Computação em Larga Escala

General Problems – Algorithmic analysis 2

António Rui Borges

Summary

- *Text processing in Portuguese*
  - *Algorithm* (*top-down approach*)
  - *Contents of* `text0.txt`
  - *Compilation command*
  - *Processing results*

Text processing in Portuguese - 1

text file as a byte stream

- text files should be processed in succession and viewed as byte streams because each UTF-8 character encompasses from one to four bytes
- operations to be carried out on stream access are among to the following: *fopen*, *fclose*, *feof*, *fread*, *fwrite*, fseek, *ftell*
- so an important operation to be assumed first is the one that extracts an UTF-8 character from the stream

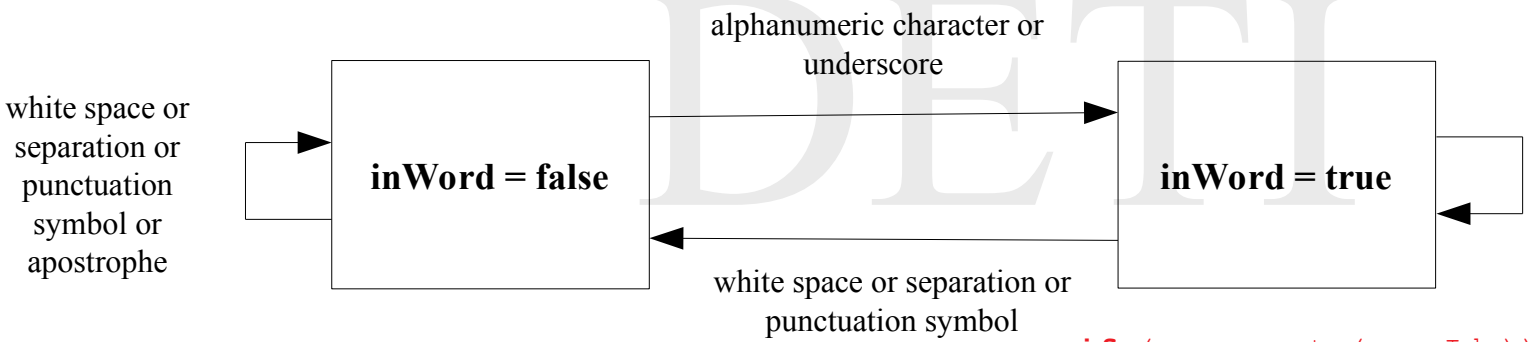
Text processing in Portuguese - 2

Algorithm

```
for each textFile
{ file = open (textFile);
  inWord = false;
  nWords = 0;
  nWordsWMultCons = 0;
  while (extractAChar (file, UTF8Char) != EOF)
    processAChar (UTF8Char, inWord, nWords, nWordsWMultCons);
  close (file);
  printResults (textFileName, nWords, nWordsWMultCons);
}
```

Text processing in Portuguese - 3

```
nWords += 1
setToZero (nConsOccur)
detMultCons = false
if (consonant (consIdx))
   nConsOccur[consIdx] += 1
```




```
if (consonant (consIdx))
   { nConsOccur[consIdx] += 1
     if (!detMultCons && (nConsOccur[consIdx] == 2))
        { nWordsWMultCons += 1
          detMultCons = true
        }
   }
```

Text processing in Portuguese - 4

Program validation

- `text0.txt`
- `text1.txt`
- `text2.txt`
- `text3.txt`
- `text4.txt`

Text processing in Portuguese - 5

Getting the contents of a text file in hexadecimal

```
[ruib@ruib-laptop countWords]$ od -A x -t x1 text0.txt
000000 63 c3 87 e2 80 93 61 41 3b 20 c3 a1 c3 81 c3 a0
000010 2c 20 c3 80 c3 a2 c3 82 27 c3 a3 c3 83 3f 21 0a
000020 65 45 2d c3 a9 c3 89 c3 a8 c3 88 2e 2e 0a 0a 20
000030 20 20 20 c3 aa c3 8a 2e 20 69 49 c3 ad 20 54 79
000040 0a 0a 0a c3 8d c3 ac c3 8c 2e 59 2e 2e 0a 5f 6f
000050 4f c3 b3 c3 93 c3 b2 c3 92 6d 21 61 62 42 e2 80
000060 a6 64 41 c3 b3 20 20 20 20 20 20 20 20 20 20 20
000070 20 20 20 c3 b5 c3 95 c3 b4 c3 94 20 20 20 2e 0a
000080 20 20 20 20 20 20 20 20 20 20 20 20 20 20 75 55
000090 c3 ba c3 9a 3f 20 74 c3 b9 c3 99 68 2e 0a
00009e
[ruib@ruib-laptop countWords]$
```

