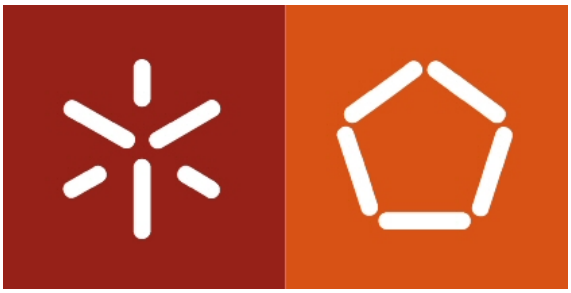


Universidade do Minho

Mestrado Integrado em Engenharia Informática

Jornal Angolano - Folha 8, v2

Trabalho Prático n.º 1 - FLex

Processamento de Linguagens

Alexandre Pacheco (A80760)
Diogo Sobral (A82523)
Inês Alves (A81368)

31 de Março de 2019

Resumo

No âmbito da Unidade Curricular de Processamento de Linguagens, foi proposto ao grupo, como forma de desenvolver os seus conhecimentos à cerca dos conteúdos lecionados nas aulas práticas, a realização de um conjunto de exercícios presentes no ficheiro fornecido. Neste ficheiro estavam presentes vários enunciados de diferentes trabalhos, sendo que o grupo ficou com o enunciado n.º 2: Jornal Angolano - Folha 8, v2.

Deste modo, o objetivo deste trabalho prático passa por aprofundar os conhecimentos relativos à Unidade Curricular em causa.

Conteúdo

1 Introdução

Estando o grupo a frequentar o 3.º ano do Mestrado Integrado em Engenharia Informática, foi-nos proposto, no contexto da Unidade Curricular de Processamento de Linguagens, que aprofundássemos os nossos conhecimentos na área. Posto isto, este projeto torna-se uma mais-valia no que toca ao aumento de experiência de uso do ambiente Linux e diversas ferramentas auxiliares, ao aumento da capacidade de escrever Expressões Regulares (ER) para descrição de padrões de frases e ainda, a partir destas, desenvolver Processadores de Linguagens Regulares, que filtrem ou transformem textos com base no conceito de regras de produção Condição-Ação. Por fim, também podemos verficar uma importância elevada na utilização do Flex para gerar filtros de texto em C.

Havendo 8 enunciados possíveis, o grupo abraçou o desafio de realizar o enunciado n.º 2: Jornal Angolano - Folha 8, v2, sendo a partir deste que se realizou todo o trabalho de aprofundamento dos conhecimentos, pondo em prática os conteúdos lecionados nas aulas práticas e teóricas da UC.

2 Normalização de Artigos

2.1 Enunciado

Como primeiro exercício proposto neste projeto, foi-nos pedido que procedêssemos à normalização e limpeza dos artigos que se encontravam no enunciado. Isto significa que o ficheiro que queremos trabalhar passará a estar organizado e pronto para ser, depois, transformado num ficheiro HTML.

Para isso, é importante vermos como se encontra o ficheiro input, a fim de entendermos, devidamente, que alterações irão ser necessárias, bem como estruturar a melhor estratégia a adotar, para realizar as modificações pretendidas.

Vejamos, a título de exemplo, o seguinte artigo que será, posteriormente, limpo e normalizado:

```
    <!-- ====================2014/2015-sera-um-ano-dificil-diz-o-presidente-igual-
    aos-outros-acrescenta-o-povo/index.html -->
<pub>
#TAG: tag:{Eduardo dos Santos} tag:{Petróleo} tag:{mensagem} tag:{preços}
#ID:{post-6243 post type-post status-publish format-standard has-post-thumbnail hentry
category-nacional tag-eduardo-dos-santos tag-petroleo tag-mensagem tag-precos}
Nacional

2015 será um ano difícil, diz o Presidente. Igual aos outros, acrescenta o Povo

PARTILHE VIA:
#DATE: [116eb] Redacção F8 | 29 de Dezembro de 2014
2015 será um ano difícil, diz o Presidente. Igual aos outros, acrescenta o Povo -
Folha 8

A baixa no preço do barril de petróleo, verificada desde Junho, está a levar o Executivo
de Eduardo
dos Santos a traçar estratégias para contornar as dificuldades desencadeadas. Ou seja,
com o preço
do petróleo em alta ou em baixa, serão sempre os mais pobres a pagar a factura.

O Presidente Eduardo dos Santos perspectivou para 2015, com uma originalidade quase
divina, um ano
difícil no plano económico, motivado pela \queda significativa do preço do petróleo
bruto", o que
vai levar à redução de algumas despesas públicas.

O \querido líder", que dirigiu nesta segunda-feira uma mensagem de ano novo à Nação,
apontou o
corte dos subsídios aos preços de combustíveis, como uma das reduções necessárias
para o próximo
ano.

\Há projectos que serão adiados e vão ser reforçados o controlo das despesas do Estado
e a
disciplina e parcimónia na gestão orçamental e financeira, para que se mantenha a
estabilidade",
disse o Presidente que está no poder há 35 anos, sublinhando que as dificuldades financeiras
não
vão interferir na política de combate à pobreza.
```

```
Habituados a (com)viver com a mentira institucional, os angolanos sabem por dura experiência
própria que o combate à pobreza continuará a ser um slogan, ao mesmo tempo que vão
aparecer mais
uns tantos multimilionários da safra presidencial.

Angola é o segundo maior produtor de petróleo da África subsaariana, depois da Nigéria,
e tem -
como sempre teve - uma economia fortemente dependente das receitas arrecadas com a
exportação
petrolífera. A baixa no preço do barril de petróleo, verificada desde Junho, está
a levar o
Executivo angolano a traçar estratégias de contorno ao actual momento.

O corte nos subsídios aos combustíveis em 2015, é uma delas, prevendo o Governo angolano
poupar
mais de 870 milhões de euros com essa medida. Com esta medida, que consta do Orçamento
Geral do
Estado (OGE) de 2015, o Governo prevê para o próximo ano \uma redução de cerca de
109,2 biliões de
kwanzas (mais de 870 milhões de euros) nos gastos com subsídios aos combustíveis",
para a mesma
quantidade de consumo de 2014.

Depois de um último ajustamento ao preço dos combustíveis, em Setembro passado, com
um aumento
médio de 25% ao consumidor no gasóleo e gasolina, na quarta-feira passada, registou-se
a um novo
reajustamento de 20% nos preços dos mesmos tipos de combustíveis.

Etiquetas: Eduardo dos SantosPetróleomensagempreços
</pub>
```

Como podemos verificar, os artigos encontram-se delimitados, superiormente e inferiormente, por *<pub>* e *</pub>*, respetivamente.

