavras que estão contidas no mesmo, entre várias outas funcionalidades. Assim, este trabalho baseou-se na exploração destas mesmas funcionalidades que o AWK nos oferece. Com o fecho da introdução deste relatório, é adequado mencionar

a estrutura do mesmo. Este relatório contém o resumo do trabalho que foi realizado durante o período dado para tal, a descrição e a implementação dos dois exercícios que compõem o trabalho prático (incluíndo a linha de pensamento seguida pelo grupo, bem como imagens que ajudam à interpretação), e no final, uma pequena conclusão que reflete o trabalho realizado.

2 Resumo

Este trabalho prático resume-se, então, à manipulação de legendas de filmes. Todas as legendas disponibilizadas seguem o seguinte formato:

```
1
00:00:48,344 --> 00:00:49,500
Chamada: recebida

2
00:00:49,707 --> 00:00:53,128
-Está tudo pronto?
-Você não tinha de me substituir.

3
00:00:53,328 --> 00:00:56,014
Eu sei, mas quero fazer um turno.

4
00:01:06,485 --> 00:01:08,943
Você gosta dele, não?
Gosta de observá-lo.
```

Figura 1: Exemplo de formato de legendas

Como se pode observar na imagem acima, cada legenda apresenta o número que a caracteriza. Na linha abaixo, podemos ver também o tempo de início e de fim da legenda. Ou seja, o tempo onde a mesma aparece e desaparece durante o percurso do filme. Se quisermos obter o tempo que a legenda manteu-se visível ao espectador, é necessário fazer a subtração entre o segundo número e o primeiro. Por fim, vemos o corpo da legenda, isto é, o texto que fica visível ao utilizador.

Este template repete-se para todas as legendas diponíveis, o que facilitou o trabalho do grupo de certo modo, o que permite percorrer todas elas de uma forma semelhante.

3 Exercício 1

Este primeiro de dois exercícios desafiou o grupo de trabalho a transformar o modo como as legendas aparecem no ficheiro original.

Assim, o grupo construiu um processador que retira os identificadores das legendas, ou seja, a primeira linha de cada uma, coloque as legendas numa única linha, juntando-as com o caracter '|', e marcar com os caracteres '=' todos os intervalos de tempo com mais de 2 segundos de silencio.

Deste modo, foi contruída esta ferramenta.

```
BEGIN {i = 0; d = "00:00:00.00.000"}

/^[0]/ {timeSilence(d,$1); d = $3;}

/^[0]/ {c = $0; i = 0;}

/^[a]/ {c = $0; i = 0;}

/[a-2A-Z]/ && i >= 0 {c = c " " $0; i = 1; next;}

/[a-2A-Z]/ && i >= 1 {c = c " | " $0; i++;}

END {print c;}

END {print c;}

function timeSilence (fimAnt, iniDep){

    split(fimAnt, a, /:/);
    split(iniDep, b, /:/);
    if(b[2] > a[2]) printf("%s --> %s =========\n", fimAnt, iniDep);
    if(b[2] == a[2]) && (b[3] - a[3]) >= 2) printf("%s --> %s =========\n", fimAnt, iniDep);
}
```

Figura 2: Código construído para exercício 1

Ao observar atentamente o ficheiro das legendas original, constatamos que todos os tempos começavam com o algarismo 0. Deste modo, de modo a não imprimir o número que caracteriza a legenda, implementamos que, se a linha começar por 0, irá ser imprimida. No entanto, é aplicada a função auxiliar timeSilence, que irá ser explicada mais tarde. De seguida, é necessário imprimir o texto da legenda. Assim, se a legenda contiver uma letra, também será imprimida. No entanto, se de seguida, for encontrada outra legenda (a sua segunda parte), irá ser imprimido o caracter '—', e só depois a segunda linha da mesma.

