

Detector d' Idiomes



Introducció

En aquest document es descriu una aplicació per a la detecció d'idiomes. L'aplicació es divideix en dues parts; per una banda tenim la llibreria que exporta un objecte amb operacions per a la detecció de l'idioma d'un text donat (apartat 1) i per l'altra, una eina de configuració del detector de l'idioma (apartat 2) que permet afegir o eliminar idiomes a la llista d'idiomes detectats per l'aplicació. L'únic requisit per afegir un nou idioma és el de disposar d'un corpus per aquest idioma en format de text pla.

L'eina se subministra amb dades per al reconeixement dels següents idiomes:

- castellà
- català
- anglès
- euskera

Instal·lació per a Windows®

El detector d'idiomes disposa d'un programa d'instal·lació estàndard que instal·la en el sistema les llibreries dinàmiques necessàries per a la seva utilització, així com els fitxers de dades necessaris per als idiomes inclosos. Inclou també:

- una eina de configuració per poder afegir o eliminar idiomes de la llista d'idiomes suportats.
- un fitxer amb la API del sistema, per poder incloure'l en els programes que l'utilitzin.

La API exporta la interfície d'un objecte que permet detectar l'idioma de tots els textos que se li vagin passant, una vegada carregades inicialment les dades dels idiomes disponibles. Aquest objecte està declarat en C++ i només permet el seu ús des de llocs on es pugui cridar C++. Si la crida a C++ pot significar un problema per integrar-se en un sistema donat, es pot afegir fàcilment a la API una funció de C estàndard que pugui realitzar tot el procés, és a dir, la iniciació i la detecció sense necessitat d'utilitzar l'objecte C++, amb el que podria ésser utilitzat per qualsevol sistema que pot cridar a una DLL estàndard de Windows.

1. Detector d'idioma

Funcionament

El detector d'idiomes està escrit en C++ i ofereix al programador un objecte CDetector que ofereix operacions per detectar l'idioma més proper del text donat a la llista d'idiomes suportats per a l'aplicació. És important tenir en compte que l'operació sempre ens indicarà l'idioma més proper d'entre els que disposa l'aplicació, però si l'idioma real del text no es troba entre ells, retornarà el més similar.

- ***Exemple de funcionament***

El funcionament del CDetector és molt senzill. Una vegada declarada la variable, s'ha de cridar a la funció d'iniciació() perquè l'objecte llegeixi el registre de Windows, la localització dels fitxers de dades i acte seguit procedeixi a carregar-los. A partir d'aquest moment ja es pot cridar a l'operació de detectar_idioma passant-li el text que es desitja avaluar. Aquesta operació retornarà per al segon paràmetre l'identificador (o abreviatura) de l'idioma més proper d'entre els existents per aquest text.

5

Breu descripció de la interfície de l'objecte CDetector

La interfície de l'objecte està definida de la següent manera:

```
class DETECTOR_API CDetector {
protected:
    // Atributos
    void *p_tri;
    std::string    prefix;

public:
    // Métodos
    CDetector();
    ~CDetector();

    int inicializar();

    int detectar_idioma( const char *texto, char
    *id);
    int detectar_idioma( std::string &texto,
    char *id);
};
```

6

A banda del constructor i destructor típics té un mètode per carregar les dades dels idiomes disponibles, **iniciació()**, que no té arguments d'entrada i que retorna un 0 si tot va correctament i una altra cosa si es produeix un error durant la càrrega.

Per la seva funció principal, la detecció de l'idioma, disposa d'un mètode anomenat **detectar_idioma**, que consta dels següents paràmetres d'entrada:

- text: text que es desitja avaluar; pot venir en dos formats, com char* o como objecte string de la llibreria estàndard. El funcionament és el mateix en ambdós casos.
- id: Codi de l'idioma detectat com el més probable durant el procés. Aquest codi ha d'ésser una de les abreviatures utilitzades per referir-se a un dels idiomes suportats pel detector.

La funció retorna 0 si tot funciona correctament i qualsevol altra cosa en cas de produir-se un error durant la seva execució.

