**5. OCR (3 people a group)**

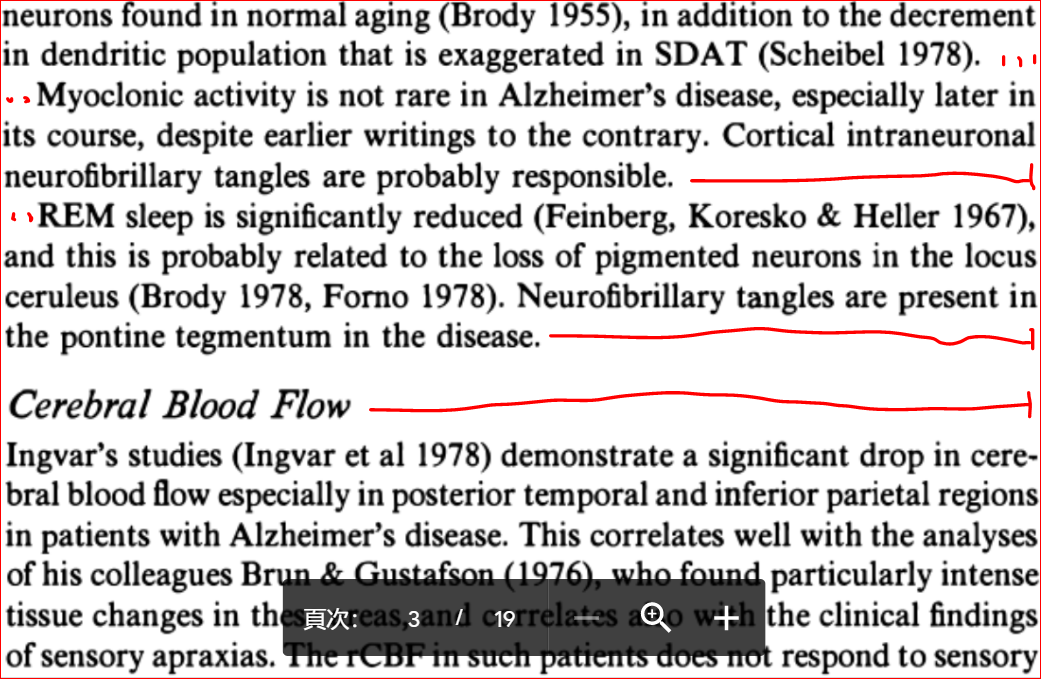
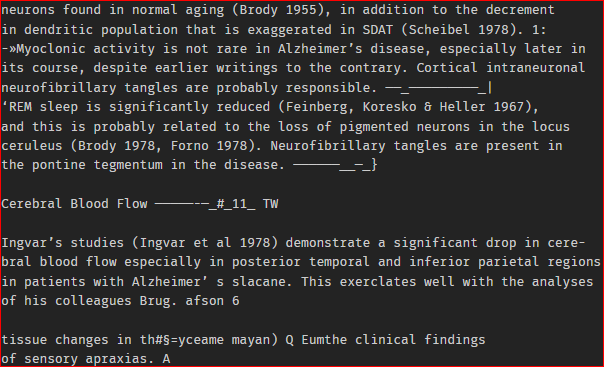
* Extract text data from image data with the OCR technique after image pre-processing.
* Skill Set: OCR (Tesseract), image processing, Machine Learning, python pdf parser and maybe some textual analysis

