Optical Character Recognition (OCR) using python

Semeh Ben Salem

semehbensalem0@gmail.com

Step1: Installing the tesseract package under Ubuntu using the following command

sudo apt install tesseract-ocr

tesseract –v pour vérifier l'installation

Step2: Using the tesseract command to extract the text.

tesseract test.png output

This is the first line of this text example.

This is the second line of the same text.

```
~/Desktop/OCR 11:07:19
$ cat output.txt
This is the first line of
this text example.

This is the second line
of the same text.
```

Chers amis, chère famille, chers tous,
En tant que gendre, je tiens à apporter
ma sympathie et à partager notre doulet
évoquant le souvenir d'un être plein de
vitalité, partageant avec tous la joie de v
l'humour sous toutes ses coutures

```
~/Desktop/OCR 11:24:50
$ cat output.txt
KIMJ W, diva/imam, MMJ loud,
[4 [ant m. ?lll.d4t., {aw i mfm
m J [Eat/hi «.1? 12¢ "DZ/dam;
m umé a ”mum dun em Fm aé
Vzfiz/ztl, [viva/u? am tau.) (,1 Jam 46. u
I'M/mum Jaw toufi; J1.) mum
```

G1 – La phrase - texte

Victor le maladroit

Le petit Victor joue au ballon dans le jardin. Sa sœur Marine mange une glace. Soudain, Victor tape dans le ballon. Marine le reçoit dans la figure. Sa glace est toute écrasée, Marine est toute barbouillée. Elle pleure. Son frère éclate de rire.

```
~/Desktop/OCR 11:23:05
$ cat output.txt
(31 , La phrase , texte

Victor le maladroit

Le petit Victorjoue au ballon dans Ie
jardin. Sa soeur Marine mange une glace.
Soudain, Victor tape dans Ie ballon.
Marine |e regoit dans la figure. Sa glace
est toute écrasée, Marine est toute

barbouillée. Elle pleure. Son frere éclate
de rire.
```

En fin de conte (diptyque)

Un jour alors que j'aidais grand-mère à faire son lit, elle me parla de tous les métiers qu'elle avait faits. Elle avait été maman, secrétaire trilingue, et princesse. Grand-mère me raconta aussi un secret. Le lit de grand-mère avait été construit à l'endroit même où il se trouve aujourd'hui.
Son rôle était d'empêcher les mauvaises créatures de sortir de ce très très vieux livre, en les gardant bloquées à l'intérieur. Mais cela ne suffisait pas il fallait que quelqu'un veille allongé dans le lit, sans jamais s'endormir sous peine de voir apparaître le plus terrible des monstres. Mais je savais que grand-mère me racontait des histoires et qu'elle n'était pas princesse, mais une sorte de surveillante, d'agent de sécurité anti-monstre. En plus une princesse n'a ni ride, ni cheveux gris. Mais peut-être aussi qu'à force de ne pas dormir les grandes personnes vieillissent plus vite.
Un jour grand-mère a fermé les yeux et la mort est sortie du livre.

```
$ cat output.txt
En fin de conxe (dipxyque)

Unjouralors quej'aidaisgrandrmérei faire son lix elle me
parla de xous les méxiers qu'elle avaix faits. Elle avaix éxé
maman, secrétaire xrilingue, ex princesse. Grandrmére me
raconta aussi un secret. Le lix de grandrmére avaix éxe
construit a l'endroit méme all N sexrouve aujould'hui.

Son réle exaix d'empécher les mauvaises créaxures de sortir
de ce xresxres vieux livre, en les gardanx bloquéesa
l'inxerieur. Maiscela ne suffisait pas i| fallait que quelqu'un
veille allonge dans le lix, sans jamaiss'endormir sous peine
de voir apparalxre le plus terrible des monsxres. Mais je
savais que grandrmére me raconxaix des hiswires ex qu'elle
n'éxaix pas princesse, ma is urle sorxe de surveillanxe, d'agenx
de sécurité antirmonstre. En plus une princesse n'a ni ride,
ni cheveux gris. Mais peuxexre aussi qu'a force de ne pas
dormir les grandes personnes vieillissenx plus vixe.

Un jour grandrmére a fermé les yeux ex la morx esx sorxie du
livre.
```

Results

•Identification correcte des textes standards présentant une typographie bien compréhensible.

•Problème pour l'identification du texte particulier en italique ou écrit à la main.

Pour installer un package avec un langage particulier

sudo apt-get install tesseract-ocr-eng

Utilisation sous python et ubuntu

- Installation du package pour les images en python
- **sudo apt-get install python-imaging**
- Installation de pip pour installer les packages pytesseract
- **sudo apt-get install python-setuptools**
- **sudo easy_install pip**
- **sudo pip install pytesseract**

```
>>> from PIL import Image
>>> import pytesseract
>>> txt=pytesseract.image_to_string(Image.open("test.png"),lang="eng")
>>> print(txt)
```