HƯỚNG DẪN SỬ DỤNG CHƯƠNG TRÌNH VNDOCR 4.0 BẢN DEMO

1/ Nguồn chương trình:

Vào link sau download : http://www.vndocr.com/

Chạy setup chương trình, nên chọn option là full installation

2/ Cơ bản:

Chương trình VnDOCR (Việt Nam Document Optical Character Recognition - phần mềm nhận dạng tài liệu tiếng Việt) là 1 trong rất nhiều chương trình OCR, ưu điểm duy nhất là có thể nhận dạng tiếng Việt OCR sử dụng các thuật toán để nhận dạng từ dữ liệu dạng bitmap (file ảnh) sang dữ liệu dạng text, do đặc thù thuật toán, các dạng file thường sử dụng là dạng file ảnh nén không tổn thất (lossless compression) dạng dữ liệu liên tục (bmp, tif. . .) Các dạng file ảnh nén có tổn thất, dùng kỹ thuật xen dòng (interleave, kiểu quét dòng của TV) thường không sử dụng được cho các OCR (các dạng file này phổ biến trên intemet như gif, jpeg. . .) Do vậy, bằng bất kỳ hình thức nào (quét ảnh, chụp ảnh, ảnh có sẵn) muốn đưa vào OCR cũng phải chuyển đổi về dạng thức file ản tương thích. Thường chọn dạng windows bitmap (.bmp) hoặc dạng TIFF (.tif) là chuẩn nhất. Mặc dù các OCR có khả năng nhận dạng trong các ảnh màu, nhưng để tăng độ chính xác và giảm tải cho hệ thống, nên chuyển về dạng thức ánh đen trắng (l bit màu, hay còn gọi là monochrome, B&W trong các chương trình xử lý ảnh) hoặc ảnh xám (greyscale, có thể chọn 4 bit hoặc 8 bit Với các nguồn tài liệu sạch, chữ rõ ràng, nên dùng dạng thức 1 bit để xử lý nhanh. Với các tài liệu cũ, in xấu, nhòe . . . nên dùng dạng thức ảnh xám để có thể xử lý nâng chất lượng trước khi đưa vào OCR. Các bộ lọc dùng để xử lý các ảnh này thường là unsharp mask, level và các bộ lọc khử nhiễu, tùy vào chương trình xử lý ảnh sử dụng.

Trong VnDOCR cũng có sẵn chức năng nâng cao chất lượng ảnh, tuy nhiên thô sơ và không thể hiệu quả bằng các trình xử lý ảnh chuyên nghiệp. Nếu bạn có hứng thú hoặc có nhu cầu làm việc lâu dài với OCR, việc thử và rút kinh nghiệm với những chức năng built-in này cũng khá thú vị. Các chức năng này nằm ở sub menu Công cụ > Công cụ ảnh.

Sau khi khai vị bằng các dạng thức ảnh và việc làm sạch ảnh 1 cách sơ bộ, ta quay ngược lại về việc quét/chụp ảnh (scan) cho mục đích OCR. ảnh cần được quét ở các dạng thức thích hợp như trên và ở độ phân giải nên dùng là 300dpi. Ở độ phân giải này đảm bảo các chữ cỡ co 8 sẽ đủ số điểm ảnh tối thiểu để nhận dạng chính xác.

Khi quét ảnh, cần điều chỉnh các tham số quét đảm bảo các yêu cầu sau:

- Độ phân giải đủ

- Các chữ không bị đứt nét (chỉnh tương phản, độ sáng tối, level . . . trong bảng điều khiển quét)

