Tesseract

🔧 Fix Me!

- Big $$$: OCR Scanned PDFs with Pytesseract and Imagemagick - A Step-by-Step Guide for Windows and Mac Yancy Dennis, Medium, 23/03/2023

old stuff...

tesseract

- https://github.com/jlsutherland/doc2text
- https://blog.modeanalytics.com/python-data-cleaning-libraries/
- https://mzucker.github.io/2016/08/15/page-dewarping.html
- tesseract sous xenial → 3.04.01
- http://www.machinalis.com/blog/ocr-with-django/
- http://www.nplug.be/ocr OCR sous linux Tesseract et gImageReader
  - sudo add-apt-repository ppa:sandromani/gimagereader
  - sudo apt-get update
  - sudo apt-get install gimagereader-gtk tesseract-ocr tesseract-ocr-fra tesseract-ocr-eng
- ...
- https://groups.google.com/forum/#!msg/tesseract-dev/dGB3cbFtGUs/x9nEu5vy_LoJ (preserve interwords spaces)
- https://mazira.com/blog/optimal-image-conversion-settings-tesseract-ocr (optmization ghostscript)
- https://doc.ubuntu-fr.org/tesseract-ocr
- http://stackoverflow.com/questions/tagged/tesseract
  - http://stackoverflow.com/questions/38921617/python-reading-an-easy-captcha-tesseract → tesseract commandé à partir de python avec os.system https://docs.python.org/3/library/os.html
  - http://stackoverflow.com/questions/22609778/how-to-preserve-document-structure-in-tesseract option preserve_interword_spaces
  - http://stackoverflow.com/questions/2363490/limit-characters-tesseract-is-looking-for
  - http://stackoverflow.com/questions/8268928/where-can-i-find-samples-of-hocr-files hocr → coordinates
  - http://stackoverflow.com/questions/15199510/blacklist-characters-are-not-ignored-by-tesseract-ocr blacklist
- ...
- http://manpages.ubuntu.com/manpages/trusty/man1/tesseract.1.html man → explication du configfile
- python textract : http://textract.readthedocs.io/en/latest/installation.html

- Python-tesseract https://pypi.python.org/pypi/pytesseract non maintenu
- https://realpython.com/blog/python/setting-up-a-simple-ocr-server/ configuration d'un serveur - compilation de tesseract + python...
- https://mlichtenberg.wordpress.com/2015/11/04/tuning-tesseract-ocr/ exemple de different config file

- https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality problèmes de qualité de scan...
- python camelot : https://hackernoon.com/announcing-camelot-a-python-library-to-extract-tabular-data-from-pdfs-605f8e63c2d5
  - https://camelot-py.readthedocs.io/en/master/
  - excalibur : https://github.com/camelot-dev/excalibur
- tabula (java) : https://tabula.technology/ + https://medium.com/better-programming/convert-tables-from-pdfs-to-pandas-with-python-d74f8ac31dc2
- https://github.com/tesseract-ocr/tesseract
  - https://github.com/tesseract-ocr/docs (tutos, abstract)
  - https://github.com/tesseract-ocr/tesseract/wiki
  - https://github.com/tesseract-ocr/tesseract/wiki/FAQ
- options :
- http://stackoverflow.com/questions/37082294/how-to-properly-ocr-typewriter-fonts-using-tesseract-and-python
  - tessedit_char_whitelist http://stackoverflow.com/questions/2363490/limit-characters-tesseract-is-looking-for

Remove gridlines :

- http://stackoverflow.com/questions/13280952/opencv-remove-gridlines-from-sudoku-puzzle
- http://stackoverflow.com/questions/27587343/improve-tesseract-detection-quality
- http://stackoverflow.com/questions/33949831/whats-the-way-to-remove-all-lines-and-borders-in-imagekeep-texts-programmatic
- http://www.multipole.org/discourse-server/viewtopic.php?t=23723

exemple (cf. http://www.imagemagick.org/script/command-line-processing.php) :

```
convert input.png \
  -negate \
  -define morphology:compose=darken \
  -morphology Thinning Rectangle:1x30+0+0 \
  -negate \
  e2.png
```