2.2 Descrição do problema

A partir da observação consciente e atenta do ficheiro que pretendemos normalizar, completamente desorganizado e desnormalizado, podemos começar a desenvolver ideias para atingir o objetivo. Ora, vemos aqui que todas as informações relevantes aparecem fora de ordem, pelo que é necessário implementar uma estrutura capaz de guardar estas informações para, posteriormente, serem colocadas no local correto.

Começando por pôr o raciocínio em prática, identificamos as seguintes informações cruciais:

- **ID:** identificador único de uma notícia
- **Tags:** etiquetas importantes da notícia
- **Categoria:** categoria da notícia
- **Título:** título da notícia
- **Data:** data da publicação da notícia
- **Texto:** corpo, propriamente dito, da notícia

Acontece que a ordem pela qual estas informações aparecem, não corresponde à ordem pela qual necessitamos que elas apareçam.

Ou seja, as informações aparecem pela seguinte ordem:

Figura 1: Ordem pela qual aparecem inicialmente as informações

E necessitamos que apareçam por esta ordem:

Figura 2: Ordem pela qual irão aparecer as informações

Numa reflexão mais atenta destas duas imagens, é possível verificar que não se encontra representada a informação referente ao texto, uma vez que esta está sempre no fim das restantes informações.

2.3 Estratégia de Implementação

Como já vimos, sabemos que estamos a analisar um artigo quando nos deparamos com a *tag* <*pub*> e, aquando da perceção desta *tag*, procede-se à escrita de "<*pub*", no ficheiro.

Tal como foi anteriormente citado, a informação requerida para transformar a notícia encontra-se desordenada, no entanto, esta aparece sempre pela mesma ordem, o que nos permite estabelecer estados e transmissões entre os mesmos. Neste sentido, após uma análise aprofundada do ficheiro *input* recebido, pudemos constatar que a informação aparecia pela seguinte ordem:

1. Aparição de <pub>
2. informação sobre as Tags → aparição de #TAG
3. informação sobre o id do post →aparição de #ID
4. informação sobre a categoria → linha que sucede a #ID
5. Título → 2 linhas após a informação da categoria
6. Data → aparição de #DATE
7. Texto → desde a linha que sucede o aparecimento de #DATE até ao fim do ficheiro, quando encontra <pub/>

Uma vez que cada informação tem os seus próprios padrões para ser recolhida e tal só deve ser feito mediante determinadas condições, decidimos criar um estado para cada informação a recolher.

O aparecimento da cláusula #TAG invoca, assim, o estado encarregue pelas *tags*. Este estado é responsável por tratar a informação de cada *tag*, de forma individual. Cada *tag* é apanhada através do uso da expressão regular:

tag:{ **texto a apanhar** }

Ora, imaginemos que temos esta frase que está a ser analisada no artigo:

#TAG: tag:{Aviação} tag:{Boeing 777-300ER} tag:{Companhia de bandeira} tag:{TAAG}

O que irá ser guardado no buffer é:

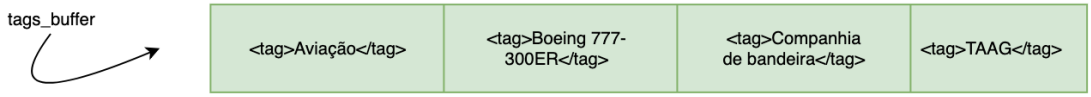


Figura 3: Representação do *buffer*

Até aqui o nosso ficheiro encontra-se representado da seguinte maneira, uma vez que a recolha das *tags* em nada afeta, para já, o ficheiro:

```
<pub
```

Quando a linha acaba, voltamos para o estado anterior, visto que a informação à cerca do *ID* da notícia pode ser encontrada sem recorrer a estados.

Após a recolha das *tags*, o próximo passo passa, então, por descobrir o *ID* da notícia. O *ID* aparece segundo o seguinte padrão:

#ID:{post-ID ... ...}

E estabelecemos a seguinte expressão regular:

```
#ID:\{[^ ]+\  {yytext[yyleng-1]='\0';fprintf(fp,"id=\"%s\">\n",yytext+5); BEGIN ID;}
```

Figura 4: Expressão regular de recolha dos *ID*s

Embora esta linha contenha muita informação, esta é irrelevante e, por isso, precisa de ser descartada. Para tal, damos início ao estado *ID* que acaba com o aparecimento de } seguido que um \n invocando o estado **CATEGORY**.

Portanto, continuando a situação, se, após aquela frase, tivermos a frase seguinte:

#ID:{post-1716 post type-post status-publish format-standard has-post-thumbnail hentry category-nacional tag-aviacao tag-boeing-777-300er tag-companhia-de-bandeira tag-taag}

O ficheiro passará a ter o seguinte aspeto:

```
<pub id="post-1716">
```

Depois desta, temos uma situação análoga à recolha das *tags*. Sabemos que a informação seguinte a guardar é a categoria da notícia, seguida do título da mesma, sendo as mesmas guardadas, na íntegra, em *buffers* e não havendo qualquer modificação no ficheiro.