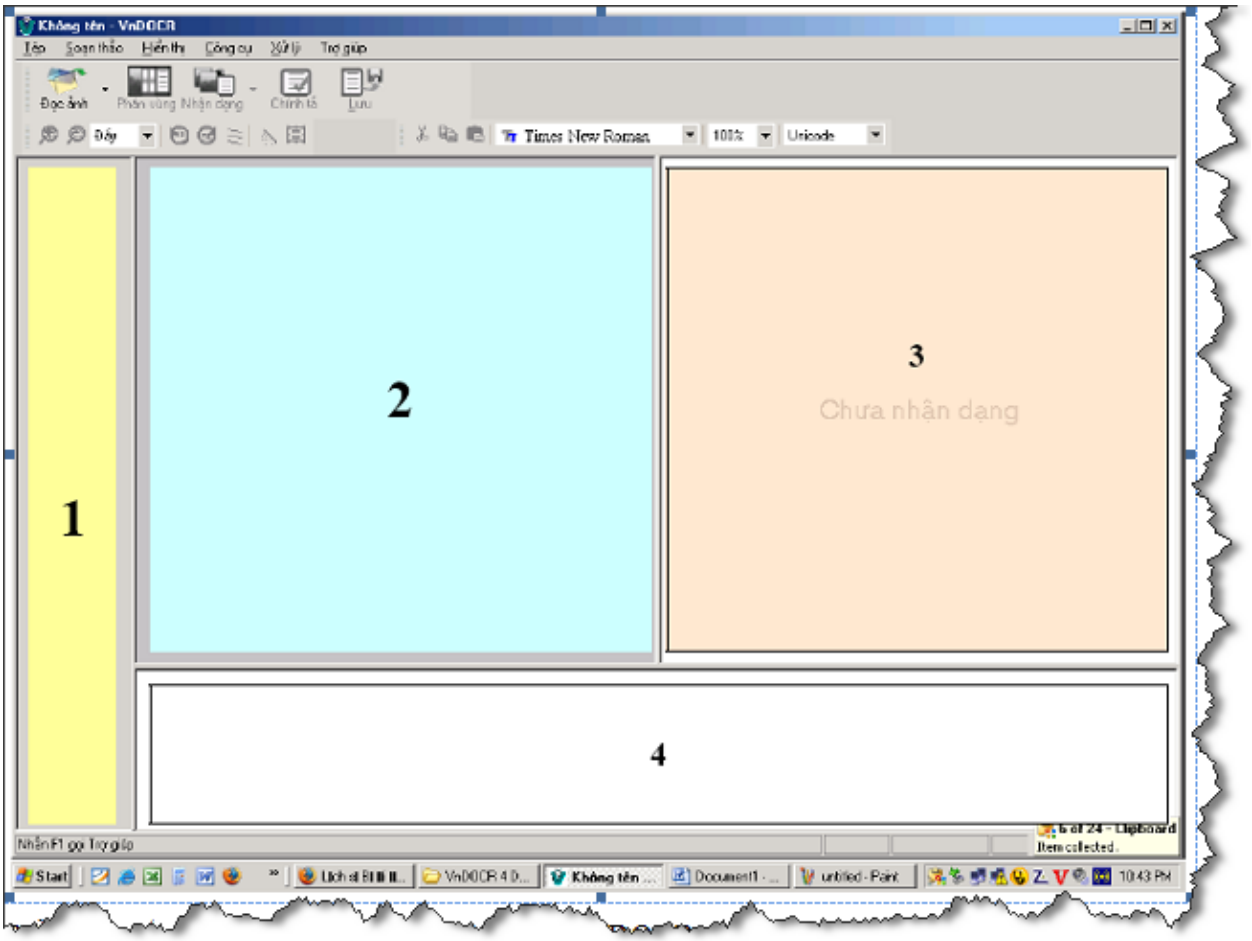
- Các vết ố, vết bẩn bị mờ đi tối đa (chỉnh chức năng xóa nhiễu trong bảng điều khiển quét)

- Ảnh kết xuất đúng dạng file cần thiết.

Làm tốt công tác đầu vào này, các công đoạn chỉnh sửa bằng trình xử lý ảnh có thể bỏ qua mà không làm ảnh hưởng nhiều đến quá trình OCR.

3/ Sử dụng chương trình:

Khi khởi động chương trình, sau màn nhắc rằng bạn chưa trả tiền, màn hình chính sẽ hiện ra như sau:



Điều cần để ý đầu tiên là thanh Toolbar của VnDOCR có các nút nhấn nhiều option (bên cạnh các nút ấn có mũi tên xuống để chọn các chế độ khác nhau)

Màn hình cơ bản có 4 khu vực:

1: Bên trái cùng là Navigation

2: Ô lớn trên, bên trái là Preview các ảnh nhập vào

3: Ô lớn trên, bên phải là màn kết quả OCR

4: Ô dài nằm dưới là Ô tham khảo.

Khi di chuyển con trỏ soạn thảo trong Ô 3, các bitmap tương ứng sẽ hiện trên Ô 4 để tiện việc đối chiếu chỉnh sửa.

Các Menu đều bằng tiếng Việt, tuy hơi chuối nhưng chắc khó có thể tìm từ hay hơn, và đủ để không cần giới thiệu về ý nghĩa. Về chức năng ta sẽ làm rõ ở phần sau.

