



# Tesseract, un OCR étonnant

*Le 20 janvier 2008, à 13:24 par Ulhume...*

Je suis tombé il y a peu sur [cet article](#) <sup>[1]</sup> ventant les bienfaits d'un "nouvel" OCR dans le paysage du libre. Après mes déboires répétés avec GOCR, j'ai donc bondi sur l'occasion.

## Origine du projet

Tesseract n'est pas tombé, loin de là, de la dernière pluie. Il s'agit en réalité d'un projet relativement ancien développé entre 1985 et 1994 par le groupe HP pour être finalement abandonné lors d'un recentrage des activités du groupe.

Ce n'est du coup qu'en 2005 que le projet reprend des couleurs lorsque certains employés décident de publier son code avec l'aide de Google et de l'Information Science Research Institute. Ces derniers apportant au vénérable outil debuggage et améliorations diverses.

Tesseract n'est devenu réellement libre qu'un peu plus tard, en Août 2006, une fois que les dernières parties propriétaires eurent été ôtées (réseau de neurone Aspirin/Migraine).

Et l'histoire ne s'arrête cependant pas là, car Tesseract, qui doit être considéré comme un "simple" moteur de reconnaissance de caractère multi-langues, est aujourd'hui en cours d'intégration dans un plus vaste projet nommé [OCRopus](#) <sup>[2]</sup>. Mené par Google, l'objectif est de donner naissance à une chaîne complète comprenant la numérisation, l'analyse de formatage (RAST), la reconnaissance de langue, la reconnaissance de caractères (Tesseract) et une correction du résultat (aspell).

Bref, de quoi changer complètement le positionnement des plate-formes libres dans le domaine l'OCR.

## Installation

C'est fou le nombre d'outils que l'on ne connaît pas qui sont déjà sagement en attente d'être installé sur nos disques (je mirrore les bases URPMI toutes les nuits). Tesseract n'échappe pas à la règle et son installation s'est simplement résumé à un :

```
|| # urpmi tesseract
```

## Premiers tests

Pour ce test, j'utilise un exemplaire de Newsweek, bien chiffonné par un mois de bourlingue, à l'inimitable papier glacé ultra-fin. C'est un peu le test critique pour moi car ce papier est à ce point transparent que à la lampe fluorescente du scanner fait apparaître en partie le texte qui se trouve derrière la page.

Première étape donc, un scan en 400DPI (ce qui semble être le bon compromis taille/reconnaissance):

```
$ scanimage --mode color --resolution300 -d
net:mon_serveur_scanner:epkowa:libusb:003:004 source
```

Ceci fait je vais juste éditer le résultat avec Gimp pour n'extraire qu'une seule colonne d'article. En effet , Tesseract n'a pour l'instant pas d'analyseur de format, et le double colonage du newswick, ainsi que la photo centrale, ne va pas vraiment l'aider. Cela donne donc la source suivante (que j'ai réduite en taille pour l'intégrer ici) :

**T**HE FAILURE OF BRUTE-FORCE SCARE TACTICS, OF course, reflected the times, too: by 2006, more voters than in 2004 were fed up with the Iraq War and terrorism alerts that had begun to sound like crying wolf. But something more fundamental was at work. Simply put, candidates do themselves little good by reminding voters of their fears and leaving it at that, for evoking fears without also raising hopes is rarely a winning strategy. "Successful candidates understand voters' fears and anxieties and speak to it," says Matthew Dowd, a former political strategist for Bush and now a contributor to ABC News. "Clinton did this in 1992, saying he understood voters' anxieties about the economy better than Bush 41 did, and [George W.] Bush did it in 2004 when he said he understood their fears and anxieties related to terrorism and security." To close the deal, these successful candidates took the next step: Bill Clinton offered hope by vowing to address economic problems more competently than George H.W. Bush was doing, and the current president offered hope when he said he would protect Americans better than John Kerry could or would. "Politicians who speak only to the fear and anxiety part without transitioning to something more optimistic don't win," says Dowd. "You can't leave voters stuck in their fear