Falando um pouco acerca da função auxiliar timeSilence, este é o método que nos permite avaliar se existe alguem tempo de silêncio com uma duração maior do que 2 segundos. Assim, esta função receve dois argumentos: o tempo final da primeira legenda, e o tempo inicial da segunda legenda. Como já foi mencionado, é aqui que estes tempos são subtraídos. Se este resultado for maior do que 2, uma série de caracteres '=' irão ser imprimidos, o que quer dizer que houve um tempo de silencio durante esta parte do filme. Se isto não se constatar, então o programa não imprimirá nada.

Esta imagem representa as legendas no ficheiro original do filme Harry Potter:

```
920
92:09:13,280 --> 02:09:18,800
-Qual tipo de magia é essa?
-Pegue a pedra!!

921
92:10:59,080 --> 02:11:06,520
Boa tarde Harry.
Presentes de seus admiradores?

922
92:11:06,520 --> 02:11:12,120
-Admiradores?
-0 que aconteceu nas masmorras entre você e o professor Quirrel...

923
92:11:12,120 --> 02:11:18,800
é um segredo absoluto. Então,
isso é claro, toda a escola já sabe disso.

924
92:11:23,160 --> 02:11:27,560
Vejo que Ron te poupou o trabalho de abrir suas rãs de chocolate.

925
92:11:27,560 --> 02:11:31,120
Foi Ron? Então ele está bem!
0 que aconteceu com Hermione?

926
92:11:31,120 --> 02:11:38,040
-Estão bem, ambos estão bem.
-Mas, o que aconteceu com a pedra?
```

Figura 3: Legendas originais do filme Harry Potter

Com a função implementada, corremos este comando AWK no terminal:

```
gawk -f Ex1.gawk harry_potter-pt.srt
```

Figura 4: Comando gawk utilizado

De imediato, recebemos o resultado final da transformação que foi executada:

```
82:87:57.560 -> 02:08:05,440 Me dé, Harry, mão gostaria voltar a ver juntos | de seu pai e sua mão de novo?
02:08:05,440 -> 02:08:11,960 Podemos devolve-los a vida, se me der o que tem.
02:08:02,700 -> 02:08:12,900 Podemos devolve-los a vida, se me der o que tem.
02:08:02,700 -> 02:08:04,500 E podemos alcança-lo juntos. | Fariámos coisas extraordinárias.
02:08:02,700 -> 02:08:04,500 E podemos alcança-lo juntos. | Fariámos coisas extraordinárias.
02:08:04,600 -> 02:08:05,600 -Mentirosol | Te matarei!|
02:08:04,500 -> 02:08:05,600 -Mentirosol | Te matarei!|
02:08:05,600 -> 02:08:05,600 -> 02:08:05,600 -Mentirosol | Te matarei!|
02:08:05,600
```

Figura 5: Legendas do filme Harry Potter modificadas

Para ter a certeza que tudo estava a funcionar corretamente, testamos o mesmo procedimento nas legendas do filme Ted:

```
344
00:28:38,395 --> 00:28:42,245
apesar do roubo de milhões de dólares,
00:28:42,245 --> 00:28:46,041
ainda não foram presos
346
00:28:46,041 --> 00:28:52,448
e possivelmente nunca serão.
90:28:52,448 --> 00:28:58,748
A maioria das leis tem alcance nacional
00:28:58,748 --> 00:29:05,946
apesar das convenções sobre o crime informático, onde a Internet
00:29:05,946 --> 00:29:11,537
não tem fronteiras e é internacional por definição.
00:29:11,537 --> 00:29:16,636
Os países não entram em acordo, o que torna esta área
351
00:29:16,636 --> 00:29:23,147
excecionalmente desafiadora de uma perspetiva legal.
90:29:23,147 --> 00:29:30,995
Mas a minha maior pergunta é esta:
333
00:29:30,995 --> 00:29:33,950
Vão sair daqui
00:29:33,950 --> 00:29:40,641
e vão ver histórias surpreendentes nos noticiários.
355
00:29:40,641 --> 00:29:44,554
Vão ler sobre programas maliciosas que fazem coisas
```