Bắt đầu, bao giờ cũng là việc mở file ảnh đã chuẩn bị ra. Bản VnDOCR 4 Demo giới hạn 3 trang ảnh (nghĩa là 3 trang trong khung 1 ,còn không giới hạn có bao nhiêu chữ trong mỗi trang ảnh, nghĩa là ta có thể ghép ảnh nhiều trang sách vào 1 trang ảnh để OCR 1 lần cho tiện, tuy nhiên kinh nghiệm bản thân tôi cho thấy chả tiện hơn tý nào)

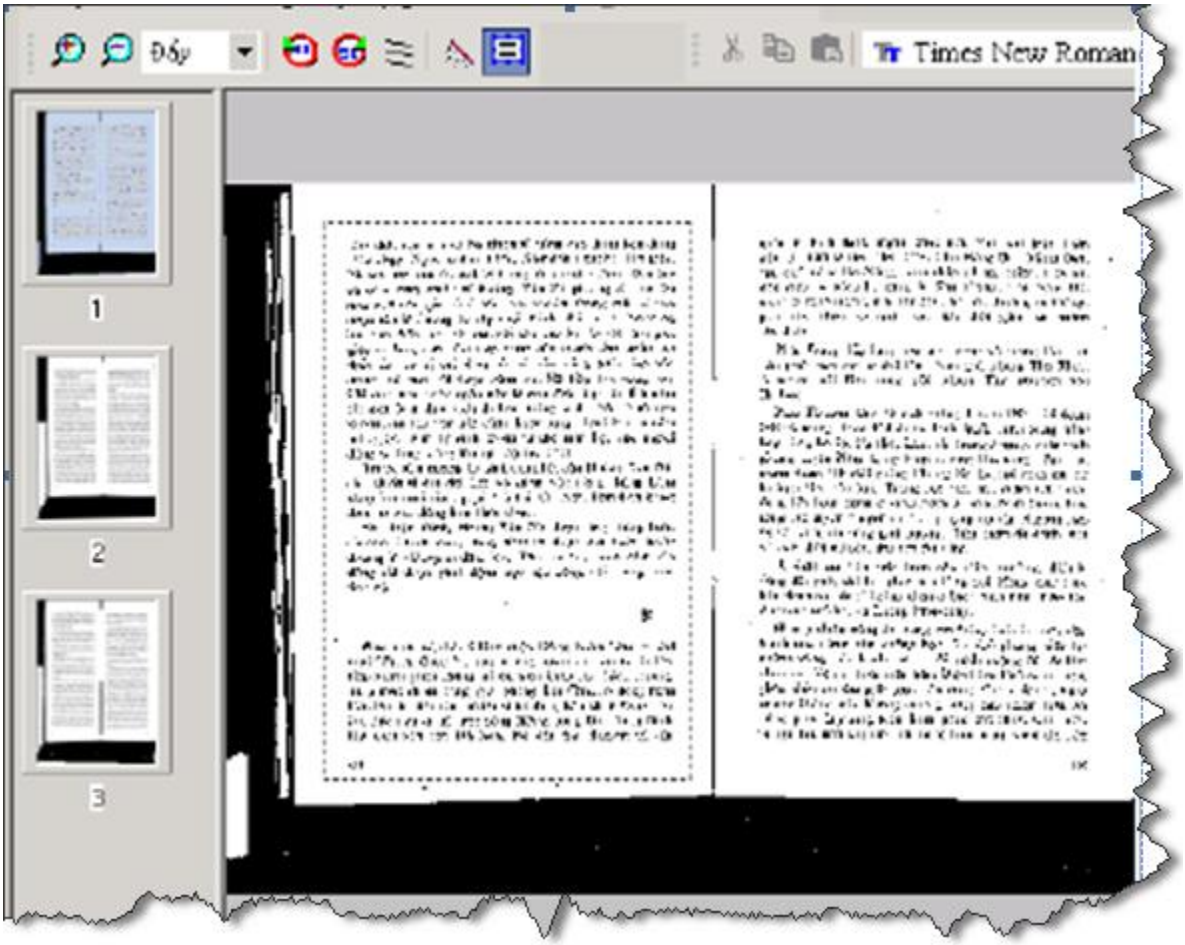
3 trang ảnh có thể là 3 file ảnh độc lập, hoặc 1 file ảnh multi-pages (cái này chỉ định dạng .tiff mới có, phục vụ cho việc chứa các chanel màu khác nhau trong công tác tách màu của chế bản điện tử).