Ensuite je vais créer deux fichiers graphique. L'un au format TIF pour tesseract, l'autre en PPM pour GOCR.

```
$ convert source source.tif
$ convert source source.ppm
```

Le temps du test est maintenant arrivé, d'abord GOCR :

```
$ gocr [3] source.ppm resultat-gocr.txt
```

Ce qui me donne en 3.92s :

```
— — —
HE FAILU_ OF BRUTE-FORCE SCA_ TACTICS, OF
course, relected _e _mes, too: by 2006, more voters
_an in 2004 were fed up ___ the Iraq War and terro_sm
alerts _at had begun to sound like c_' ng wolf. But some-
thing more _ndamentaJ was at work. Simply put, candi-
dates do _emselves litUe good by reminding voters of their fears
and leaving it at _at, for evoking fears m_out also raising hopes
is rarely a mnning strategy. ''Success_l candidates unders_nd
voters' fears and anxl__es and speak to it,'' says Ma_ew Dowd, a
fo_er political strate?ist for Bush and now a cont_' butor to ABC
33ews. ''Clinton did this in 1992, saying he understood voters' anxi-
e_es about the economy be_er _an Bush 4l did, and _George W.J
Bush did it in 2004 when he said he understood _eir fears and
a__ties related to terro_sm and securi_:' To close _e deal, Ulese
success_l candidates took _e next step: Bill Clinton o_ered hope
by vom'ng to address economic problems more competenUy _an
George H.W. Bush was doing, and _e cu_ent president o_ered
hope when he said he would protect Americans be_er _an _ohn
Ker_ could or would. ''Politicians who speak only to the fear and
anxie_ pa_ m'_out transi_oning to something more op_mistic
don't m'n;' says Dowd. ''You can't leave voters stuck in their fear
```

Maintenant le même test avec tesseract :

```
$ time tesseract source.tif result-tesseract
```

Cette fois le résultat tombe en 3.12ss ce qui donne déjà Tesseract comme étant 1.25 fois plus rapide qu'GOOCR :

```
HE FAILURE OF BRUTE·FORCE SCARE TACTICS, OF course, reflected the times, too: by 2006, more voters than in 2004 were fed up with the Iraq War and terrorism alerts that had begun to sound like crying wolf But something more fundamental was at work. Simply put, candidates do themselves little good by reminding voters of their fears and leaving it at that, for evoking fears without also raising hopes is rarely a winning strategy. "Successful candidates understand voters' fears and anxieties and speak to it," says Matthew Dowd, a former political strategist for Bush and now a contributor to ABC News. "Clinton did this in 1992, saying he understood voters' anxieties about the economy better than Bush 41 did, and [George W] Bush did it in 2004 when he said he understood their fears and anxieties related to terrorism and security." To close the deal, these successful candidates took the next step: Bill Clinton offered hope by vowing to address economic problems more competently than George H.W Bush was doing, and the current president offered hope when he said he would protect Americans better than John Kerry could or would. "Politicians who speak only to the fear and anxiety part without transitioning to something more optimistic don't win," says Dowd. "You can't leave voters stuck in their fear
```

## Conclusion

Le résultat est juste sans appel. Le texte produit par Tesseract est tout simplement parfait contrairement ?

Plus rapide donc et surtout bien pour des résultats bien meilleur (même si cela demande à être re-testé en français) tesseract semble d'un coup de baguette magique promettre aux plate-formes libres (ou pas d'ailleurs, car il fonctionne aussi sous Windows) une numérisation de document d'une qualité enfin digne de ce nom. Une très très bonne nouvelle en vérité.

### Liens:

[1] <http://ubunteros.tuxfamily.org/spip.php?article148#forum790>

[2] <http://code.google.com/p/ocropus/>

[3] <http://pwet.fr/man/linux/commandes/gocr>