Documentation du programme homonym.pl

Éric Guichard, Enssib / ENS / CNRS (Dante)

14 août mars 2017

Ce script écrit en Perl est construit pour fonctionner de façon autonome et pour être sollicité par des programmes écrits en d'autres langages (python, etc.).

1 Présentation

Il est historiquement conçu pour repérer des noms propres orthographiés de façon variable : Brezinski et Brzezinski, AE. . .et AE. . .ou Ausserbach et Außerbach, par exemple.

Il intègre aussi les variations dans les diacritiques (Frederic et Frédéric, Nuñès et Nunès), les ajouts d'une lettre ou d'un signe (indentique au lieu de identique, Raphaèle ou Raphaëlle, Jean Pierre et Jean-Pierre), les interversions de deux lettres (inspiration et inpsiration), les erreurs de frappe clavier (arome au lieu d'atome, r et t étant proches sur le clavier azerty), et de façon générale les variations d'une seule lettre dans des mots de longueur égale (Chaize et chaise, comme atome et arome).

Il ne repère donc que les *très lègères variations* au sein de mots préalablement sélectionnés (par exemple des noms propres de diverses langues occidentales).

2 Méthode

Dans tous les cas, les formes initiales des mots sont conservées.

La routine `suppressiondiacritikandco` les réduit à un état minimal type ASCII (ß → ss, é → e, Æ → AE). Elle supprime aussi espaces et insécables, simplifie tous les tirets. Il est aisé de la compléter au besoin. Ensuite, tous les mots sont basculés en minuscules.

La démarche s'effectue en trois temps.

1. Repérage des mots avec même graphie sans diacritiques ni accents. On lance alors une *alerte*.
2. Repérage des mots *de même longueur* pour rechercher d'éventuelles inversions de deux lettres (conpsiration au lieu de conspiration). Est alors sollicitée la routine `debutdefragmentcommunadeuxmots`, qui recherche le début commun à deux mots. La routine `motmisalenvers` fait ce qu'elle dit, elle sera utile pour trouver les fins communes à deux mots en utilisant la routine précédente. On pourra alors aisément repérer le début commun à deux mots (éventuellement vide), leur fin commune (évt. vide), et les milieux propres à chaque mot.

   Voici le détail de la démarche, repérée dans le script par les commentaires notés #CAS 2a, #CAS 2b, etc.

   — Si au moins 3 lettres consécutives distinguent les milieux de mots, on suppose qu'ils sont vraiment différents. Pas d'alerte.

— Si les deux mots sans diacritiques sont identiques à une inversion près d'un couple de lettres, on lance une *alerte*.
— Si les couples de lettres sont distincts, même après inversion, on suppose encore que nos mots sont différents.
— Si une seule lettre distingue les deux mots, on lance malgré tout une *alerte*.

3. Ajout par erreur ou ambiguïté d'une lettre (ex. : indentité au lieu de identité, Tzara ou Zara). Ce point est assez délicat. Le piège est que le début et la fin communs aux deux mots ne redonnent pas toujours le plus petit mot. Par exemple, *ecole* et *ecolle* ont pour début commun *ecol* et pour fin commune *le*. En bref, on ne peut reconstituer les mots avec leur début, leurs milieux respectifs et leur fin.

   Une *alerte* est lancée si les deux mots sont proches à une lettre/signe près. C'est la seule partie du script où l'utilisat/eur/rice est invité/e à retrouver le signe supplémentaire.

3 Réponses produites par le script

Dans la grande majorité des cas, l'alerte signale en majuscules les signes qui posent problème.

3.1 Premiers exemples d'alertes

— Réponse à Außerbach Aussebrach : *Außerbach et Aussebrach sont identiques à une inversion près : ausseRBach*
— Réponse à identité indentité : *identité et indentité se distinguent à un signe près : iNdentite*

3.2 Autres exemples d'alertes (cas 1 et 2)

Dans ces cas, la différence est mise en évidence par des majuscules.
— École et ecolé ont les mêmes formes simplifiées
— École et Eocle sont identiques à une inversion près : eCOle
— Une lettre distingue ecolTe et ecolLe
— Eocle et Eole se distinguent à un signe près : eoCle
— Jean-Pierre et jean Pierre se distinguent à un signe près : jean-pierre [1]

3.3 Exemples d'alertes du cas 3

Ici, les différences ne sont pas visualisées.
— Raphaëlle et Raphaèle se ressemblent à une lettre doublée près, au milieu du grand mot et absente du plus petit mot.
— École et Ecoles se ressemblent à la lettre terminale près, absente du plus petit mot.
— École et recole se ressemblent à l'initiale près, absente du plus petit mot.
— Nunèçö et NUnec se ressemblent à la lettre terminale près, absente du plus petit mot.

1. Cas rare où le tiret n'est pas très visible, car il n'existe pas sous forme majuscule.

4 Usage

4.1 Documentation succincte

perl homonym.pl -help lance cette documentation. Pour info, le début du code explique en détail le fonctionnement de ce script.

perl homonym.pl Idem

perl homonym.pl -f fichier (avec chemin d'accès si le fichier n'est pas dans le même dossier que ce programme) : signale, pour tous les mots du fichier (séparés par un passage à la ligne ou RC), les très fortes ressemblances. Le résultat apparaît dans la fenêtre du terminal.

perl homonym.pl -g fichier idem, mais produit le résultat dans un fichier de même nom que l'initial, complété du suffixe .homonym (non garanti si le fichier d'origine contient des espaces ou des accents ni dans des univers Windows).

perl homonym.pl mot1 mot2 mot3 (etc.) signale, pour tous les mots de la liste (2 à 2), les très fortes ressemblances (erreurs typo, variations d'accents, de casse, interversion de 2 lettres, etc.).

NOTE : si le fichier est exécutable, on peut omettre la commande perl du début : `./homonym.pl` etc.

4.2 Exemples d'usage et fichiers associés

— `perl homonym.pl cher chère`
Réponse : *cher et chère se ressemblent à la lettre terminale près, absente du plus petit mot.*

— Le fichier `mots.txt` contient les mots suivants, chacun sur une ligne : *Brezinski chat chaise Chèze chaize chate cher chèr Brzezinski Außerbach Aussebrach Au-Berbach identité indentité Raphaëlle Raphaèle École ecolé Ecoles recole ecolte écolage Eocle Eole Nunèçö Ecolle NUnec Ecopli Jean-Pierre jean Pierre*
`perl homonym.pl -g mots.txt` fabrique le fichier `mots.txt.homonym` qui commence ainsi :
Brezinski et Brzezinski se distinguent à un signe près : brZezinski
chat et chate se ressemblent à la lettre terminale près, absente du plus petit mot
Une lettre distingue chaiSe et chaiZe

Enjoy !
Eric.Guichard@enssib.fr