Figura 6: Legendas originais do filme Ted

Correndo o comando gawk, é devolvido o seguinte resultado:

```
Correndo o comando gawk, 6 devolvido o seguinte resultado:

0:24:54,709 -> 00:24:59,277 vamos falar sobre vocés.
0:24:59,277 -> 00:25:08,075 23% de vocés esteve no Starbucks
0:24:59,277 -- 00:25:08,075 23% de vocés esteve no Starbucks
0:25:08,075 -- 00:25:10,778 As coisas ficam mais interessantes.
0:25:08,075 -- 00:25:10,778 As coisas ficam mais interessantes.
0:25:10,180 -- 00:25:10,778 As coisas ficam mais interessantes.
0:25:10,180 -- 00:25:27,809 Não é uma ciência exata, mas é bem precisa.
0:25:11,180 -- 00:25:27,809 Não é uma ciência exata, mas é bem precisa.
0:25:27,809 -- 00:25:23,181 Eu pude identifacar um hotel onde estiveram hospedados 761 de vocés 0:25:27,809 Não é uma ciência exata, mas é bem precisa.
0:25:24,920 -- 00:25:25,181 Eu pude identifacar um hotel onde estiveram hospedados 761 de vocés 0:25:49,220 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180 -- 00:25:55,180
```

Figura 7: Legendas do filme Ted modificadas

Como se pode observar nas legendas processadas acima, não existe qualquer tempo de silêncio presente nas legendas do filme Ted. Isto deve-se ao facto de, neste caso, quando uma legenda acaba, a seguinte aparece no ecrã, instantaneamente. Outro facto acerca destas legendas é que o seu corpo apenas apresenta uma linha. Ou seja, não existe, no filme Ted, legendas com 2 linhas, como acontecia no caso do filme do Harry Potter.

Todas as legendas sofreram as transformações acima referidas. Assim sendo, o grupo concluiu o primeiro de dois exercícios do trabalho prático.

4 Exercício 2

O método que adotamos para resolver este problema foi o de, uma vez que as legendas têm todas um identificador que a caracteriza, alterar os timestamps consoante estes IDs. Ou seja, se existe legendas com o mesmo identificador em ambos os ficheiros, isto quer dizer que estas legendas são iguais (apesar de estarem em linguagens diferentes).

Assim, a função que implementamos pode receber 7 argumentos : identificador da primeira legenda no ficheiro original, identificador da primeira legenda no ficheiro que vai ser consultado e alterado, identificador da última legenda no ficheiro original, identificador da última legenda a ser alterada no ficheiro, os nomes do ficheiro original, do ficheiro alterado, e por fim, do ficheiro onde tudo vai ser guardado.

```
BEGIN { if(ARGC--7){ ARGC-8; ARGV[7]-ARGV[6]}

if(ARGV[1] =- ARGV[2] && ARGV[3] =- ARGV[4]){

alteraTimestamp(ARGV[5], ARGV[6])

}

*else{

if(ARGV[3] =- ARGV[4]) {

alteraTimestamp(ARGV[1], ARGV[2], ARGV[3], ARGV[4], ARGV[6]).
}

else{

ficheiro(ARGV[1], ARGV[2], ARGV[3], ARGV[4], ARGV[6]).
}

*function ficheiro(11, 12, f1, f2){

durl=f1-11

dur2-f2-12

scale=dur1/dur2

sinf=f1:5*scale =-11

f(t)=t*scale=shift
}

*function alteraTimestamp(11, i2, f1, f2, original, alterado){

## Retira-se o identificador da legenda, de cada ficheiro

## Retira-se o identificador da legenda, de cada ficheiro

## Se o elemento junto ao inicio da linha for diferente de 0, então guardar o conteúdo da linha em "linha!", do ficheiro original

## Se o elemento junto ao inicio da linha for diferente de 0, então guardar o conteúdo da linha em "linha!", do ficheiro original

## Se o elemento funto ao inicio da linha for diferente de 0, então guardar o conteúdo da linha em "linha2", do ficheiro original

## Se o elemento junto ao inicio da linha for diferente de 0, então guardar o conteúdo da linha em "linha2", do ficheiro original

## Se o elemento junto ao inicio da linha for diferente de 0, então guardar o conteúdo da linha em "linha2", do ficheiro original

## Se o elemento junto ao inicio da linha for diferente de 0, então guardar o conteúdo da linha em "linha2", do ficheiro original

## Se o elemento junto ao inicio da linha for diferente de 0, então guardar o conteúdo da linha em "linha2", do ficheiro original

## Se o elemento junto ao inicio da linha for diferente de 0, então guardar o conteúdo da linha em "linha1", do ficheiro original

## Se o elemento junto ao inicio da linha for diferente de 0, então guardar o conteúdo da linha em "linha2", do ficheiro original

## Se o elemento junto ao inicio da linha em diferente de 0, então guardar o conteúdo da linha em "linha2", do ficheiro original

## Se o elemento junto ao inicio da linha em diferente de 0, então guardar o conteúdo da linha em "linha2", do ficheiro original

## Se o
```