Para ler a categoria, apenas temos que ler uma linha completa e, no fim, iniciar o estado **TITLE** para encontrar o título.

O título aparece entre duas linhas sem texto. De notar que a análise do ficheiro revelou que existem notícias que não possuem título e, nestes casos, aparece apenas uma linha solta e uma linha apenas com caracteres . Assim, temos o título quando encontramos o padrão:

\n **texto** \n \n

Caso este não apareça, então temos o padrão:

-+\n

e aparecerá no ficheiro a frase "Sem título".

Em ambas as situações voltamos ao estado inicial.

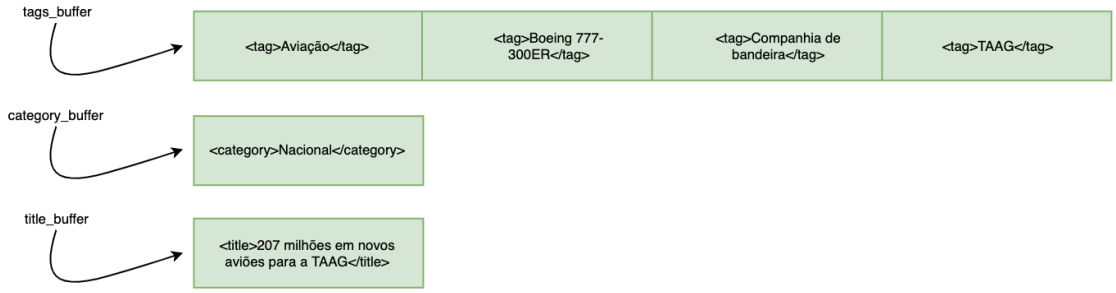


Figura 5: Representação dos *buffers*

Ao título, sucede a informação da data com a aparição da expressão "#DATE"que inicializa o estado **DATE**. Neste estado, apenas apanhamos a informação após o aparecimento do primeiro] que se refere à data. Por esta altura, já temos todos os *buffers* preenchidos com a informação necessária como podemos ver na figura 5. Uma vez que a informação não aparece pela ordem pretendida, é nesta altura que usamos todos os *buffers* para ordenar devidamente a informação.

Então, o nosso ficheiro tem agora um aspeto mais completo e organizado:

```
 <pub id="post-1716">
<title>207 milhões em novos aviões para a TAAG</title>
<author_date>Redacção F8 | 2 de Outubro de 2014</author_date>
<tags>
  <tag>Aviação</tag> <tag>Boeing 777-300ER</tag> <tag>Companhia de bandeira</tag>
<tag>TAAG</tag>
</tags>
<category>Nacional</category>
```

Para finalizar, quando é encontrado um \n trocamos de estado para o estado **TEXT**, que fica ativo até encontrarmos um <pub/>.

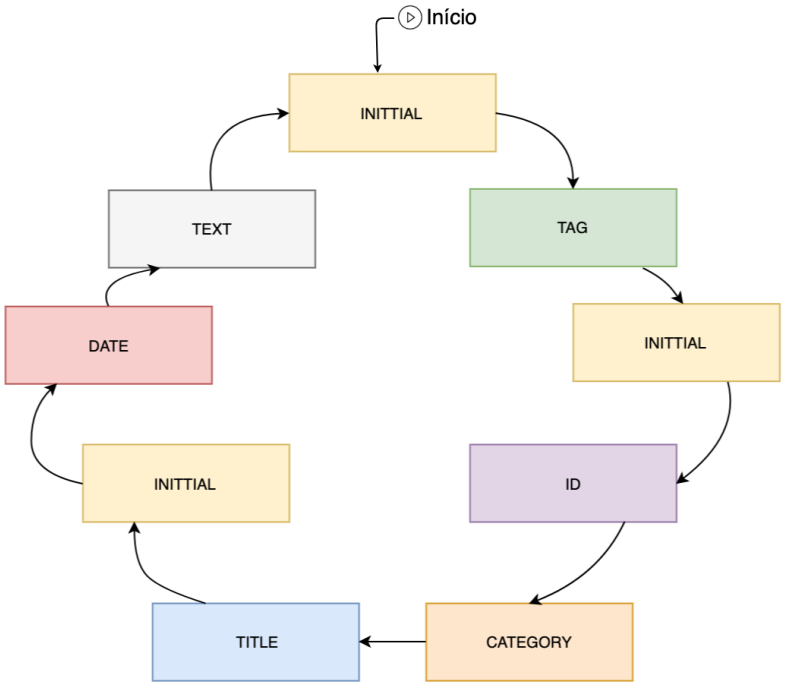


Figura 6: Ciclo de vida dos estados de uma notícia

2.4 Resultados obtidos

Explicado todo o processo de desenvolvimento deste exercício, podemos ver, agora, o resultado final:

```
    <pub id="post-1716">
   <title>207 milhões em novos aviões para a TAAG</title>
   <author_date>Redacção F8 | 2 de Outubro de 2014</author_date>
   <tags>
     <tag>Aviação</tag> <tag>Boeing 777-300ER</tag> <tag>Companhia de bandeira</tag>
<tag>TAAG</tag>
   </tags>
   <category>Nacional</category>
   <text>
[aaa16]

A transportadora aérea angolana TAAG vai contrair um empréstimo de 207 milhões de
euros para
adquirir duas aeronaves Boeing 777-300ER, de uma encomenda de três, segundo autorização
presidencial.

De acordo com um despacho presidencial de 1 de oOutubro, aquela empresa pública angolana
pode agora
avançar com este financiamento, para o pagamento antecipado da aquisição das duas
aeronaves, com
entregues previstas para Dezembro de 2015 e Março de 2016.

Trata-se de um contrato assinado entre a TAAG e a Boeing a 27 de Março de 2012, envolvendo
a
```

aquisição de três novos aparelhos, o primeiro dos quais já está ao serviço da companhia desde
Junho último.