### 給第四組Img Paper Paragraph Segmentation 組長黃喬永

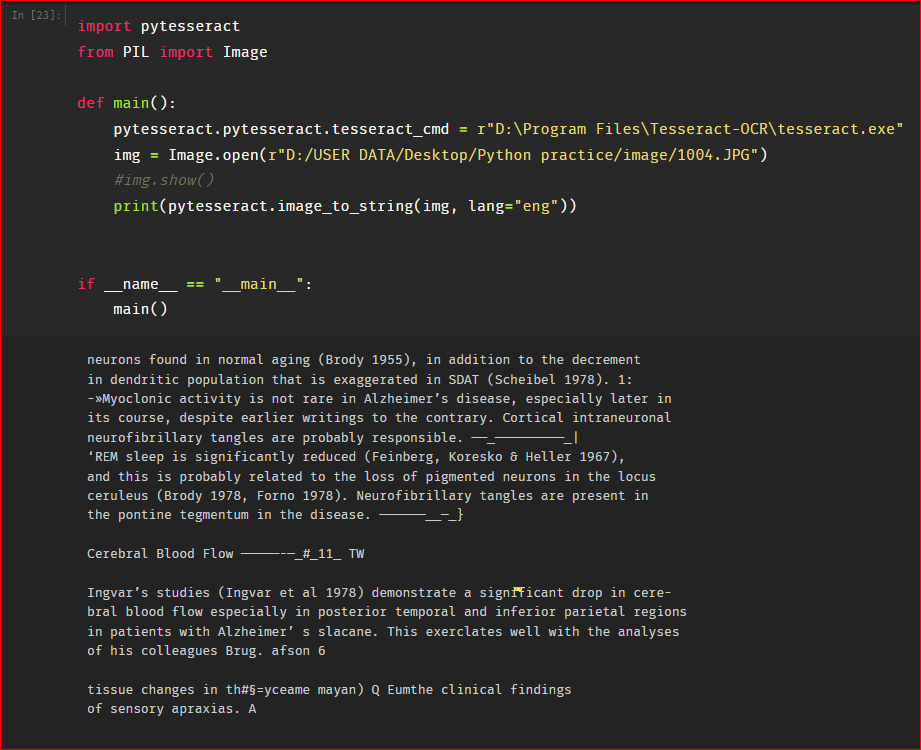
OCR 英文 印刷文字辨識 ，字元識別， 論文段落，依照閱讀論文的瀏覽順序，去切割出每個subtitle底下的文字區塊。

1. **Image論文，每一個文字區塊，邊緣完全切除不要留任何空白。以便進行 OpenCV 印刷文字image辨識。**

有一些要先跟第四組Img Paper Paragraph Segmentation 討論，希望他們文字區塊邊緣不要留空白，1.each paragraph. 2.subtitle and content in each block.3. metadata, main title, and subtitle     存檔方式第1號論文Image區塊 ：1001 、 1002 、1003、1004、………..1999..

Pytesseract辨識效果相當不錯，似乎不需要使用Pillow作影像前處理，只需注意文字區塊邊緣不要留任何空白。



1. **剪下的Image論文區塊可以以編碼命名檔案名稱。**

第1號論文Image區塊 ：1001 、 1002 、1003、1004、………..1999..

第2號論文Image區塊： 2001 、 2002 、2003、2004、………..2999..

**C.論文影像辨識Pillow(依據影像品質做前處理，可能用不到)**

1.需要使用Pillow •BLUR (模糊) •CONTOUR (輪廓)

2.•EDGE\_ENHANCE (邊緣增強)•EDGE\_ENHANCE\_MORE(邊緣更增強)

•SHARPEN (銳利化)

3.將圖片RGB的數值做轉換

from PIL import Image

img = Image.open("RGB.jpg")

width, height = img.size

for y in range(height):

    for x in range(width):

        rgba = img.getpixel((x,y))

        rgba = (255 - rgba[0], # R

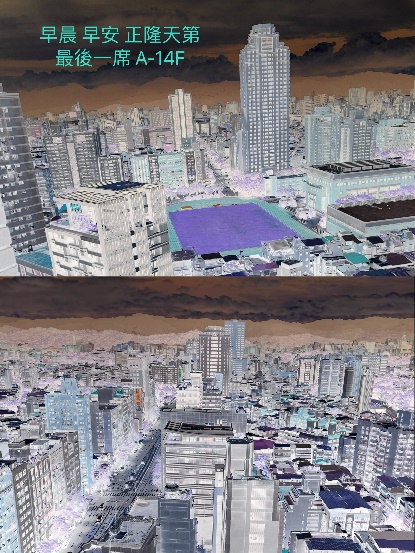
            255 - rgba[1], # G

            255 - rgba[2] ) # B

        img.putpixel((x,y), rgba)

img.show()

img.save("RGB\_new.png")

**D.Q&A:**

可能不用openCV，但為什麼要用Machine Learning ??

Segmentation Label要用在哪??

OCR文字區塊辨識後的應用? 1.相同的研究方法 2.使用模型種類2.同類型摘要論文3. 同類型論文結論

是否論文依編碼1001 、 1002，而不要使用論文名稱?