UNIVERSIDADE DO MINHO
ESCOLA DE ENGENHARIA

Processamento de Linguagens

TRABALHO PRÁTICO NR 1

Ana Marta Santos Ribeiro A82474
Carla Isabel Novais da Cruz A80564
Jéssica Andreia Fernandes Lemos A82061

MIEI - 3º Ano - 2º Semestre

Braga, 31 de março de 2019

1 Resumo

Este relatório tem como objetivo documentar o primeiro trabalho prático proposto na unidade curricular de *Processamento de Linguagens* utilizando como auxílio a ferramenta *Flex* para gerar filtros de texto.

Assim, primeiro será apresentada uma introdução ao projeto, bem como uma descrição do seu enunciado. De seguida é explicado toda a implementação da solução, nomeadamente as estruturas de dados utilizadas, os filtros de textos necessários, o modo de criação dos ficheiros HTML e a criação de índices. Serão ainda ilustrados exemplos de utilização. Finalmente será realizada uma apreciação crítica ao trabalho desenvolvido.

Índice

2 Introdução

No âmbito da Unidade Curricular de *Processamento de Linguagens*, foram desenvolvidos filtros de modo a separar um ficheiro *.txt* fornecido em diversos HTML. Para tal foi necessário aplicar o conhecimento adquirido relativamente a *Flex* e expressões regulares. De modo a que este trabalho fosse realizado, também foi essencial saber manipular com facilidade a biblioteca *Glib*.

Iremos apresentar toda a linha de pensamento e estratégias utilizadas para a concretização dos objetivos pretendidos pelo exercício 2 neste trabalho prático que consiste em criar o HTML correspondente a cada notícia do Jornal Angolano. Será ainda criado uma lista de tags encontradas em todas estas e as respetivas ocorrências.

3 Estrutura do Relatório

Neste relatório serão apresentados todos os passos desde a apresentação do enunciado e descrição do problema, passando pela apresentação da forma como implementamos a solução, até aos exemplos de utilização e do resultado final obtido. Assim, serão indicadas as estruturas de dados que optamos bem como os filtros de texto utilizados de modo a armazenarmos a informação necessária para a criação dos ficheiros HTML e dos índices.

4 Análise e especificação

4.1 Enunciado e Descrição do problema

O enunciado escolhido para este projeto foi o segundo, Jornal Angolano - Folha 8 v2. Assim, é necessário limpar e normalizar cada um dos artigos contidos no ficheiro *.txt*, bem como criar um ficheiro HTML para cada notícia. Para além disso, é preciso criar um índice para cada tag com os títulos dos artigos, isto é, associar a cada um, o link para as notícias a que pertence. Foi ainda necessário elaborar outro índice com todas as tags encontradas e o respetivo número de ocorrências.

É importante referir, que optamos por implementar este índice com um link para aquele que contém associado todos os artigos de uma tag.

4.2 Objetivos com a realização do Trabalho Prático

Este trabalho prático, com a resolução do exercício proposto, tem como principais objectivos:

- Aumentar a capacidade de escrever Expressões Regulares para a descrição de padrões de frases;
- Filtrar e transformar textos com base no conceito de regras de produção *Condição-Ação*;
- A utilizaçao do *Flex* para gerar filtros de texto em C.

5 Implementação da Solução

5.1 Estrutura de Dados

Para a realização deste trabalho implementamos três estruturas, nomeadamente uma lista ligada, uma árvore que em cada posição contém uma lista ligada e uma tabela de hash. Estas foram implementadas recorrendo à utilização da biblioteca *Glib*, que se revelou eficaz no armazenamento de dados.

De modo a construir o HTML pela ordem correta, foi necessário armazenar a informação antes de a escrever para o ficheiro. Como cada um possui diversas tags, optamos por armazená-las numa lista ligada.

De forma a elaborar o índice com as tags e o número de ocorrências de cada uma, decimos criar uma árvore binária em que a chave é a tag (char*) e o valor é uma lista ligada, com a informação relativa aos artigos em que esta se encontra. A lista contém o título e o identificador do artigo, como se pode observar de seguida. É importante referir, que o número de ocorrências de cada tag corresponde ao número de elementos da lista ligada.

```
1 typedef struct listArticles {
2     char *article;
3     char *title;
4 } *ListArticles;
```

Como o ficheiro *.txt* pode conter várias notícias repetidas, foi necessário não as considerar. Assim, decidimos criar uma tabela de hash com o identificador de todos os artigos já processados, de modo a verificar se uma notícia deve ser ignorada ou processada. Neste caso, optamos pela tabela de hash, uma vez que a procura torna-se muito eficaz.

5.2 Filtros de Texto

Para a resolução dos problemas apresentados e de modo a ser possível o desenvolvimento de uma solução para o problema proposto, foi necessário filtrar as diversas notícias do ficheiro *.txt* fornecido. As decisões tomadas na escolha dos filtros a utilizar foi deveras importante pois as expressões regulares utilizadas com o *Flex* serão utilizadas na limpeza e normalização de todo o documento.