Com este despacho, assinado pelo Presidente José Eduardo dos Santos, a TAAG é \autorizada" a
celebrar acordos de financiamento com o HSBC Bank, Banco de Negócios Internacional, Afrexim Bank e
um sindicato de bancos, nos montantes de 130.199.651 e 131.449.151 de dólares.

Estes valores, o equivalente a 207,2 milhões de euros, \destinam-se ao pagamento antecipado" da
aquisição das duas aeronaves ao fabricante norte-americano, lê-se no mesmo despacho.

Estas aeronaves têm capacidade para transportar 225 passageiros em classe económica, 56 em
executiva e 12 em primeira classe, possibilitando o acesso a telemóvel e internet a bordo.

A administração da TAAG disse anteriormente que o investimento nesta encomenda visa \consolidar os
destinos atuais", face a \algumas irregularidades no cumprimento de horário" e outras dificuldades
logísticas, podendo depois avançar com novas alternativas de destinos.

Esta autorização de financiamento surge na mesma semana em que { como o Folha 8 noticiou { foi
acordado, no Dubai, que a companhia aérea de bandeira angolana vai ser gerida pela congénere
Emirates ao abrigo de um acordo de parceria estratégica.

O acordo entre o Ministério dos Transportes de Angola e a administração da Emirates envolve um
contrato de gestão de topo da TAAG pela companhia dos Emirados Árabes Unidos. A Emirates passará a
nomear o Presidente do Conselho de Administração da TAAG e mais três administradores executivos, de
um total de nove elementos.

\Nascerá uma nova TAAG, que se pretende alinhada com os padrões e o estado da arte a nível
mundial", afirmou o ministro dos Transportes angolano, Augusto da Silva Tomás.

Angola vai indicar cinco dos elementos para a administração da transportadora aérea, mas de acordo
com informação transmitida no final da assinatura do acordo, a gestão corrente da TAAG será
assegurada por uma Comissão Executiva composta pelos administradores executivos nomeados pela
Emirates para as áreas Comercial, Operacional, Financeira e Administrativa.

</text>
</pub>

3 Criação de HTML

3.1 Enunciado

Após a limpeza e normalização dos artigos, é agora proposto ao grupo que implemente a criação de um ficheiro HTML por cada notícia presente nestes.

Para começar, é importante entender como está a estuturação e a vizualização gráfica dos artigos normalizados. A título de exemplo, vejamos a seguinte notícia que se encontra no enunciado deste projeto:

```
<pub id="post-6243">
   <title>2015 será um ano difícil, diz o Presidente. Igual aos outros, acrescenta
o Povo</title>
   <author_date>Redacção F8 | 29 de Dezembro de 2014</author_date>
   <tags>
     <tag>Eduardo dos Santos</tag> <tag>Petróleo</tag> <tag>mensagem</tag> <tag>preços</tag>
   </tags>
   <category>Nacional</category>
   <text>
2015 será um ano difícil, diz o Presidente. Igual aos outros, acrescenta o Povo -
Folha 8

A baixa no preço do barril de petróleo, verificada desde Junho, está a levar o Executivo
de Eduardo
dos Santos a traçar estratégias para contornar as dificuldades desencadeadas. Ou seja,
com o preço
do petróleo em alta ou em baixa, serão sempre os mais pobres a pagar a factura.

O Presidente Eduardo dos Santos perspectivou para 2015, com uma originalidade quase
divina, um ano
difícil no plano económico, motivado pela \queda significativa do preço do petróleo
bruto", o que
vai levar à redução de algumas despesas públicas.

O \querido líder", que dirigiu nesta segunda-feira uma mensagem de ano novo à Nação,
apontou o
corte dos subsídios aos preços de combustíveis, como uma das reduções necessárias
para o próximo
ano.

\Há projectos que serão adiados e vão ser reforçados o controlo das despesas do Estado
e a
disciplina e parcimónia na gestão orçamental e financeira, para que se mantenha a
estabilidade",
disse o Presidente que está no poder há 35 anos, sublinhando que as dificuldades financeiras
não
vão interferir na política de combate à pobreza.

Habituados a (com)viver com a mentira institucional, os angolanos sabem por dura experiência
própria que o combate à pobreza continuará a ser um slogan, ao mesmo tempo que vão
aparecer mais
uns tantos multimilionários da safra presidencial.

Angola é o segundo maior produtor de petróleo da África subsaariana, depois da Nigéria,
e tem {
```

```
como sempre teve { uma economia fortemente dependente das receitas arrecadas com a
exportação
petrolífera. A baixa no preço do barril de petróleo, verificada desde Junho, está
a levar o
Executivo angolano a traçar estratégias de contorno ao actual momento.

O corte nos subsídios aos combustíveis em 2015, é uma delas, prevendo o Governo angolano
poupar
mais de 870 milhões de euros com essa medida. Com esta medida, que consta do Orçamento
Geral do
Estado (OGE) de 2015, o Governo prevê para o próximo ano \uma redução de cerca de
109,2 biliões de
kwanzas (mais de 870 milhões de euros) nos gastos com subsídios aos combustíveis",
para a mesma
quantidade de consumo de 2014.

Depois de um último ajustamento ao preço dos combustíveis, em Setembro passado, com
um aumento
médio de 25% ao consumidor no gasóleo e gasolina, na quarta-feira passada, registou-se
a um novo
reajustamento de 20% nos preços dos mesmos tipos de combustíveis.

   </text>
</pub>
```

Posto isto, é fácil entender que cada notícia se encontra delimitada pela *tag "pub"*, em que *<pub id="...">* indica o início da notícia e *</pub>* indica que a notícia chegou ao fim. Note-se que em "..."deverá encontrar-se um identificador único de cada notícia, pelo que é um número variável de notícia para notícia.

