

Gestion d'un annuaire

Introduction

On dispose d'un annuaire recensant la liste des employés d'une entreprise.

Dans cette entreprise, le standard est géré de façon automatique par un logiciel. Il suffit que le correspondant énonce le nom d'une personne pour que le logiciel reconnaisse le nom de cette personne.

Un certain nombre d'employés ont cependant des noms qui se prononcent de la même façon.

L'objectif de cet exercice est de fournir la liste des noms des employés dont la prononciation correspond à celle énoncée par le correspondant.

Algorithme phonétique d'indexation

Dans le traitement du langage naturel et plus particulièrement du traitement de l'oral, un algorithme phonétique d'indexation des mots est un algorithme qui indexe les mots par leur prononciation. L'objectif basique est que les noms ayant la même prononciation soient codés avec la même chaîne de manière à pouvoir trouver une correspondance entre eux malgré des différences mineures d'écriture.

Dans la suite de cet exercice, nous mettrons en œuvre l'algorithme du *Soundex* qui est largement utilisé dans le monde numérique. Il repose sur un principe de base qui consiste à codifier le mot en éliminant les lettres en double, les lettres muettes et en rapprochant les sons de certaines lettres.

Exemples : "Jurafsky", "Jarofsky", "Jarovsky" et "Jarovski" seront tous représentés par exemple par J612 pour un codage correspondant à l'anglais¹
"Robert" et "Rupert" sont représentés par : "R163", tandis que "Rubin" donne R150²
"Baragwanath" donne B625
"Woolcock" donne W420

De tels algorithmes peuvent être utilisés par des moteurs de recherche, des correcteurs orthographiques. L'algorithme du *Soundex* est également utilisé pour faire des recherches selon l'épellation phonétique des noms dans les bases du recensement fédéral des Etats-Unis.

Travail demandé

Connaissant :

- les noms des employés de l'entreprise

écrire un programme permettant :

- de saisir un nom
- et d'afficher la liste des employés de l'entreprise dont les noms se prononcent de la même façon que le nom saisi

L'algorithme phonétique d'indexation mis en œuvre dans le programme est une version simplifiée de l'algorithme du *Soundex*, il est décrit ci-après.

Il est **clair que** ce programme doit faire appel à des fonctions intermédiaires permettant de mettre en œuvre les différentes étapes de cet algorithme.

¹ <http://www.loria.fr/~perrier/teaching/progTAL/fonctions.pdf>

² <http://fr.wikipedia.org/wiki/Soundex>

L'algorithme du *Soundex* – version simplifiée

Il existe de nombreuses variantes de l'algorithme de calcul du code *Soundex*, nous vous en proposons une version simplifiée³ :

1. Mise en forme
 - Supprimer les éventuels espaces initiaux à gauche du mot
 - Mettre le mot en majuscules
2. Enlever les lettres muettes
 - Conserver la première lettre de la chaîne
 - Supprimer toutes les occurrences des lettres : A, E, H, I, O, U, W, Y (sauf si c'est la première lettre du nom)
3. Coder les sons proches

Attribuer une valeur numérique aux lettres sauf la première de la manière suivante :

Version pour l'anglais :	Version pour le français :
1 = B, F, P, V	1 = B, P
2 = C, G, J, K, Q, S, X, Z	2 = C, K, Q
3 = D, T	3 = D, T
4 = L	4 = L
5 = M, N	5 = M, N
6 = R	6 = R
	7 = G, J
	8 = X, Z, S
	9 = F, V

4. Enlever les doublons
 - Si un chiffre est répété plusieurs fois consécutivement, n'en garder qu'une occurrence.
5. Ramener à un code de longueur 4
 - Ne garder que les quatre premiers caractères, si le code généré fait moins de 4 caractères, le compléter par des 0.

³ On trouvera une version complètement francisée de cet algorithme aux adresses :

<http://sqlpro.developpez.com/cours/soundex/>
<https://info.univ-lemans.fr/~carlier/recherche/soundex.html>