Para filtrar as tags, o identificador do post e a data_autor, o racícionio aplicado foi semelhantes para os três, começando por filtrar estas palavras do ficheiro pois a informação necessária estaria depois desta inicial. Por exemplo, para a tag, utilizamos TAG[ ]*:[ ]*, em que ignorará isto e os espaços que possam aparecer a seguir e aplica-se a ação de iniciar o estado DEFtag. Dentro disto são aplicadas expressões que irão filtrar o que está entre as duas chavetas, que serão as tags e estas são guardadas na estrutura.

Para a filtragem do título e da categoria, o racícionio nestes foi parecido, sendo o que está entre dois \n distintos a informação a reter. Por exemplo, após a seleção do identificador do post, quando for encontrado um \n após o término da chaveta, será inicializado o DEFcategoria. Esta termina quando encontra outro \n. Quando este é encontrado coloca a categoria na estrutura pretendida e é feito o BEGIN do DEFtitulo. Este guarda o título e executa o BEGIN INITIAL quando outro \n é capturado.

De seguida, é filtrado o texto e caso este contenha etiquetas, estas são filtradas. Quando a notícia chega ao fim, esta é escrita num ficheiro HTML .

Com o auxílio das estruturas apresentadas anteriormente e a utilização destes filtros, irá ser possível a obtenção de um ficheiro HTML e de índices, que serão descritos futuramente.

5.3 Criação dos ficheiros HTML

Com o intuito de gerar um ficheiro HTML para cada notícia à medida que efetuamos a leitura destes criamos um ficheiro deste tipo, cujo nome será o identificador da mesma. Para a elaboração destes ficheiros recorremos à função *createHTMLFile*. Após a criação deste, podemos escrever as informações referentes à notícia que foram obtidas através dos filtros anteriormente apresentados. Assim sendo o ficheiro HTML de cada post irá conter o título, a categoria, a data, as tags que este contêm bem como o texto da notícia. Desta forma, com o auxílio da função *articleToHTML* acedemos aos dados armazenados bem como à estrutura que contém as tags desse artigo. É de destacar que nesta função temos em atenção se já existia um ficheiro HTML dessa notícia, de modo a não se criar o mesmo novamente. Optamos por no fim da notíca criar um link para ser possível regressar ao índice principal, através do auxílio do *href*.

5.4 Índices

5.4.1 Índice de Tags

De modo a construir um ficheiro HTML com todas as tags existentes e o respetivo número de ocorrências, recorremos às estruturas anteriormente referidas. Para tal optamos por apresentar a informação na forma de uma tabela, sendo a primeira coluna o número de ocorrências e a segunda a tag. Tendo em conta que pretendiamos ao selecionar uma determinada tag aceder a um índice que contém todos os títulos dos artigos a que esta pertence foi necessário criar um link através do *href*.

5.4.2 Índice de Títulos

Como já foi referido anteriormente, criamos ficheiros HTML para cada uma das tags que será constituído pelos títulos de cada notícia que a contém. De modo a aceder às notícias foi criado um link. Para além disso, damos a possibilidade de regressar ao índice anterior.

6 Exemplos de utilização

De modo a demonstrar os pontos referidos ao longo do relatório apresentamos de seguida o índice de tags, que apresenta o número de ocorrências bem como a respetiva tag.

Tags

Occurrences	Tag
1	+adres
1	11 de novembro
1	13 anos
1	1º de Agosto
1	2012
3	2014
2	2015
2	2017
1	2018
3	27 de maio
1	40 anos
1	404 quilates
1	42 anos
1	5 de Outubro
1	82 anos
1	98%
1	abastecimento
1	Abel Chivukuvuku
1	Abel Chivukuvuku. mentiras
1	aberração
1	aborto
1	absolvição

Figura 1. Índice de Tags

Selecionando a tag que se pretende acedemos ao seguinte índice que contém todos os títulos das notícias com essa mesma tag.

Articles - Polícia

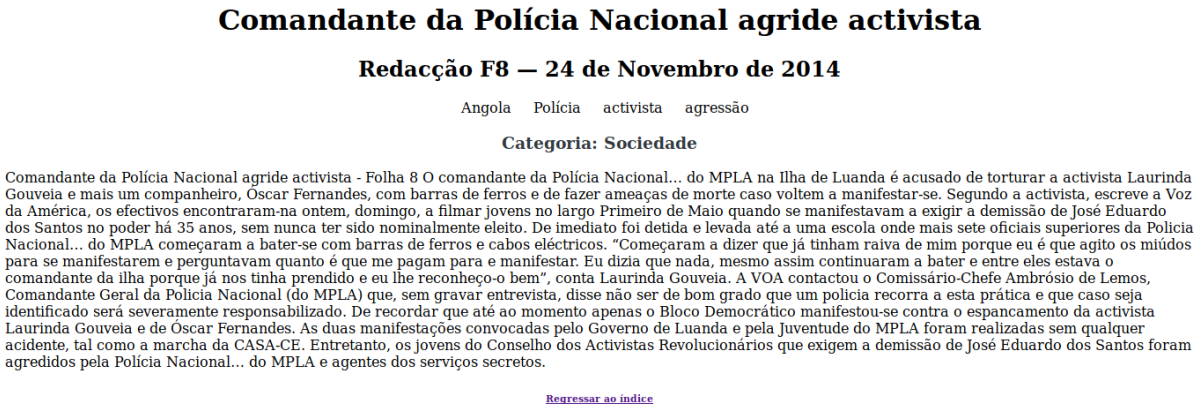
Indice:

- Boletins de voto roubados em Manica
- Comandante da Polícia Nacional agride activista
- A Martina Sinfona
- SIC apresentou supostos sequestradores
- Testemunhas descrevem o assassino de Rufino
- Tortura e homicídio: sofrer e morrer às mãos da Polícia
- Um polícia injustiçado
- O aborto, o MPLA e a Polícia
- Agredir jornalistas está
- Andam a tirar os dísticos
- CASA-CE denuncia violência dos donos do reino, o MPLA
- Polícia agride jornalista

Regressar ao índice

Figura 2. Índice de títulos das notícias com a tag "Polícia"

Agora é possível selecionar um dos artigos ou regressar ao índice principal. Selecionando uma das notícias, no caso apresentado a "Comandante da Polícia Nacional agride activista", é possível visualizar a notíca toda.



Comandante da Polícia Nacional agride activista

Redacção F8 — 24 de Novembro de 2014

Angola Polícia activista agressão

Categoria: Sociedade

Comandante da Polícia Nacional agride activista - Folha 8 O comandante da Polícia Nacional... do MPLA na Ilha de Luanda é acusado de torturar a activista Laurinda Gouveia e mais um companheiro, Óscar Fernandes, com barras de ferros e de fazer ameaças de morte caso voltem a manifestar-se. Segundo a activista, escreve a Voz da América, os efectivos encontraram-na ontem, domingo, a filmar jovens no largo Primeiro de Maio quando se manifestavam a exigir a demissão de José Eduardo dos Santos no poder há 35 anos, sem nunca ter sido nominalmente eleito. De imediato foi detida e levada até a uma escola onde mais sete oficiais superiores da Policia Nacional... do MPLA começaram a bater-se com barras de ferros e cabos eléctricos. "Começaram a dizer que já tinham raiva de mim porque eu é que agito os miúdos para se manifestarem e perguntavam quanto é que me pagam para e manifestar. Eu dizia que nada, mesmo assim continuaram a bater e entre eles estava o comandante da ilha porque já nos tinha prendido e eu lhe reconheço-o bem", conta Laurinda Gouveia. A VOA contactou o Comissário-Chefe Ambrósio de Lemos, Comandante Geral da Policia Nacional (do MPLA) que, sem gravar entrevista, disse não ser de bom grado que um policia recorra a esta prática e que caso seja identificado será severamente responsabilizado. De recordar que até ao momento apenas o Bloco Democrático manifestou-se contra o espancamento da activista Laurinda Gouveia e de Óscar Fernandes. As duas manifestações convocadas pelo Governo de Luanda e pela Juventude do MPLA foram realizadas sem qualquer acidente, tal como a marcha da CASA-CE. Entretanto, os jovens do Conselho dos Activistas Revolucionários que exigem a demissão de José Eduardo dos Santos foram agredidos pela Polícia Nacional... do MPLA e agentes dos serviços secretos.

Regressar ao índice

Figura 3. Notícia do ficheiro post-4980.html

7 Conclusão

Este projeto permitiu-nos consolidar a matéria lecionada nas aulas relativamente a *Flex*. Para além disto, constatamos o seu grande poder expressivo e versatibilidade. Esta ferramenta revelou-se muito eficaz para filtrar e tratar informação.

Numa primeira fase tornou-se fundamental desenvolver os filtros de texto necessários para filtrar e armazenar as informações necessárias para a resolução do problema, o que se tornou complicado inicialmente dado que nem todos os artigos possuíam a mesma estrutura. Para além disto, tornou-se imperativo ter cuidado com vários aspetos tais como artigos repetidos, sem identificadores, títulos ou tags. Desta forma, foi essencial ter em atenção a escolha das estruturas de dados a implementar no projeto.

Por fim, geramos os ficheiros HTML para cada uma das notícias e desta forma conseguimos criar um índice com todas as tags bem como as ocorrências de cada, sendo possível aceder ao índice onde se encontram todos os artigos que contêm uma determinada tag.

Em suma, destaca-se a importância da escolha das estruturas de dados que tem impacto direto na eficácia do programa bem como das expressões regulares tendo em conta que manipulam toda a informação das notícias.