Para além disso, é também pedido que se junte um link para o ficheiro com texto "título"em cada índice de "*tag*".

3.2 Descrição do Problema

Aquando da realização deste exercício, foi-nos possível destacar três estados, para além do estado inicial: **PUB**, **TAGS** e **DISCARD**, que estão representados no seguinte diagrama:

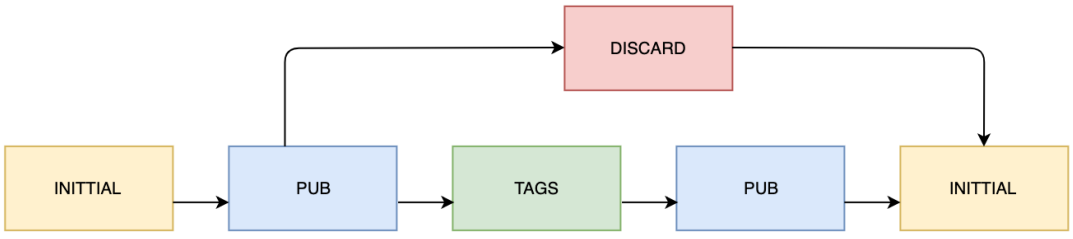


Figura 7: Ciclo de possíveis estados para cada notícia

Sendo assim, podemos afirmar que, dentro de cada notícia, iremos necessitar de encontrar as suas *tags* e o corpo, propriamente dito, da notícia (*text*).

De notar ainda que existem notícias repetidas, pelo que, com o inutito de eliminar repetições, criou-se o estado *discard*.

Passemos agora a uma breve descrição destes estados:

- **PUB:** significa que estamos perante uma notícia;
- **TAGS:** significa que estamos perante uma área de *tags* que pertencem a uma dada notícia;
- **DISCARD:** significa que foi encontrada uma notícia repetida, pelo que a sua análise não passaria de um desperdício de memória e tempo.

3.3 Estratégia de Implementação

Para a resolução deste exercício, foi necessária a criação de uma estrutura de dados nova para armazenar as *tags* encontradas. Para isso recorreu-se a uma árvore (*GTree*), uma vez que esta possibilitava a ordenação dos seus nodos conforme pretendíamos, facilitando assim a organização dos dados.

Começamos então por delimitar a nossa área de trabalho: a notícia. Como já foi referido, esta encontra-se delimitada por *tags*.

```
\<pub\  {BEGIN PUB;}
```

A primeira ação que é preciso ter em conta é a repetição de notícias, ou seja, caso o *ID* da notícia em que estamos a trabalhar já tenha sido lido antes, esta notícia não será tida em conta. Isto significa que, aquando da leitura da notícia nos encontramos no estado *PUB* e, assim que nos deparamos com uma situação de repetição, passamos para o estado *discard*, ignorando esta.

```
<PUB>id=\"[^\"]+ {if (searchPost(yytext+4)) {inithtml(yytext+4);} else {BEGIN DISCARD;}}
```

Deste modo, uma questão importante a reter aqui é: **como é que nos conseguimos aperceber desta repetição?** Ora, da mesma forma que utilizamos uma *GTree* para guardar as diversas *tags*, foi também criada uma que guarda os *ID*s das notícias. Então, quando nos deparamos com um *ID* que já foi previamente identificado, sabemos que já obtivemos informações à cerca dessa notícia, descartando-a e assegurando que as notícias não se repetem. Caso contrário, insere o *ID* da notícia na árvore e é criado um ficheiro html para a mesma com o respetivo nome. Para este efeito, dependemos plenamente da função *searchPost* que informa se o *ID* da notícia já foi averiguado antes ou não e da função *inithtml* para criar o ficheiro, inicializar as cláusulas do HTML e escrever num *buffer* toda a informação que mais tarde será necessária para fazer a ponte de ligação dos diferentes *posts* às respetivas *tags*.

De seguida, obtemos informações à cerca do título da notícia, da data da sua emissão, categoria da mesma e, obviamente, das suas *tags*. Estas são guardadas numa árvore, isto é, assim que nos deparamos com a *tag "<tags>"*, sabemos que, a partir dali, iremos encontrar uma ou mais *tags* referentes àquela notícia. Sendo assim, foi criada uma função *insertTagsLink* cuja função é guardar as *tags* na árvore. Notemos que esta árvore possui como chave uma *tag* e cada chave possui uma lista que contém os títulos das notícias que fazem uso daquela *tag*, juntamente com o *link* de acesso de cada uma dessas notícias.

```
void insertTagsLink(char * tag){
    GSList * list = g_tree_lookup(tags,tag);
    char * str = strdup(buffer_post);
    char * key = strdup(tag);

    list = g_slist_prepend(list,str);
    g_tree_insert(tags,key,list);
}
```

Por fim, sabemos que quando é encontrada a *tag* <*text*> irá dar-se início ao corpo da notícia, sendo que este apenas termina quando se encontra a *tag* </*text*>. Após esta, é encontrada a *tag* </*pub*>, simbolizando, como já foi referido, o fim de toda a notícia.

