Optical character recognition using Tesseract-OCR

Assignment 5

Optical character recognition or optical character reader (OCR) is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example: from a television broadcast). Tesseract is an optical character recognition engine for various operating systems. Originally developed by Hewlett-Packard as proprietary software in the 1980s, it was released as open source in 2005 and development has been sponsored by Google since 2006.

Working process:

1. Install pytesseract, pillow.

2. Download Tesseract-OCR (from https://github.com/UB-Mannheim/tesseract/wiki) and install.

3. Add tesseract.exe path from installation folder to a variable
pytesseract.pytesseract.tesseract_cmd='your\\path\\ tesseract.exe'

4. Perform getting text from pictures using pytesseract.

5. Choose any social media source to retrieve at least 100 pictures (e.g. posters).

6. Analyze collected data. You can use any tools from previous assignments.