U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

MESTRADO INTEGRADO EM ENGENHARIA INFORMÁTICA E COMPUTAÇÃO | 2º ANO
EIC0027 | SISTEMAS OPERATIVOS | 2015/2016

TRABALHO PRÁTICO Nº 1

Eliminador de ficheiros duplicados

Introdução e objetivos

Num sistema de ficheiros, verifica-se frequentemente a existência de ficheiros regulares duplicados, isto é, de várias cópias de um mesmo ficheiro, existentes em diferentes diretórios. A duplicação de ficheiros conduz a desperdício de espaço e, potencialmente, a inconsistências de informação, quando se fazem, inadvertidamente, alterações em apenas uma das cópias.

O objetivo final deste trabalho é desenvolver uma aplicação que permita detetar a existência de ficheiros duplicados e que mantenha apenas um exemplar desses ficheiros, substituindo as restantes ocorrências por "hard links" para aquele exemplar.

Através da sua realização, pretende-se proporcionar a familiarização com a programação de sistema, em ambiente Linux, envolvendo, principalmente, a manipulação de ficheiros e diretórios, o desenvolvimento de aplicações multiprocesso e invocação de programas externos.

Especificação do trabalho

A aplicação a desenvolver deve satisfazer os seguintes requisitos:

- O código executável terá o nome **`rmdup`** e será lançado em execução através do comando **`rmdup <dir>`** em que **`<dir>`** representa o nome de um diretório que contém os ficheiros a processar. Este diretório pode conter sub-diretórios, constituindo uma árvore de directórios.
- Deve ser uma aplicação multiprocesso, lançando em execução um processo por cada diretório da árvore de diretórios, incluindo o diretório indicado como argumento da linha de comando.
- Para detetar os ficheiros duplicados, todos os processos devem guardar num ficheiro de texto comum, **`files.txt`**, informação sobre os ficheiros regulares detetados nos diretórios por eles processados: nome do ficheiro, tamanho, permissões de acesso, data de modificação, …
- Considera-se que um ficheiro é um duplicado de outro se se verificarem simultaneamente as seguintes condições:
  - ambos são ficheiros regulares;
  - ambos têm o mesmo nome;
  - ambos têm as mesmas permissões de acesso;
  - ambos têm o mesmo conteúdo.
- Após todos os processos detetores de ficheiros regulares terem terminado, o processo inicial deve identificar todos os conjuntos ficheiros duplicados e, para cada conjunto, substituir os ficheiros mais recentes por um “hard link” para o ficheiro mais antigo, que não será tocado. Nota: a substituição deve ser feita por forma a evitar perdas de dados.
- No final, no diretório **`<dir>`**, deve ser criado um ficheiro de texto, **`hlinks.txt`**, contendo uma listagem dos "hard links" que foram criados.

Em anexo, é apresentado um exemplo das alterações introduzidas na estrutura de ficheiros de um diretório **`d1`**, após a execução do comando **`rmdup d1`**.

Notas sobre o desenvolvimento

- Fazer testes de erro nas chamadas ao sistema e usar sempre a opção de compilação **`-Wall`**, por forma a garantir que a compilação dos programas não dá origem a avisos (*warnings*).
- Começar por desenvolver:
  - uma aplicação, **`lsdir`**, que gera uma listagem dos ficheiros regulares de um diretório;
  - outra aplicação, multiprocesso, que percorre a árvore de diretórios, lançando em execução um processo **`lsdir`** por cada diretório encontrado.
- Procurar tirar partido dos utilitários da *shell* (ex: **`sort`** ou outros).
- Não podem ser usadas chamadas **`system`**, nem o equivalente **`execlp("sh","sh","-c", …)`**.
- A formatação do ficheiro **`files.txt`** pode ser escolhida livremente.
- Ter em conta que pode haver diretórios para os quais não exista permissão de acesso.

Entrega do trabalho

- Data limite para a entrega do trabalho: 2016/04/18, às 12:00h.
- Oportunamente serão publicadas algumas regras para a entrega do trabalho, na página de "Sistemas Operativos", no Moodle da Universidade do Porto.