Vejamos agora como fucniona a escrita no ficheiro. Para começar, é importante recordarmos que toda a informação escrita entre as *tags* <*pub* e </*pub*> é escrita no ficheiro.
Além disso, todas as cláusulas encontradas, excetuando as que se encontram entre <tag> e </tag> e <text> e </text>, são escritas no ficheiro da seguinte maneira:

cláusula **texto que estava previamente entre as cláusulas**

Como exemplo deste procedimento, vejamos a seguinte imagem:

```
<p><b>Data: </b>Redacção F8 – 29 de Dezembro de 2014</p>
```

Figura 8: Código HTML usado para representar uma data

De seguida, ao encontrarmos a *tag* <*title*>, no ficheiro irá aparecer graficamente da seguinte forma:

título da notícia

A título de exemplo, para este caso, temos este código HTML gerado:

```
<h1>2015 será um ano difícil, diz o Presidente. Igual aos outros, acrescenta o Povo</h1>
```

Figura 9: Código HTML usado para representar um título

Já no caso da *tag* <*tags*> para além de ser escrita da forma descrita acima, inicia também uma lista com <ul>, terminando a mesma com .

lista de tags

Posto isto, o código HTML usado para este exemplo é:

```
<ul>
    <li>Eduardo dos Santos</li> <li>Petróleo</li> <li>mensagem</li> <li>preços</li>
</ul>
```

Figura 10: Código HTML usado para representar a lista de tags

```
<PUB>\<tags\> {fprintf(fp,"<b>Tags: </b>\n\t<ul>"); BEGIN TAGS;}
<TAGS>\<\/tags\> {fprintf(fp,"</ul>"); BEGIN PUB;}
```

No que toca à *tag* <*tag*>, aquando da escrita em ficheiro, reflete-se o seguinte comportamento:

palavra que se encontra à sua frente

Temos, então, o seguinte código HTML:

```
<li>Eduardo dos Santos</li>
```

Figura 11: Código HTML usado para representar uma tag presente na lista de tags

Correspondendo a este comportamento, temos o seguinte código:

```
<TAGS>\<tag\>[^\<]+ {fprintf(fp,"<li>%s",yytext); insertTagsLink(yytext+5);}
<TAGS>\<\/tag\> {fprintf(fp, "</li>");}
```

Por sua vez, quando nos apercebemos da existência da *tag* <*text*>, a ação tomada passa por escrever todas as palavras que se encontram a seguir desta *tag*, no entanto, há dois casos que são importantes de referir:

- quando é encontrado um .\n
- quando são encontrados um ou mais de dois \n seguidos

Nestas situações, o procedimento escolhido é escrever e, de seguida, .

```
<TEXT>\.[\n]+ {fprintf(fp, ".</p><p>");}
<TEXT>\n[\n]+ {fprintf(fp, "</p><p>");}
```

3.4 Resultados obtidos

Nesta secção iremos apresentar os resultados obtidos após a resolução e desenvolvimento deste exercício.

Posto isto, é fundamental percebermos que temos três pontos cruciais:

- Ficheiro input
- Ficheiro HTML resultante
- Visualização gráfica do ficheiro HTML

3.4.1 Ficheiro Input

Este ficheiro é passado como input com o objetivo de lhe ser aplicado o ficheiro *ex2.l*, que contém as ações a tomar. Tomemos como exemplo o ficheiro que se encontra no início do esclarecimento deste exercício, ou seja, a notícia cujo *ID* é 6243.

3.4.2 Ficheiro HTML resultante

Conteúdo do ficheiro HTML resultante, após lhe serem aplicadas as condições e expressões regulares já referidas neste relatório.

```
<HTML> <BODY> <meta charset='UTF-8'/><pub id="post-6243">
   <h1>2015 será um ano difícil, diz o Presidente. Igual aos outros, acrescenta o
Povo</h1>
   <p><b>Data: </b>Redacção F8 | 29 de Dezembro de 2014</p>
   <b>Tags: </b>
<ul>
     <li>Eduardo dos Santos</li> <li>Petróleo</li> <li>mensagem</li> <li>preços</li>
   </ul>
   <p><b>Categoria: </b>Nacional</p>
   <p>
2015 será um ano difícil, diz o Presidente. Igual aos outros, acrescenta o Povo -
Folha 8</p><p>A baixa no preço do barril de petróleo, verificada desde Junho, está
a levar o Executivo de Eduardo dos Santos a traçar estratégias para contornar as
dificuldades desencadeadas. Ou seja, com o preço do petróleo em alta ou em baixa,
serão sempre os mais pobres a pagar a factura.</p><p>O Presidente Eduardo dos Santos
perspectivou para 2015, com uma originalidade quase divina, um ano difícil no plano
económico, motivado pela \queda significativa do preço do petróleo bruto", o que vai
levar à redução de algumas despesas públicas.</p><p>O \querido líder", que dirigiu
nesta segunda-feira uma mensagem de ano novo à Nação, apontou o
corte dos subsídios aos preços de combustíveis, como uma das reduções necessárias
para o próximo ano.</p><p>\Há projectos que serão adiados e vão ser reforçados
o controlo das despesas do Estado e a disciplina e parcimónia na gestão orçamental
e financeira, para que se mantenha a estabilidade", disse o Presidente que está no
poder há 35 anos, sublinhando que as dificuldades financeiras não vão interferir na
política de combate à pobreza.</p><p>Habituados a (com)viver com a mentira institucional,
os angolanos sabem por dura experiência própria que o combate à pobreza continuará
a ser um slogan, ao mesmo tempo que vão aparecer mais uns tantos multimilionários
da safra presidencial.</p><p>Angola é o segundo maior produtor de petróleo da África
subsaariana, depois da Nigéria, e tem - como sempre teve - uma economia fortemente
dependente das receitas arrecadas com a exportação petrolífera. A baixa no preço do
barril de petróleo, verificada desde Junho, está a levar o Executivo angolano a traçar
estratégias de contorno ao actual momento.</p><p>O corte nos subsídios aos combustíveis
em 2015, é uma delas, prevendo o Governo angolano poupar
mais de 870 milhões de euros com essa medida. Com esta medida, que consta do Orçamento
Geral do Estado (OGE) de 2015, o Governo prevê para o próximo ano \uma redução de
cerca de 109,2 biliões de kwanzas (mais de 870 milhões de euros) nos gastos com subsídios
aos combustíveis", para a mesma
quantidade de consumo de 2014.</p><p>Depois de um último ajustamento ao preço dos
combustíveis, em Setembro passado, com um aumento médio de 25% ao consumidor no gasóleo
e gasolina, na quarta-feira passada, registou-se a um novo
reajustamento de 20% nos preços dos mesmos tipos de combustíveis.</p><p>   </p>
</BODY> </HTML>
```

3.4.3 Visualização gráfica do ficheiro HTML

Por fim, podemos, então, ver o resultado gráfico do ficheiro HTML gerado.

