OCR識別:

使用到的模型主要為檢測文本區塊的CTPN以及作為不定長OCR識別的CRNN +CTC。

主要程式依據的為第二個reference:

程式碼改動的地方包括:

1. 從python 2.7 到python 3.6 部分語法的修正變動
2. Pytorch 0.1.12到1.3.1的語法修正變動
3. 在train/pytorch\_train/train.py內對CTC loss的改動
4. 在model.py裡126~137行間增加對crnn識別出的文本進行結構化的提取以及輸出成文字檔(nlptest.csv)。
5. 增加applications/medicalRecord.py，此py檔為將病歷進行結構化識別。

To do:

1. 英文單詞間空白的切分，在crnn.keys的字典中加入空白的標註，訓練集裡也需空白的標註，但此方法需另外準備資料集；或者將模型識別後的單字用其他NLP的方式做分割。
2. 部分區塊垂直文本的識別改善
3. 繁體中文數據集對模型的調校訓練
4. 病歷並非如身分證之類的完全結構化單一模板，也許可以先找出關鍵詞於影像中的相對位置，如識別出”病名”的文字框的中心點x座標為靠左/置中(表1/表2)，以此得知對應內容的位置(於其右/其下)，挑出處於此區塊的所有文本，再對此區域的文字進行二次識別。

表1 表2

 