Sau khi mở xong, ta bắt đầu bằng việc xem xét chất lượng ảnh. Nhanh nhất là ta ấn nút nhận dạng, nhận thử 1 trang rồi xem kết quả thế nào

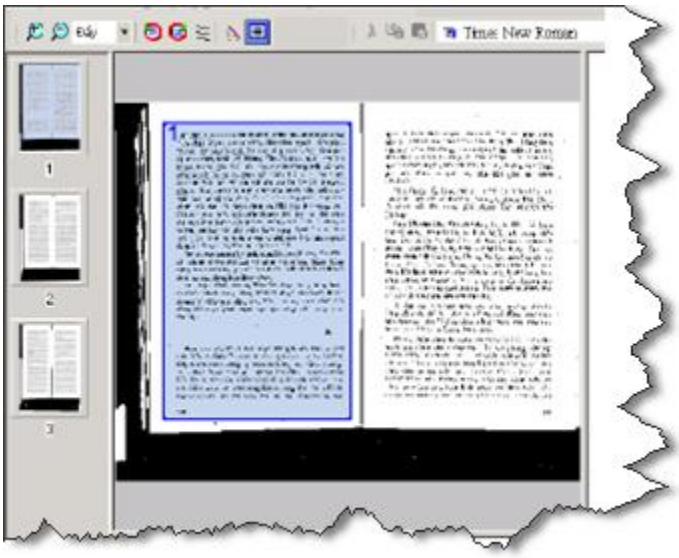
Nếu kết quả sai ít (tất nhiên tùy người và tùy hoàn cảnh, việc đánh giá thế nào là ít có thể là từ 1 lỗi đến 100 lỗi, cái này các bạn phải tự quyết lấy thôi) thì ta bắt đầu luôn việc chính là nhận dạng.

Khởi đầu công việc là phân vùng ảnh. Chương trình OCR nào cũng có khả năng tự phân vùng, nhưng nếu bạn để chương trình làm, các đoạn văn nó ném sang cửa sổ số 3 sẽ là một đám mà khi nối lại xong, bạn sẽ hối hận là tại làm sao không tự đánh máy ngay, chả cần đến scanner và OCR mà vẫn nhanh hơn được. Vì thế, tốt nhất ta tự phân vùng bằng tay, bằng cách ấn vào nút "Tạo vùng mới", là cái nút nho nhỏ :

Rồi, thật cẩn thận, kéo nó quanh cái vùng mà bạn nhận là "trang"



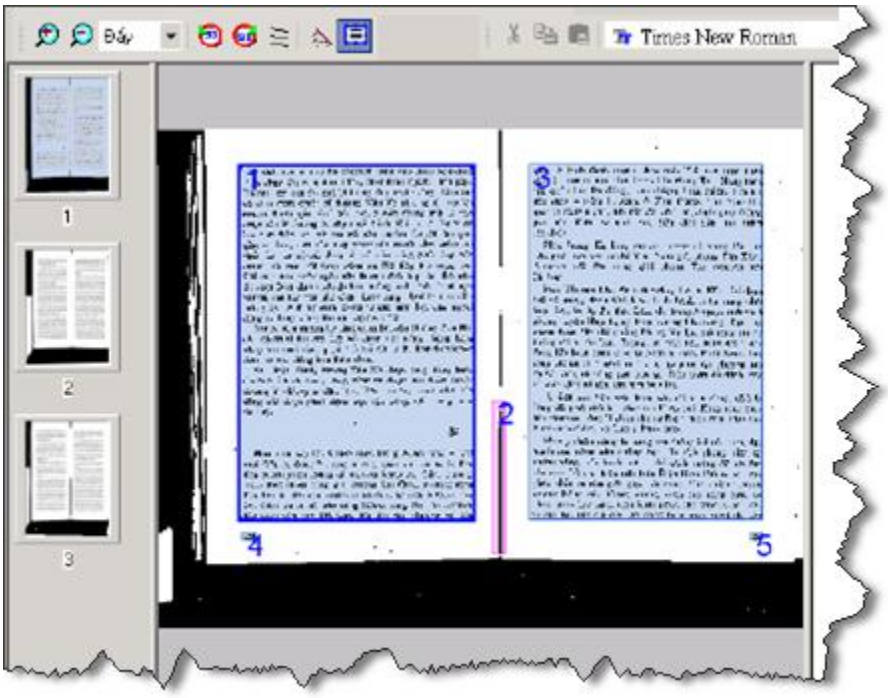
Nó sẽ đánh dấu vùng của ta như thế này:



Cứ làm thế lần lượt, theo thứ tự của các trang sách, cho đến trang cuối.

Nếu lỡ tay, bạn có thể di chuyển chuột đến vùng biên chọn vừa làm và nhấn nút, sẽ có các nút hiện lên để bạn di chuyển hoặc chỉnh lại kích cỡ vùng chọn. Nếu lúng túng quá, bạn cứ ấn nút delete để xóa và chọn lại. Luống cuống nữa, thì vào mênh "Xử lý" > "Đánh dấu vùng bằng tay", sẽ có chức năng xóa giúp bạn. ..