Text processing in Portuguese - 6

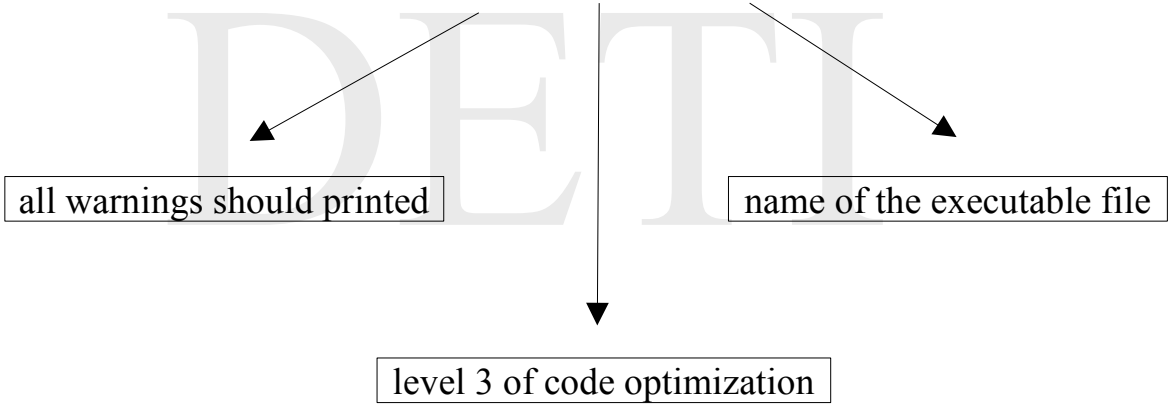
Contents of `text0.txt` (graphic representation + UTF-8 encoding)

```
c  Ç    —       a  A  ;  sp á    Á    à    ,  sp À    â    Â    `  ã    Ã    ?  !  nl
63 c387 e28093 61 41 3b 20 c3a1 c381 c3a0 2c 20 c380 c3a2 c382 27 c3a3 c383 3f 21 0a
e  E  -  é    É    è    È    .  .  nl
65 45 2d c3a9 c389 c3a8 c388 2e 2e 0a
nl
0a
sp sp sp sp ê    Ê    .  sp I  I  í    sp T  y  nl
20 20 20 20 c3aa c38a 2e 20 69 49 c3ad 20 54 79 0a
nl
0a
nl
0a
Í    ì    Ì    .  Y  .  .  nl
c38d c3ac c38c 2e 59 2e 2e 0a
_  o  O  ó    Ó    ò    Ò    m  !  a  b  B  …       d  A  ó    sp sp sp sp sp sp sp sp
5f 6f 4f c3b3 c393 c3b2 c392 6d 21 61 62 42 e280a6 64 41 c3b3 20 20 20 20 20 20 20 20
sp sp sp sp sp sp õ    Õ    ô    Ô    sp sp sp .  nl
20 20 20 20 20 20 c3b5 c395 c3b4 c394 20 20 20 2e 0a
sp sp sp sp sp sp sp sp sp sp sp sp sp sp u  U  ú    Ú    ?     t  ù    Ù    h  .  nl
20 20 20 20 20 20 20 20 20 20 20 20 20 20 75 55 c3ba c39a 3f 20 74 c3b9 c399 68 2e 0a
```

Text processing in Portuguese - 7

Compilation

```
[ruib@ruib-laptop countWords]$ gcc -Wall -O3 -o countWords countWords.c
```



Text processing in Portuguese - 8

```
[ruib@ruib-laptop countWords]$ ./countWords text0.txt text1.txt text2.txt text3.txt text4.txt

File name: text0.txt
Total number of words = 17
Total number of words with at least two instances of the same consonant = 2

File name: text1.txt
Total number of words = 1184
Total number of words with at least two instances of the same consonant = 207

File name: text2.txt
Total number of words = 11027
Total number of words with at least two instances of the same consonant = 1999

File name: text3.txt
Total number of words = 3369
Total number of words with at least two instances of the same consonant = 508

File name: text4.txt
Total number of words = 9914
Total number of words with at least two instances of the same consonant = 1322

[ruib@ruib-laptop countWords]$
```