Exemple de crida en C

El següent programa en C és un exemple que utilitza el detector d'idioma per a la detecció de l'idioma del fitxer passat com paràmetre.

```
#include "detector.h"
#include <iostream>
#include <fstream>
#include <string>
#include <cstdlib>

int main ( int argc, char *argv[] ) {
    CDetector detector;

    // Comprobar argumentos
    if ( argc != 2 ) {
        std::cerr << "\n Uso: " << argv[0] << "
<fitxer>\n\n";
        return 1;
    }
}
```

```

    if ( detector.inicializar() < 0 ) {
        std::cerr << "ERROR: No puedo inicializar el
detector" << std::endl;
        return 2;
    }

    std::ifstream fich(argv[1]);
    char buf[1025];
    std::string texto;
    while (!fich.eof()) {
        fich.read(buf, 1024);
        texto += buf;
    }

    char id[10];
    detector.detectar_idioma(texto,id);
    std::cout << "Idioma detectado: " << id <<
'\n';

    return 0;
}

```

9

Eina de configuració per al detector d'idiomes (ConfigDetector)

Aquesta eina s'encarrega de llegir els fitxers de dades dels idiomes disponibles, mostrant una llista amb tots els trobats. A partir d'aquest moment permet eliminar idiomes existents o afegir-ne de nous.

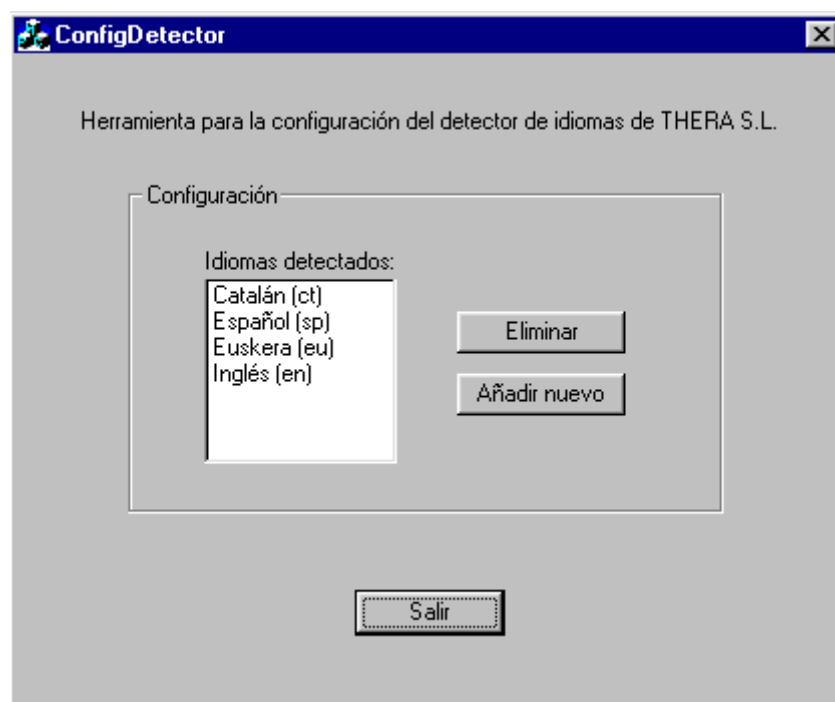
Es pot accedir a aquesta eina des del menú d'inici de Windows fent:

```

Inicio>>Programas>>Thera>>Detector
Idiomas>>ConfigDetector

```

10



11

Afegir nou

Aquesta operació ens permet afegir un nou idioma per incorporar-lo al nostre detector. Se li ha d'especificar el nom complet de l'idioma, la abreviatura (màx. 3 lletres) associada i el fitxer que contingui el corpus de l'idioma introduït. L'abreviatura no pot coincidir amb cap de les existents ja que aquesta abreviatura serà utilitzada pel detector d'idiomes per donar el resultat. Sobre el fitxer de càrrega, el format d'aquest haurà d'ésser format de text amb codificació de 1 byte (Windows, ISO-8859, ASCII, etc)

NOTA: De moment no se suporta la codificació utf8 ni cap altre que permeti unicode; malgrat això els resultats obtinguts serien bastant correctes. L'únic que passa és que en el moment de calcular les probabilitats, no es té en compte les codificacions especials de caràcters que ocupen més d'un byte, i per tant, l'estadística pot quedar una mica desvirtuada. En un futur proper es donarà suport a aquest format.

Eliminar

Elimina un idioma d'entre els disponibles pel detector. El fitxer de dades utilitzat per aquest també serà eliminat, així com tota referència a aquest idioma.

12

Thera, Centre de Llenguatges i Computació, S.L.

Tel 93 403 45 58 – Fax 93 403
C/ Adolf Florensa s/n Ed. Florensa
08028-BARCELONA