Còn nếu bạn lười, ấn nút phân vùng tự động, số trang ta có sẽ buồn cười như sau:



Bạn thấy không, từ 2 trang sách 13x18 trong con mắt của ta, nó lại phân thành 5 trang. Sẽ khó chịu là khi OCR, nó cứ từ trang 1 tới 5 mà nối, thế là trong các áng văn của chúng ta có đám chữ giun" (nhận ở "trang 2" tự động) và 2 số trang (trang 4 và trang 5 tự động).

Tình hình sẽ còn tệ hơn nhiều nếu trang sách thật của ta có nhiều đoan, ví dụ 1 đoạn chú thích cuối trang chẳng hạn.

Đánh dấu xong rồi ta lại ở 2 trường hợp

Trường hợp đầu tiên bạn sẽ gặp khi đọc cái tài liệu củ chuối này, hoặc một thứ tương tự đó là trường hợp bạn chưa hề "dạy dỗ" gì OCR để giảm thời gian sửa lỗi, tăng xác suất chính xác của chương trình.

Như ta biết, mỗi loại tài liệu thì có một ông chế bản, một ông in. Mỗi ông chế bản quyết định sổ kiểu chữ, cỡ chữ, phân trang. . . mỗi ông in cho một kiểu nhòe chữ hoặc đứt nét chữ, một kiểu giấy bẩn mực khác nhau. Mà OCR không phải là thứ siêu đến mức đoán được ý chí, tình cảm và túi tiền của chừng ấy ông chế bản và ông nhà in trên đời, nên ta phải dạy nó thích ứng với từng quyển sách.

Đầu tiên, OCR là ý chí của ông viết sốp (soft), nên ta phải thử xem mặc định ông viết sốp dạy nó thế nào. Ta cứ ấn nút nhận dạng, cho nhận dạng tất cả các trang đã.

Sau đó, vào Sub Menu "Xử lý" > "Học"

Cái này sẽ hiện ra:

Tất nhiên bấm menu "Học" lại ra dialog "Dạy" là một sự hay, nhưng đó không phải bản chất. Hãy nhìn vào bố cục của bảng, ta thấy nó gồm 3 phần nhỏ, bên tay trái: là phần mà OCR tự "xẻo" từng chữ ra để nhận dạng. Bên tay phải, hiện phần ảnh nơi OCR "xẻo,, được chữ. Hàng dưới là 1 loạt các "dụng cụ giảng dạy".

Như ta thấy trên hình, chữ "ế" trong tài liệu này đã được OCR cắt đúng, nhưng không nhận diện được. Lý do dễ hiểu là vì trong tài liệu này (tài liệu F3 16 tập l) cả font chữ và cách in làm dấu gắn liền đỉnh chữ (có thể thấy ở chữ ễ trên đó).

Còn hình thù cổ quái bên trái chữ ế, nếu bạn nhấn vào đó sẽ thấy đó là đuôi chữ ư . Chính vì OCR nhận dạng kiểu đó nên ở phần trên đã nhắc rằng bằng các cách, hình của chữ trong ảnh phải liên tục (không bị đứt nét) nếu ta không muốn có tài liệu chữ giun.

Dông dài thế đã *đủ.* Giờ ta dạy OCR bằng cách ấn vào tùng chữ ở phía trái, xem bên phải để tự đánh giá nó là chữ gì, rồi đánh chữ đó vào combo box "Chữ" ở phía dưới. Đánh xong chữ vào Ô này chưa phải đã xong, bạn phải tự phân loại nó là ký tự gì (trong bản chữ cái tiếng anh như a, b, c thì chọn "Chữ Anh", ă, â, đ. . . . chọn "Chữ Việt thường" : . . . rồi ấn Dạy"

Lưu ý đặc biệt: **Chữ gõ vào combo box phải chuyển bộ đánh sang code TCVN - ABC. nếu không chữ nhân dạng ra sẽ kỳ quái lắm.**

Công việc chưa kết thúc ở phần đầu danh sách bên trái, là phần mà các ký tự OCR không thể nhận diện được. Bạn phải kéo xuống phần cuối cùng, nơi mà OCR đã nhận dạng được ký tự, để kiểm tra xem nó có nhận sai hay không. Nó như thế này:

Như bạn thấy, trong dạng tài liệu của quyển sách ví dụ, F3 16 tập 1, do dấu dính vào đỉnh chữ nên các chữ rõ ràng là ả, ắ, ả . . . đều bị quy chụp là chữ â. Giờ ta phải mất công dạy lại OCR để nó nhận diện cho đúng. Kéo xuống chút nữa, bạn sẽ thấy chữ d và đ bị nhầm lẫn rất nhiều, do dấu gạch ngang của chủ đ trong font chữ của sách này quá nhỏ. Và cuối cùng, trước khi thoai, đừng quên ấn nút ghi" để ghi lại các sửa đổi. Hãy dùng các option mặc định của chương trình, vì khi tôi thử thay đổi, bản dạng này coi như điếc đối với sự dạy dỗ công phu của ta Với các phần cắt ra là dấu, phải hết sức để ý kẻo dạy sai:

Như khung ví dụ trên, dấu không được nhận dạng. Đầu tiên bạn phải chọn option "Dấu chữ Việt" trước, sau mới chọn d ấu tương ứng là “dấu ả” rồi ấn Dạy. Nếu làm không đúng thứ tự này, dấu ả sẽ được thay bằng chữ "dấu”.

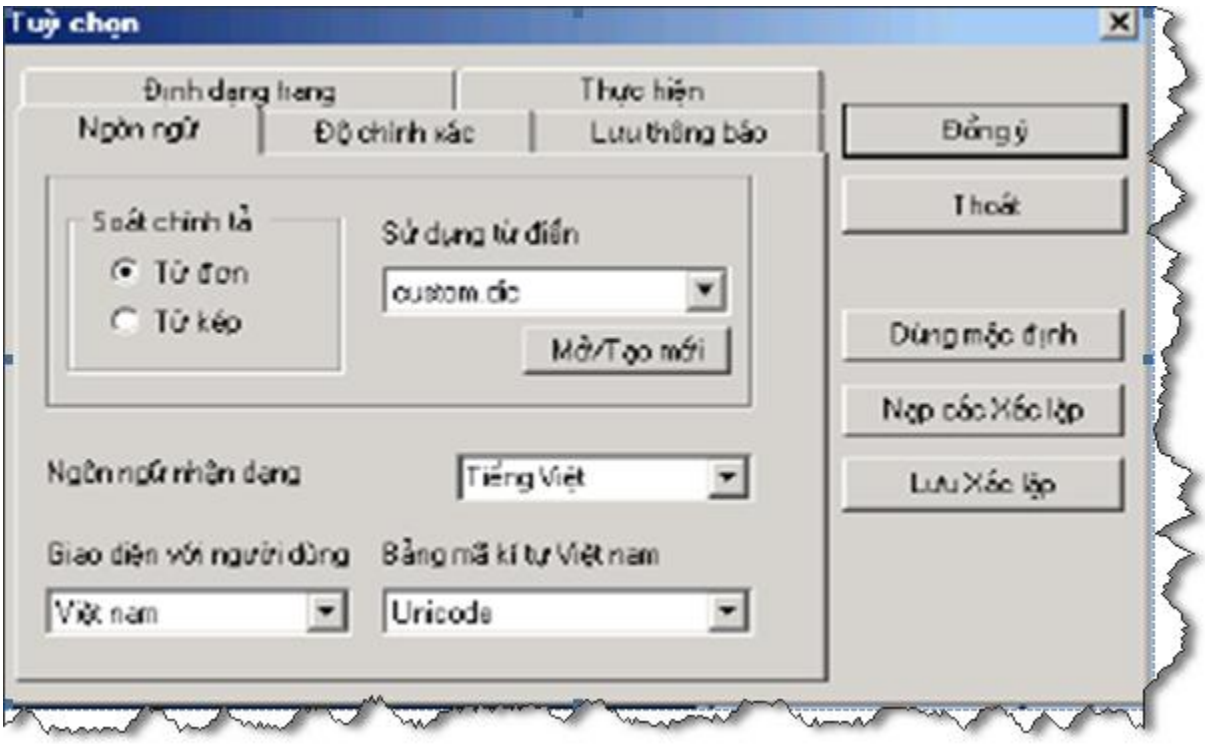
Thật kiên nhẫn làm tốt phần "dạy" này, ta đạt được độ chính xác OCR tới 98% trên cái tài liệu quét kiểu củ chuối, giảm rất nhiều công kiểm lỗi. Hy vọng các bạn đạt được độ chính xác cao hơn. Tuy nhiên, với quyển sách 200 trang, nếu bạn làm công tác dạy dỗ thật kỹ lưỡng trên kho ảng 10 đến 20 trang đầu, xác suất trúng đích đã là tối đa, dạy dỗ thêm cũng không tăng thêm được độ chính xác. Tất nhiên đó là nói với tình trạng những quyển sách in kiểu cũ như F3 16 này. Với các tài liệu mới in chữ to, ít vết bẩn và dùng font hợp lý, thậm chí không cần dạy dỗ tý nào (tôi thử chụp màn word rồi đưa sang, 100% mà không cần dạy). Như vậy, việc có dạy hay không và dạy như thế nào phải tùy thuộc vào quyển sách và độ cầu kỳ của bạn!