ANEXO

Antes da execução do comando **`rmdup d1`** :

```
user007@ubuntu:~/SOPE/d1$ ls -laiR
.:
total 28
936431 drwxrwxr-x  4 user007 user007 4096 Mar  2 17:03 .
934234 drwxrwxr-x 18 user007 user007 4096 Mar  2 15:41 ..
936439 drwxrwxr-x  4 user007 user007 4096 Mar  2 16:48 d2
936440 drwxrwxr-x  3 user007 user007 4096 Mar  2 15:42 d3
936444 -rw-r--r--  1 user007 user007 9993 Mar  2 17:03 f1.pdf
936445 -rw-r--r--  1 user007 user007 7595 Mar  2 15:46 f2.dat
936446 -rw-r--r--  1 user007 user007 2210 Mar  2 15:50 f3.txt

./d2:
total 28
936439 drwxrwxr-x 4 user007 user007 4096 Mar  2 16:48 .
936431 drwxrwxr-x 4 user007 user007 4096 Mar  2 16:03 ..
936441 drwxrwxr-x 2 user007 user007 4096 Mar  2 15:49 d4
936442 drwxrwxr-x 2 user007 user007 4096 Mar  2 15:50 d5
936569 -rw-r--r-- 1 user007 user007 9993 Mar  3 16:48 f1.pdf
936567 -rw-r--r-- 1 user007 user007 4577 Mar  3 15:47 f4.txt
936568 -rw-r--r-- 1 user007 user007 6713 Mar  3 15:47 f5.txt

./d2/d4:
total 16
936441 drwxrwxr-x 2 user007 user007 4096 Mar  2 15:49 .
936439 drwxrwxr-x 4 user007 user007 4096 Mar  2 15:48 ..
937250 -rw-r--r-- 1 user007 user007 7595 Mar  4 15:49 f2.dat
937251 -rw-rw-r-- 1 user007 user007 9817 Mar  4 15:49 f7.txt

./d2/d5:
total 16
936442 drwxrwxr-x 2 user007 user007 4096 Mar  2 15:50 .
936439 drwxrwxr-x 4 user007 user007 4096 Mar  2 15:48 ..
937752 -rwxr-xr-- 1 user007 user007 7898 Mar  5 15:50 f8
937753 -rwxr-xr-- 1 user007 user007 8910 Mar  5 15:50 f9

./d3:
total 20
936440 drwxrwxr-x 3 user007 user007 4096 Mar  2 16:25 .
936431 drwxrwxr-x 4 user007 user007 4096 Mar  2 16:03 ..
936443 drwxrwxr-x 2 user007 user007 4096 Mar  2 16:03 d6
936457 -rw-r--r-- 1 user007 user007 5678 Mar  7 16:24 f1.pdf
936458 -rw-r--r-- 1 user007 user007 7711 Mar  7 16:25 f5.txt

./d3/d6:
total 20
936443 drwxrwxr-x 2 user007 user007 4096 Mar  2 16:03 .
936440 drwxrwxr-x 3 user007 user007 4096 Mar  2 15:42 ..
936580 -rw-rw-r-- 1 user007 user007 9993 Mar  5 15:51 f1.pdf
937257 -r--r--r-- 1 user007 user007 6713 Mar  4 15:51 f5.txt
936431 -rw-rw-r-- 1 user007 user007 9817 Mar  7 15:52 f7.txt
user007@ubuntu:~/SOPE/d1$
```

Após a execução do comando **`rmdup d1`** :

```
user007@ubuntu:~/SOPE/d1$ ls -laiR
.:
total 28
936431 drwxrwxr-x  4 user007 user007 4096 Mar  2 17:03 .
934234 drwxrwxr-x 18 user007 user007 4096 Mar  2 15:41 ..
936439 drwxrwxr-x  4 user007 user007 4096 Mar  2 16:48 d2
936440 drwxrwxr-x  3 user007 user007 4096 Mar  2 15:42 d3
936569 -rw-r--r--  1 user007 user007 9993 Mar  2 17:03 f1.pdf
936445 -rw-r--r--  1 user007 user007 7595 Mar  2 15:46 f2.dat
936446 -rw-r--r--  1 user007 user007 2210 Mar  2 15:50 f3.txt
938001 -rw-r--r--  1 user007 user007  696 Mar  2 15:50 files.txt
938007 -rw-r--r--  1 user007 user007  127 Mar  2 15:50 hlinks.txt

./d2:
total 28
936439 drwxrwxr-x 4 user007 user007 4096 Mar  2 16:48 .
936431 drwxrwxr-x 4 user007 user007 4096 Mar  2 16:03 ..
936441 drwxrwxr-x 2 user007 user007 4096 Mar  2 15:49 d4
936442 drwxrwxr-x 2 user007 user007 4096 Mar  2 15:50 d5
936569 -rw-r--r-- 1 user007 user007 9993 Mar  3 16:48 f1.pdf
936567 -rw-r--r-- 1 user007 user007 4577 Mar  3 15:47 f4.txt
936568 -rw-r--r-- 1 user007 user007 6713 Mar  3 15:47 f5.txt

./d2/d4:
total 16
936441 drwxrwxr-x 2 user007 user007 4096 Mar  2 15:49 .
936439 drwxrwxr-x 4 user007 user007 4096 Mar  2 15:48 ..
936445 -rw-r--r-- 1 user007 user007 7595 Mar  4 15:49 f2.dat
936431 -rw-rw-r-- 1 user007 user007 9817 Mar  4 15:49 f7.txt

./d2/d5:
total 16
936442 drwxrwxr-x 2 user007 user007 4096 Mar  2 15:50 .
936439 drwxrwxr-x 4 user007 user007 4096 Mar  2 15:48 ..
937752 -rwxr-xr-- 1 user007 user007 7898 Mar  5 15:50 f8
937753 -rwxr-xr-- 1 user007 user007 8910 Mar  5 15:50 f9

./d3:
total 20
936440 drwxrwxr-x 3 user007 user007 4096 Mar  2 16:25 .
936431 drwxrwxr-x 4 user007 user007 4096 Mar  2 16:03 ..
936443 drwxrwxr-x 2 user007 user007 4096 Mar  2 16:03 d6
936457 -rw-r--r-- 1 user007 user007 5678 Mar  7 16:24 f1.pdf
936458 -rw-r--r-- 1 user007 user007 7711 Mar  7 16:25 f5.txt

./d3/d6:
total 20
936443 drwxrwxr-x 2 user007 user007 4096 Mar  2 16:03 .
936440 drwxrwxr-x 3 user007 user007 4096 Mar  2 15:42 ..
936580 -rw-rw-r-- 1 user007 user007 9993 Mar  5 15:51 f1.pdf
937257 -r--r--r-- 1 user007 user007 6713 Mar  4 15:51 f5.txt
936431 -rw-rw-r-- 1 user007 user007 9817 Mar  7 15:30 f7.txt
user007@ubuntu:~/SOPE/d1$
```

Notas:

- Os nomes duplicados foram coloridos, *a posteriori*, com a mesma cor, para mais fácil identificação.
- Alguns dos nomes duplicados não correspondem a ficheiros duplicados, tendo em conta os critérios anteriormente indicados.
- Na coluna da direita, os "i-nodes" correspondentes a ficheiros duplicados estão coloridos; os que foram alterados, através da criação de um "hard link", estão coloridos e sublinhados.