Optimal alignment and plagiarism detection software

Code Execution Manual

Diego REYNA REYES & Ryan VAN GREUNEN

January 15th, 2023

Delivered files

The delivered .zip file contains the following files:

./TDm2_REYNAREYES_VANGREUNEN/documents/report.pdf: The report containing the answers to questions 1, 2 and 4 of the instruction website.

./TDm2_REYNAREYES_VANGREUNEN/documents/manual.pdf: This manual explaining how to run and test our code for questions 3 and 4.

./TDm2_REYNAREYES_VANGREUNEN/code/q3/main_q3.c: Source code for our implementation of question 3.

./TDm2_REYNAREYES_VANGREUNEN/code/q3/TD2_q3.c: A modified version of the TD2.c file shared on the instructions.

./TDm2_REYNAREYES_VANGREUNEN/code/q3/texte1.txt: The .txt file containing texte1 as shared with the instructions.

./TDm2_REYNAREYES_VANGREUNEN/code/q3/texte2.txt: The .txt file containing texte2 as shared with the instructions.

./TDm2_REYNAREYES_VANGREUNEN/code/q4/main_q4.c: Source code for our implementation of question 4.

./TDm2_REYNAREYES_VANGREUNEN/code/q4/TD2_q4.c: A modified version of the TD2.c file shared on the instructions. It's important to note that this file its slightly different to the one found in ./TDm2_REYNAREYES_VANGREUNEN/code/q3/TD2_q3.c.

./TDm2_REYNAREYES_VANGREUNEN/code/q4/t1.txt: The .txt file containing t1 as shared with the instructions.

./TDm2_REYNAREYES_VANGREUNEN/code/q4/t2.txt: The .txt file containing t2 as shared with the instructions.

Compiler and directory

Compiler: This code was tested and compiled using the following compilers:

- gcc (Ubuntu 7.5.0-3ubuntu1-18.04) 7.5.0

```
gcc (Ubuntu 7.5.0-3ubuntu1~18.04) 7.5.0
Copyright (C) 2017 Free Software Foundation, Inc.
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
```

- gcc-10 (Ubuntu 10.3.0-1ubuntu1-18.04-1) 10.3.0

```
gcc-10 (Ubuntu 10.3.0-1ubuntu1~18.04~1) 10.3.0
Copyright (C) 2020 Free Software Foundation, Inc.
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
```

  Given this is the most recent version of the used compiler that works on Linux, this version will be used in all the following instructions.

- gcc (Rev4, Built by MSYS2 project) 12.2.0

```
gcc (Rev4, Built by MSYS2 project) 12.2.0
Copyright (C) 2022 Free Software Foundation, Inc.
This is free software; see the source for copying conditions.  There is NO
warranty; not even for MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
```

Directory: Expecting the provided .zip file was extracted on folder **./**

Compiling and testing question 3

Compiling command: To compile the code written for this question we use the following command:

```
gcc-10 ./TDm2_REYNAREYES_VANGREUNEN/code/q3/main_q3.c -o ./q3_exe
```

The structure of the command is:

gcc-10 path/to/main_q3.c -o path/to/output/q3_exe

Testing command: To test the code written for this question we use the following command:

```
./q3_exe ./TDm2_REYNAREYES_VANGREUNEN/code/q3/texte1.txt ./TDm2_REYNAREYES_VANGREUNEN/code/q3/texte2.txt
```

The structure of the command is:

path/to/q3_exe path/to/txt_file1 path/to/txt_file2

When ran on the provided texte1.txt and texte2.txt we expect the following output:

```
zuramaru@zuramaru-VirtualBox:~/Documents/INF4202B$ gcc-10 ./TDm2_REYNAREYES_VANGREUNEN/code/q3/main_q3.c -o ./q3_exe
zuramaru@zuramaru-VirtualBox:~/Documents/INF4202B$ ./q3_exe ./TDm2_REYNAREYES_VANGREUNEN/code/q3/texte1.txt ./TDm2_REYNAREYES_VANGREUNEN/code/q3/texte2.txt
Score of allignment: 577.
Size of text 1: 1520
Size of text 2: 1799
Similarity score: 82.62 %
|-----------------------------------------------------------------|-----------------------------------------------------------------|
| Source Wikipedia.        La        di          s          t a   | Source Wikipedia modifie par un etudiant du cours IT-4301E, trait |
|    nc            e de Leve       ns      ht   e       i  n     | ement algorithmique de l information. La distance de d editon est |
|  une distance         mathematique donnant une mesure de la simil | une distance au sens mathematique donnant une mesure de la simil |
| arite entre deux chaines de caracteres. Elle est egale au nombre  | arite entre deux s    equ enc     e  s. Elle est egale au nombre  |
| minimal de caracteres qu il faut supprimer, inserer ou remplac er | minimal de caracteres qu il faut supprimer, inserer ou substituer |
|  pour passer d une chai n e a l autre. Elle a ete proposee par Vl |  pour passer d une sequence a l autre. Elle a ete proposee par Vl |
| adimir Levenshtein en 1965. Elle est       egalement connue sous  | adimir Levenshtein en 1965. Elle est ainsi egalement connue sous  |
| les noms de distance d   edi   t ion ou de d      e      formatio | les noms de distance de Levenshtei n ou de distance de deformatio |
| n dynamique temporelle, notamment  en reconnaissance de formes et | n dynamique temporelle  d  a   ns le  d  om ai   n e de l         |
|  particulierement en reconnaissance vocale1,2. Cette distance est |   a                reconnaissance de formes. Cette distance est |
|  d  autan  t plus grande que le nombre de differences entre les d |  est une fonction croissante du nombre de differences entre les d |
| eux chaines est grand. La distance de Levenshtei n peut etre cons | eux sequ ences       . La distance d   edi   t ion peut etre cons |
| ideree comme une generalisation de la distance de Hamming         | ideree comme une generalisation de la distance de Hamming (donnee |
|                                 . On peut montrer en particulier q |  par le nombre de position en lesquelles les deux sequences posse |
| ue la distance de Hamming est un majorant de la distance de Leven | dent des caracteres differents). On peut montrer en particulier q |
| shtei n.Definition          : on appelle distance de Levenshtei n |   t ion.Definition formelle : on appelle distance d   edi   t ion |
|  entre deux mots M et P le cout minimal pour alle r de  M a  P en |  entre deux mots M et P le cout minimal t  r ansform er M en P en |
|  effectuant les operations elementaires      s            uivant | effectuant les operations elementaires, dites d edition, suivant  |
| es : i) substitution d un caractere de M en  un caractere de P ;  | es : i) substitution d un caractere de M par un caractere de P ;  |
| ii) ajou t    dans M d un caractere de P ; iii) suppression    d | ii) insertion dans M d un caractere de P ; iii) suppression (ou d |
|          un caractere de M. On associe ainsi a chacune de ces o | eletion) d un caractere de M. On associe ainsi a chacune de ces o |
| perations un cout. Le c ou  t     est toujours egal a 1, sauf dan | perations un cout. On choisit souvent un cout  egal a 1  pour tou |
|   s le  c   a    s d un  e    substitution de caracteres identiqu | tes les operations excepte la substitution de caracteres identiqu |
| es.               Exemples : si M = "examen" et P = "examen",   | es qui a un cout nul.Exemples : si M = "examen" et P = "examen",  |
| alors LD (M, P) = 0, parce qu aucune operation n a ete realisee.  | alors Lev(M, P) = 0  parce qu aucune operation n a ete realisee.  |
| Si M = "examen" et P = "examan", alors LD (M, P) = 1, parce qu il | Si M = "examen" et P = "examan", alors Lev(M, P) = 1, parce qu il |
|  y a eu un  remplaceme nt (changement du e en a), et que l on ne  |  y a eu une substitution  (changement du e en a), et que l on ne  |
| peut pas en faire           m   o                  ins            | peut pas en faire une transformation de M en P avec un moindre co |
|                                                                   | ut                                                              |
|-----------------------------------------------------------------|-----------------------------------------------------------------|
```

Compiling and testing question 4

Compiling command: To compile the code written for this question we use the following command:

```
gcc-10 ./TDm2_REYNAREYES_VANGREUNEN/code/q4/main_q4.c -o ./q4_exe
```

The structure of the command is:

gcc-10 path/to/main_q4.c -o path/to/output/q4_exe

Testing command: To test the code written for this question we use the following command:

```
./q4_exe ./TDm2_REYNAREYES_VANGREUNEN/code/q4/t1.txt ./TDm2_REYNAREYES_VANGREUNEN/code/q4/t2.txt
```

The structure of the command is:

path/to/q4_exe path/to/txt_file1 path/to/txt_file2

When ran on the provided t1.txt and t2.txt we expect the following output:

```
zuramaru@zuramaru-VirtualBox:~/Documents/INF4202B$ gcc-10 ./TDm2_REYNAREYES_VANGREUNEN/code/q4/main_q4.c -o ./q4_exe
zuramaru@zuramaru-VirtualBox:~/Documents/INF4202B$ ./q4_exe ./TDm2_REYNAREYES_VANGREUNEN/code/q4/t1.txt ./TDm2_REYNAREYES_VANGREUNEN/code/q4/t2.txt
Score of allignment: 789.
Size of text 1: 1525
Size of text 2: 2020
Similarity score: 77.74 %
|-----------------------------------------------------------------|-----------------------------------------------------------------|
| Source Wikipedia                                                | Source Wikipedia modifie par un etudiant du cours IT-4301E, trait |
|                                .                                | ement algorithmique de l information.                           |
|-----------------------------------------------------------------|-----------------------------------------------------------------|
| La distance de L evenshtein une distance         mathematique don | La distance de d editon est une distance au sens mathematique don |
| nant une mesure de la similarite entre deux chaines de caracteres | nant une mesure de la similarite entre deux s    equ enc    e  s |
| . Elle est egale au nombre minimal de caracteres qu il faut suppr | . Elle est egale au nombre minimal de caracteres qu il faut suppr |
| imer, inserer ou remplac er pour passer d une chai n e a l autre. | imer, inserer ou substituer pour passer d une sequence a l autre. |
|  Elle a ete proposee par Vladimir Levenshtein en 1965. Elle est | Elle a ete proposee par Vladimir Levenshtein en 1965. Elle est a |
|     egalement connue sous les noms de distance d   edi   t ion o | insi egalement connue sous les noms de distance de Levenshtei n o |
| u de d      e      formation dynamique temporelle, notamment  en | u de distance de deformation dynamique temporelle  d  a   ns le  |
| reconnaissance de formes et particulierement en reconnaissance vo | d  om ai   n e de l         a              reconnaissance de |
| cale1,2. Cette distance est d  autan  t plus grande que le nombre | formes. Cette distance est une fonction croissante du nombre    |
|  de differences entre les deux chaines est grand. La distance de | de differences entre les deux sequ ences       . La distance d  |
| Levenshtei n peut etre consideree comme une generalisation de la | edi   t ion peut etre consideree comme une generalisation de la |
| distance de Hamming                                             | distance de Hamming (donnee par le nombre de position en lesquell |
|                                                  . On pe        | es les deux sequences possedent des caracteres differents). On pe |
| ut montrer en particulier que la distance de Hamming est un major | ut montrer en particulier que la distance de Hamming est un major |
| ant de la distance de Levenshtei n.                             | ant de la distance d   edi   t ion.                             |
|-----------------------------------------------------------------|-----------------------------------------------------------------|
|                                                                 | Et pourquoi ne pas raconter des ballivernes entre temps pour dete |
|                                                                 | cter les lecteurs attentifs.                                    |
|-----------------------------------------------------------------|-----------------------------------------------------------------|
|                                                                 | Et inserer un paragraphe qui n'a rien a voir avec le chmiliblic p |
|                                                                 | our tromper le chaland !                                        |
|-----------------------------------------------------------------|-----------------------------------------------------------------|
| Definition         : on appelle distance de Levenshtei n entre d | Definition formelle : on appelle distance d   edi   t ion entre d |
| eux mots M et P le cout minimal pour alle r de  M a  P en effectu | eux mots M et P le cout minimal t  r ansform er M en P en effectu |
| ant les operations elementaires      s           uivantes : i)  | ant les operations elementaires, dites d edition, suivantes : i)  |
| substitution d un caractere de M en  un caractere de P ; ii) ajou | substitution d un caractere de M par un caractere de P ; ii) inse |
|  t    dans M d un caractere de P ; iii) suppression     d        | rtion dans M d un caractere de P ; iii) suppression (ou deletion) |
|    un caractere de M. On associe ainsi a chacune de ces operation | d un caractere de M. On associe ainsi a chacune de ces operation |
| s un cout. Le c ou  t     est toujours egal a 1, sauf dan  s le  | s un cout. On choisit souvent un cout  egal a 1  pour toutes les  |
| c   a    s d un  e    substitution de caracteres identiques     | operations excepte la substitution de caracteres identiques qui a |
|          .                                                      |  un cout nul.                                                   |
|-----------------------------------------------------------------|-----------------------------------------------------------------|
| Exemples : si M = "examen" et P = "examen", alors LD (M, P) = 0, | Exemples : si M = "examen" et P = "examen", alors Lev(M, P) = 0  |
| parce qu aucune operation n a ete realisee. Si M = "examen" et P | parce qu aucune operation n a ete realisee. Si M = "examen" et P |
| = "examan", alors LD (M, P) = 1, parce qu il y a eu un  remplacem | = "examan", alors Lev(M, P) = 1, parce qu il y a eu une substitut |
| e nt (changement du e en a), et que l on ne peut pas en faire   | ion  (changement du e en a), et que l on ne peut pas en faire une |
|        m   o                  ins       .                     |  transformation de M en P avec un moindre cout.                 |
|-----------------------------------------------------------------|-----------------------------------------------------------------|
|                                                                 | Pas super complet ce cours...                                   |
|-----------------------------------------------------------------|-----------------------------------------------------------------|
```