Xong phần dạy, bạn nhớ Ghi nhé.
Quên mất một điều: những chữ không hoàn chỉnh, hoặc xấu quá, chính ta luận mãi mới ra, thì bạn đừng dạy OCR nhé, kẻo nó tẩu hỏa nhập ma!

Sau khi dạy xong, bạn lần theo đường dẫn tìm file train.trn trong thư mục của VnDOCR, backup file này lại. Kinh nghiệm cho thấy thỉnh thoảng file này vẫn tồn tại những chương trình coi như "chữ thầy trả thầy", kết quả OCR sai thậm tệ. Đó là lúc lôi file backup này thay cho file hiện hành để làm việc tiếp. Chắc đây là lỗi của bản demo.

Kiên trì dạy cho chương trình đến khi bạn cảm thấy xác suất lỗi là chấp nhận được, hoặc không thể tăng được nữa, ta quay sang bắt tay vào công việc chính: nhận dạng văn bản, sửa lỗi và đưa sang file text. Bắt đầu mở các file ảnh từ đầu, chọn vùng bằng tay và nhấn nút nhận dạng, chọn tất cả các trang.
Kết quả đến ngay.

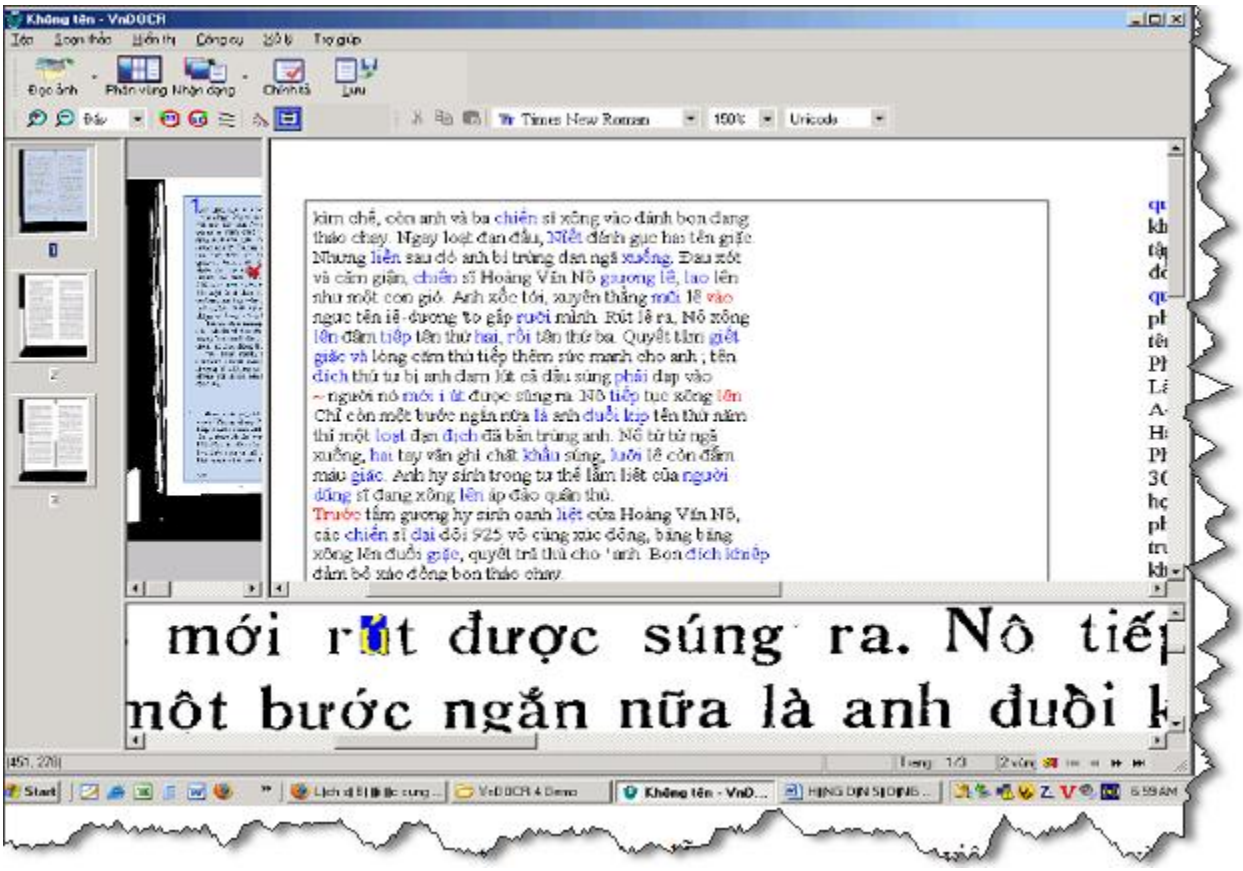
Tuy nhiên trước khi nhấn nút "Nhận dạng" ta chớ quên chỉ định cho OCR sang bảng mã Unicode (VnDOCR 4.0 xuất luôn được Unicode, đỡ mất công chuyển mã và sửa lỗi sau chuyển mã)



