Bisoños Usuarios de GNU/Linux de Mallorca y Alrededores | Bergantells Usuaris de GNU/Linux de Mallorca i Afegitons

Traduccions i screen scraping (9204 lectures)

Per **Antoni Aloy López**, ***aaloy*** (http://trespams.com)

Creado el 14/10/2007 13:01 modificado el 06/11/2007 18:39

Amb un grapat de línees de codi python podem tenir un sistema de traducció per les nostres aplicacions sempre que tinguin connexió a Internet.

Actualització

En Juan Hevilla ens ha enviat un petit programa[1] fet amb wxPython que permet fer traduccions de l'anglès al castellà del que tinguem al portapapers (es permeten modificacions) que aprofita el codi d'aquest exemple. És a més un bon exemple de com es poden fer aplicacions amb WxPython. Gràcies Juan!
Traducción: Internostrum[2]

Actualización

Juan Hevilla nos ha enviado una pequeña utilidad[1] realizada en wxPython que nos permite traducir el texto que tengamos en el portapapeles del inglés al castellano (se permiten modificaciones) que aprovecha el código de este ejemplo. Es además un buen ejemplo de cómo realizar apliaciones wxPython. ¡Gracias Juan!.

Google[3], Altavista[4], Gencat[5], o Internostrum[6], per exemple ens ofereixen serveis de traducció per la web, que ens poden anar molt bé per a creació d'aplicacions que necessitin tirar de traduccions ràpides com a ajuda a l'usuari.

Les traduccións del Gencat i Internostrum són especialment bones, encara que Gencat no és precisament ràpid.

La idea és accedir mitjançant un script a algun d'aquests serveis, enviar la traducció que volem fer i després recuperar el resultat de traducció per mostrar-ho a la nostra aplicació.

.

Per l'exemple utilitzaré Google, ja que és un dels més complexes, no tant per la quantitat de paràmetres que fa servir (Gencat per exemple en té més) sinó perquè permet veure com podem enviar també capçaleres http, ja que Google ens donarà un error de pàgin prohibida si no identificam el navegador.

Les llibreries de Python que ens serviran per accedir a la pàgina de traducció són *urllib* i *urllib2*. En pàgines que no necessiten mas cosa *urllib* ja és suficient, però si necessitam com en el cas de Google, enviar informació addicional o autentificar-nos, necessitarem la llibreria *urllib2*

El que feim és doncs crear un diccionari amb els paràmetres del POST que hem de passar, *params* en el nostre exemple, i transformar-lo en una cadena apta per la seva transmissió per Internet, això ho feim amb

```
data = urllib.urlencode(params)
```

Com que hen d'identificar el navegador, hem de passar un diccionari amb les capçaleres http que vulguem afegir

```
headers = { 'User-Agent' : user_agent }
```

per l'*user agent*, he anat a la utilitat d'identificació de Konqueror i he generat i copiat la identificació

```
 user_agent = 'Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.6 (like Gecko) (Kubuntu)'
```

Una vegada fet això ja sols ens queda fer la connexió i obtenir la pàgina de resultats

```
        request = urllib2.Request(url, data,headers)
        response = urllib2.urlopen(request)
        pagina = response.read()
```

Ens ens queda tractar la pàgina obtinguda i extreure'n la traducció que volem, per això utilitzarem una llibreria anomenada BeautifulSoup,[7], una petita meravella de codi, que ens permet tractar amb HTML i XML. En el nostre cas basta carregar el resulta obtingut amb la codificació pertinent i cercar el marcador (tag) que té la resposta

```
        soup = BeautifulSoup(''.join(pagina), fromEncoding='utf-8')
        traduccio =soup.find('div',{'id':'result_box'})
        return traduccio.renderContents().strip()
```

La resta de codi ja sols és un poc de gestió d'errors i poca cosa, més. Fixau-vos com podem generar fàcilment un arxiu temporal pe deixar la pàgina si les coses no van bé. Con se sol dir, Python ve amb les bateries incloses.

Esper que us sigui d'utilitat. Els altres serveis de traducció són molt semblants i es deixen com a exercici pel lector.

```
#!/usr/bin/env python
# -*- coding: UTF-8 -*-
# Autor: aaloy
# ----------------------------------------------------------------------
"""
    Emprant google com a traductor
"""
import urllib
import urllib2
from BeautifulSoup import BeautifulSoup
from tempfile import mkstemp
def traductor(text,direccio):
    url = "http://www.google.com/translate_t"
    params ={
    'hl':'en',
    "ie":"UTF8",
    'text':text,
    "langpair":direccio,
    }
    # Hem d'indidicar el user agent o ens donarà un error Forbidden
    user_agent = 'Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.6 (like Gecko) (Kubuntu)'
    data = urllib.urlencode(params)
    headers = { 'User-Agent' : user_agent }
    try:
        # Si anam per get
        print "%s?%s" % (url,data)
        # Però anirem per post
        request = urllib2.Request(url, data,headers)
        response = urllib2.urlopen(request)
        pagina = response.read()
        # I ara l'screeen scrapping
        soup = BeautifulSoup(''.join(pagina), fromEncoding='utf-8')
        traduccio =soup.find('div',{'id':'result_box'})
        return traduccio.renderContents().strip()
    except urllib2.HTTPError, msg:
        print "Error en la conexión, has puesto el user_agent?"
        print msg
    except Exception, e:
        # Si hi ha error deixam el contingut dins una pàgina per a revisar-ho
        print e
        id,name = mkstemp(prefix='google_tr_', suffix='.html')
        print "Error, s'ha escrit el resultat a %s " % name
        arx = open(name,'w')
        arx.write(soup.prettify())
        arx.close()
    return None

if __name__=="__main__":
    print traductor("una patata caliente servida fría",'es|en')
```

```
http://www.google.com/translate_t?langpair=es%7Cen&text=una+patata+caliente+servida+fr%C3%ADa&ie=UTF8&hl=
A buck served cold
```

Lista de enlaces de este artículo:

1. http://bulma.net/~aaloy/tt.tgz
2. http://www.internostrum.com/navegador.php?direccio=ca-es&url=http%3A%2F%2Fbulma.
3. http://www.google.com/translate_t
4. http://babelfish.altavista.com/
5. http://traductor.gencat.net/jsp/go2text.jsp?locale=ca
6. http://www.internostrum.com/
7. http://www.crummy.com/software/BeautifulSoup/documentation.html

E-mail del autor: aaloy _ARROBA_ bulma.net
Podrás encontrar este artículo e información adicional en: http://bulma.net/body.phtml?nIdNoticia=2403