tesseract (OCR)

- Objet : Installation et utilisation de Tesserac
- Niveau requis :
  débutant
- Commentaires : *Installer et utiliser tesseract pour la reconnaissance de caractères (OCR)*
- Débutant, à savoir : Utiliser GNU/Linux en ligne de commande, tout commence là !.
- Suivi :
  à-tester
    - Création par chalu 01/07/2017
    - Testé par <...> le <...> Fix Me!
- Commentaires sur le forum : Lien vers le forum concernant ce tuto [1)]

Introduction

Installer et utiliser le logiciel tesseract pour effectuer une reconnaissance de caractère à partir d'une image *.png ou d'un fichier *.pdf

Installation

```
apt-get install tesseract-ocr tesseract-ocr-fra tesseract-ocr-osd
```

Par défaut le pack de langue anglaise (eng) est installé et est utilisé lors de la reconnaissance des caractères.

On peut installer d'autres langues, par exemple l'espagnol si on veut exploiter un document dans cette langue, il suffit d'installer le paquet : tesseract-ocr-spa.

Pour installer toutes les langues installer le paquet **tesseract-ocr-all**

Interface graphique

Pour avoir une interface graphique en français, choisir OCRfeeder (en français) :

```
apt-get install ocrfeeder unpaper
```

ou gimagereader (en anglais)

Images

Pour manipuler les images, on installe imagemagick

```
apt-get install imagemagick
```

Scanner le document

Vous pouvez scanner votre document pour obtenir une image avec suffisamment de qualité en utilisant le logiciel de votre choix, par exemple SimpleScan. Le mieux est de choisir le format png.

L'important est de choisir une résolution assez élevée de 300 à 500 dpi, voire 600 dpi

Reconnaissance du texte d'un fichier PDF

Convertir

On commence par convertir le fichier *.pdf en image *.png :

```
convert -density 500 PDFtest.pdf -quality 100 test.png
```

Il y aura autant d'images en sortie que de pages du pdf nommé PDFtest.pdf.
Les noms de ces images seront test-0.png, test-1.png ...etc

Il peut y avoir des messages d'erreurs mais cela n'empêche la reconnaissances de caractères.

Reconnaissance de texte

Pour effectuer la reconnaissance de texte de la première image :

```
tesseract -l fra test-0.png output1
```

Ici la langue du document est spécifiée avec l'option -l fra.

Si rien n'est indiqué, c'est la langue anglaise qui est utilisée (eng)

Pour indiquer l'utilisation de deux langues par exemple français et allemand utilisez l'option : -l fra+deu.

Par défaut le fichier en sortie sera au format *.txt, on trouvera donc un fichier output1.txt à ouvrir avec n'importe quel éditeur de texte.

Pour effectuer la reconnaissance de texte de la deuxième image :

```
tesseract -l fra test-1.png output2
```

Reconnaissance du texte d'une image *png

La taille de l'image est un élément clé pour la reconnaissance des caractères

Voir l'exemple donné dans ce message d'un fil du forum où l'on voit bien l'influence de la taille de l'image sur la reconnaissance de caractères.

Automatisation avec des scripts

On peut ajouter des actions personnalisées dans thunar (gestionnaire de fichiers de XFCE) qui permettent d'avoir avec un clic droit sur le fichier une entrée de menu permettant de choisir une action à réaliser sur ce fichier.

Script sur png

Le script suivant (avec tous ses défauts, c'est mon premier script) permet d'effectuer la reconnaissance des caractères sur une image *.png et ouvre libreoffice (writer) pour lire ou modifier le texte.

```
#!/bin/bash
tesseract -l fra "$1" "${1%%.*}"
lowriter "${1%%.*}.txt"
exit 0
```

Il suffit de copier coller le texte avec mousepad (ou un autre éditeur de texte) et de l'enregistrer en lui donnant par exemple le nom PNG-2-ocr-lo. Ensuite un clic droit sur le fichier > Propriétés > Permissions pour le rendre exécutable en cochant la case ad-hoc.

Dans thunar > Editer > Configurer les actions personnalisées > clic sur le bouton pour ajouter une action.
On complète le nom de l'action par exemple PNG 2 OCR libreoffice et on remplit la ligne de commande avec :

```
/le-chemin-qui-mène-au-script/PNG-2-ocr-lo %f
```

On peut choisir une icone pour l'action.

Dans l'onglet “conditions d'apparition” on coche “Fichiers image” et on complète la ligne “motif de fichiers” avec :
.png;.PNG

On valide et voilà. On a l'action qui est proposée sur les fichiers *.png

Script roc

Exemple d'utilisation de tesseract en sélectionnant une partie de l'affichage à l'écran (page web, fichier image, etc....)

```
#!/bin/bash

## sélection d'une zone sur l'écran pour conversion ocr

##fichier de sortie
sortie=/chemin/vers/zone_ocr.txt

##choix langue  (fra eng etc...)
langue=fra

cd ~
import -quality 300 -depth 1000 ~/tmp.jpeg
tesseract  -l $langue ~/tmp.jpeg tmp  2> /dev/null && rm -f ~/tmp.jpeg
cat ~/tmp.txt >> $sortie && rm -f ~/tmp.txt
```

Utilisation

Le script lance un sélecteur de souris en forme de croix. Donc on sélectionne la zone que l'on veut passer à l'ocr et on retrouve (avec plus ou moins de fidélité) le texte dans le fichier indiqué dans la variable “sortie”. La qualité de copie obtenue peut être réglée par les valeurs de -quality et -depth. Voir le man import. A noter que le script est actuellement configuré pour ajouter, à la suite dans le même fichier, les sélections successives.

Source :

Phlinux

- r.o.c.-sur-une-partie-de-l-ecran

Sources

- https://github.com/tesseract-ocr/tesseract/wiki/Command-Line-Usage
- Tesseract-OCR sur le site Ubuntu-fr
- http://www.joyofdata.de/blog/a-guide-on-ocr-with-tesseract-3-03/

[1)]

N'hésitez pas à y faire part de vos remarques, succès, améliorations ou échecs !