Bảng tùy chọn này nằm ở menu "Công cụ" > "Tùy chọn"

Sau khi nhấn "Nhận dạng", tất cả các phần của màn hình sẽ có nội dung. Ta chỉnh sửa luôn sau khi OCR là hợp lý nhất, do còn ảnh trang sách và công cụ đối chiếu của VnDOCR sẽ giúp ta làm rất nhanh.

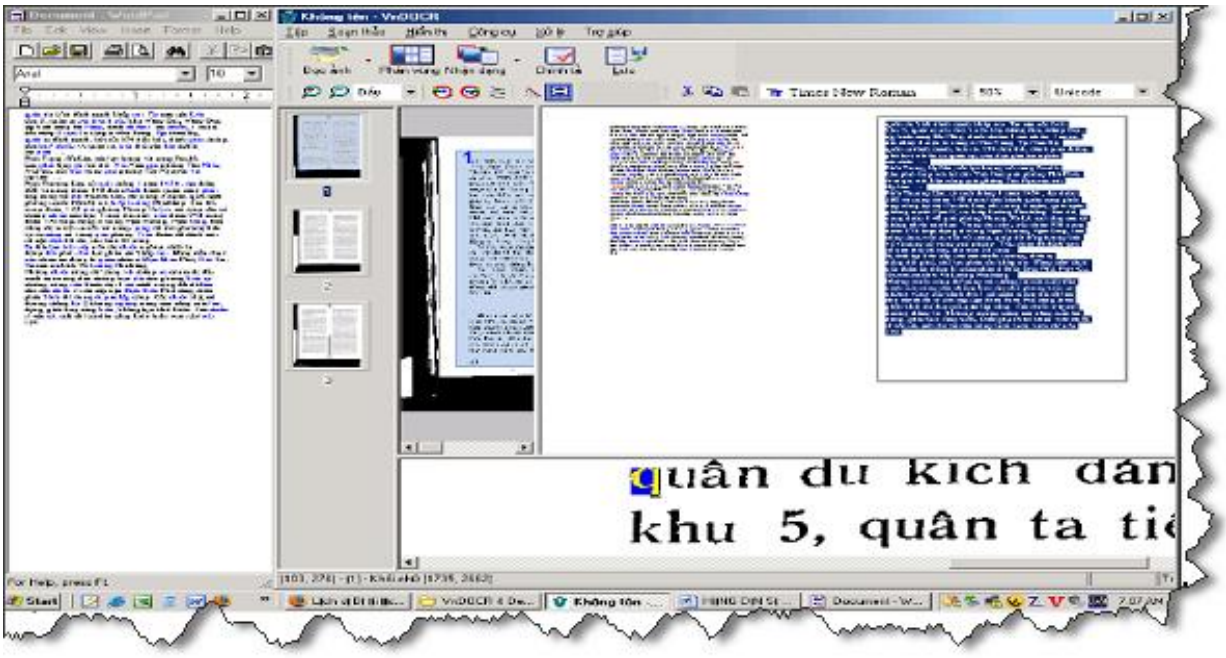
Đây là màn hình chỉnh sửa, khi trong cửa sổ text (số 3) đặt zoom ở 150%, thu nhỏ cửa sổ bitmap lại để lấy chỗ:



Việc highlight phần tương ứng với con trỏ text sẽ làm dễ dàng hơn rất nhiều việc đối chiếu và sửa lỗi

Sau khi sửa lỗi xong, ta sẽ chuyển text sang 1 file text. Vì bản demo không cho copy, và bỏ chức năng save ra text file, nên ta sử dụng drag-n-drop để chuyển text sang Word

Cụ thể, ta mở 1 cửa sổ Wordpad, co kéo để cửa sổ này song song với cửa sổ OCR. ở cửa sổ text (số 3) lấy zoom 50%, đánh dấu từng trang text và kéo - thả sang Wordpad



Thế là Xong.