2015 será um ano difícil, diz o Presidente. Igual aos outros, acrescenta o Povo

Data: Redacção F8 — 29 de Dezembro de 2014

Tags:

- Eduardo dos Santos
- Petróleo
- mensagem
- preços

Categoria: Nacional

2015 será um ano difícil, diz o Presidente. Igual aos outros, acrescenta o Povo - Folha 8

A baixa no preço do barril de petróleo, verificada desde Junho, está a levar o Executivo de Eduardo dos Santos a traçar estratégias para contornar as dificuldades desencadeadas. Ou seja, com o preço do petróleo em alta ou em baixa, serão sempre os mais pobres a pagar a factura.

O Presidente Eduardo dos Santos perspectivou para 2015, com uma originalidade quase divina, um ano difícil no plano económico, motivado pela "queda significativa do preço do petróleo bruto", o que vai levar à redução de algumas despesas públicas.

O "querido líder", que dirigiu nesta segunda-feira uma mensagem de ano novo à Nação, apontou o corte dos subsídios aos preços de combustíveis, como uma das reduções necessárias para o próximo ano.

"Há projectos que serão adiados e vão ser reforçados o controlo das despesas do Estado e a disciplina e parcimónia na gestão orçamental e financeira, para que se mantenha a estabilidade", disse o Presidente que está no poder há 35 anos, sublinhando que as dificuldades financeiras não vão interferir na política de combate à pobreza.

Habituados a (com)viver com a mentira institucional, os angolanos sabem por dura experiência própria que o combate à pobreza continuará a ser um slogan, ao mesmo tempo que vão aparecer mais uns tantos multimilionários da safra presidencial.

Angola é o segundo maior produtor de petróleo da África subsaariana, depois da Nigéria, e tem – como sempre teve – uma economia fortemente dependente das receitas arrecadas com a exportação petrolífera. A baixa no preço do barril de petróleo, verificada desde Junho, está a levar o Executivo angolano a traçar estratégias de contorno ao actual momento.

O corte nos subsídios aos combustíveis em 2015, é uma delas, prevendo o Governo angolano poupar mais de 870 milhões de euros com essa medida. Com esta medida, que consta do Orçamento Geral do Estado (OGE) de 2015, o Governo prevê para o próximo ano "uma redução de cerca de 109,2 biliões de kwanzas (mais de 870 milhões de euros) nos gastos com subsídios aos combustíveis", para a mesma quantidade de consumo de 2014.

Depois de um último ajustamento ao preço dos combustíveis, em Setembro passado, com um aumento médio de 25% ao consumidor no gasóleo e gasolina, na quarta-feira passada, registou-se a um novo reajustamento de 20% nos preços dos mesmos tipos de combustíveis.

Figura 12: Resultado final

4 Criação de uma lista de TAGS

4.1 Enunciado

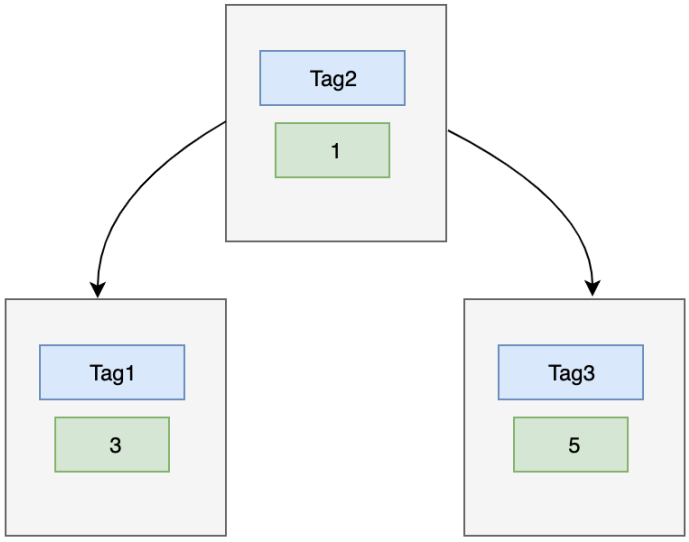
No último exercício proposto neste enunciado era-nos pedida a criação de uma lista de todas as *tags* encontradas no ficheiro, fazendo-se acompanhar do respetivo número de ocorrências.

Para isso, é importante reparar que o ficheiro que irá ser usado é o ficheiro que nos foi dado inicialmente, ou seja, é o ficheiro que não se encontra normalizado, nem limpo.

4.2 Descrição do problema

Ora, após já estarmos familiarizados com este projeto, foi fácil entender que abordagem tomar para realizar este exercício.

Resumidamente, o ponto fundamental neste exercício é ter uma estrutura de dados capaz de guardar todas as *tags* e ainda, cada uma dessas *tags*, ter uma variável associada capaz de fazer a contagem do seu número de ocorrências.



Para além disso, a fim de obter uma visualização gráfica apelativa, a ideia do grupo foi construir uma tabela que ilustrava, precisamente, todas as *tags* e os seus números de ocorrência.

4.3 Estratégia de Implementação

Percebido o objetivo deste exercício e planeado todo o desenvolvimento necessário para a correta realização do mesmo, foi decidido, de forma uniforme e consciente, assim como no exercício anterior, que a estrutura de dados adotada seria a *GTree*. Tal como sabemos, esta é ordenada alfabeticamente, pelo que melhora, substancialmente, a organização dos dados.