Figura 8: Código do exercício 2

A nossa implementação começa com uma cadeia de ifs e elses. Isto, traduzido, quer dizer que: se i1 for igual a i2, e f1 for igual a f2, então a função ficheiro não é necessária ser executada. No entanto, se a primeira condição for falsa mas f1 igual a f2, então não se deve, novamente, executar a função ficheiro, uma vez que daria um erro de divisão por 0. Por fim, se todos estas condições forem falsas, aí é executada a função ficheiro, disponibilizada pelo professor da disciplina.

Depois disto, parte-se para a função alteraTimestamp, que recebe todos os dados disponíveis como argumento. Aqui, percorre-se as legendas dos dois ficheiros de uma forma simultânea. Assim, retira-se o identificador de ambas as legendas, e compara-se. Se estes forem iguais e estiverem dentro do limite criado pelos argumentos ([i1, f1] para a legenda do ficheiro original, [i2, f2] para a legenda do ficheiro a alterar) e copia-se o timestamp da legenda do ficheiro original. Com todos os dados necessários, imprime-se todos eles: identificador da legenda do ficheiro alterado, novo timestamp e, também, a legenda do ficheiro alterado (o texto da legenda não sofre qualquer alteração neste processo).

Todas as legendas percorridas, obtemos que, dentro dos limites impostos pelo utilizador, as legendas com identificadores iguais possuem o mesmo tempo de início e de fim.

Como o grupo não conseguiu implementar a 100% as funções deste exercício, escolhemos traduzir a nossa linha de pensamento através de pseudo-código, como se pode observar na função alteraTimestamp. Deste modo, não somos capazes de fornecer um exemplo do resultado final, à semelhança de como foi feito no exercício anterior.

5 Conclusões

Com este trabalho prático, adquirimos e, maioritariamente, aprofundamos os nossos conhecimentos acerca da ferramenta AWK. Com isto, constatamos toda a utilidade que esta ferramenta nos pode oferecer, sendo que agora, com estes exercícios, conseguimos dominar melhor esta linguagem de programação de processamento de texto. Podemos afirmar que no primeiro exercício o grupo não sentiu grandes dificuldades na sua implementação, no entanto, o mesmo não se passou no exercício dois, pelas razões já referenciadas acima. Em suma, estamos satisfeitos com o trabalho realizado até então, sendo que o grupo se sente preparado para, após este desenvolvimento dos conhecimentos em AWK que irão ser importantes no futuro, aprender e estudar uma nova ferramenta de trabalho, na qual se vai basear o próximo projeto prático.