Assim, a estratégia de implementação deste exercício passou por escrever uma expressão regular que, ao ser capaz de identificar a existência de uma *tag*, a insere na árvore criada, *GTree *tags*, através da função *insereTag*, que recebe uma *tag*.

```
<TAG>tag:\{[^\}]+\} {nstr=strdup(yytext+5); nstr[yyleng-6]='\0'; insereTag(nstr);}
```

Figura 13: Expressão Regular responsável por identificar as diversas *tags*

Vejamos melhor o raciocínio utilizado nesta expressão. Começamos por nos aperceber que, quando fosse encontrada a expressão "#TAG:", iriamos ter, a seguir, uma ou mais *tags*, sendo

esta a expressão que desencadeia o alerta para o início da captura das *tags*. Cada uma destas *tags* encontra-se precedida da expressão *"tag:"*. Sendo assim, para obtermos as palavras que, de facto, se referem às *tags* que desejamos recolher, é necessário que apenas comecemos a reconhecer estas palavras após *"tag:"*.

tag:{**palavra a identificar**}

A fim de facilitar o exercício, optou-se por colocar a função *insereTag* como responsável por contar o número de ocorrências das diversas *tags* encontradas. Consequentemente, a função começa por verificar se a *tag* encontrada já existe ou não. Caso exista, apenas incrementa a variável encarregue de fazer a contagem das ocorrências dessa *tag*, mas, caso essa *tag* ainda não faça parte da árvore, ou seja, caso seja a primeira ocorrência encontrada da mesma, a função insere, ordenadamente, a *tag* e inicializa o seu contador com valor 1.

```
void insereTag(char * tag){
    gpointer value = g_tree_lookup(tags,tag);
    int * i;
    if (value == NULL){
        i = (int *) malloc(sizeof(int));
        *i = 1;
        g_tree_insert(tags,(gpointer)tag,(gpointer)i);
    }
    else{
        i = (int *) value;
        int k = *i;
        k++;
        (*i) = k;
    }
}
```

Figura 14: Função que insere uma *tag* na árvore

Por fim, tal como tinha sido planeado e já mencionado anteriormente neste relatório, foi unânime a decisão de criar uma tabela para ilustrar o resultado final deste exercício. Para isso, foi criada a função *initTable()*, cujos espaços a preencher são, ignorando a redundância, preenchidos com auxílio da função *imprimir_tag*.

```
void initTable(){
    fprintf(fp, "<style>table, th, td { border: 1px solid black;} </style>");
    fprintf(fp, "<table>");
    fprintf(fp, "<tr>");
    fprintf(fp, "<th>Tag</th>");
    fprintf(fp, "<th>Número de Ocorrências</th>");
    fprintf(fp, "</tr>");
}
```

Figura 15: Função responsável por criar a tabela

```
gboolean imprimir_tag(gpointer key, gpointer value, gpointer data){
    fprintf(fp, "<tr>");
    fprintf(fp, "<td>%s</td>\n",key);
    fprintf(fp, "<td>%d</td>",*((int *)value));
    fprintf(fp, "</tr>");

    return FALSE;
}
```

Figura 16: Função responsável por preencher as colunas da tabela

Assim, esta última função é utilizada na função pré-definida do **GLib** *g_tree_foreach*, que recebe a árvore onde se encontram guardadas todas as *tags* e a função *imprimir_tag*. É desta forma que é possível fazer a travessia da árvore e aplicar, a cada nodo, o segundo argumento desta função.

Por conseguinte, a função *imprimir_tag* vai fazer uso de uma *key*, a *tag* que se encontra no nodo que está a ser trabalhado, e de um *value*, que corresponde ao valor que se encontra associado à *tag*, ou seja, o número de ocorrências dessa *tag*. Posteriormente, com todas estas implementações, é criada a tabela desejada.

4.4 Resultados obtidos

Assim sendo, o resultado final deste exercício, do ponto de vista do grupo, tornou-se bastante agradável visualmente.

Uma vez que existiam demasiadas *tags*, encontra-se aqui representada apenas uma parte da tabela resultante:

Tag	Número de Ocorrências
"encruzilhada"	1
"santismo	1
+adres	1
100 anos	2
11 de novembro	2
127 anos	1
13 anos	2
14 anos	1
15 anos	2
15+2	3
1886	2
1961	4
1975	2
1977	2
1993	2
1995	2
1º de Agosto	4
20 anos	6
200 melhores universidades do mundo	2
200 milhões	1
2008	2
2010	2
2012	6
2014	13
2015	11
2016	12
2017	24

Figura 17: Resultado final

5 Conclusão e Análise Crítica

Perante este primeiro trabalho prático da Unidade Curricular de Processamento de Linguagens, o grupo deparou-se com diversos obstáculos. A fim de serem ultrapassados, procuramos sempre seguir as linhas de raciocínio apresentadas nas aulas práticas, chegando sempre, deste modo, a um consenso à cerca da abordagem a tomar.

Apesar dos entraves apresentados, a elaboração deste projeto revelou-se, sem dúvida, uma mais valia no que toca à consolidação de conhecimentos, levando todos os elementos do grupo a terem uma parte ativa no desenvolvimento do mesmo, promovendo assim o trabalho de equipa. Para isso, foi necessário uma aprendizagem atenta à cerca de Expressões Regulares e da linguagem de filtragem de texto FLex, a fim de facilitar a realização, o mais correta possível, deste trabalho.

Com o enunciado que nos ficou destinado, enunciado n.º 2: Jornal Angolano - Folha 8, v2, consideramos que foi possível demonstrar o nosso entendimento em relação ao tema.

Posto isto, fazemos uma avaliação bastante positiva do nosso desempenho neste projeto já que, apesar das dificuldades encontradas aquando de todo o desenvolvimento deste trabalho, o grupo as superou com distinção, respondendo de forma autónoma e correta a todas as questões apresentadas e todos os propósitos iniciais foram alcançados, revelando um trabalho consistente e bem estruturado.