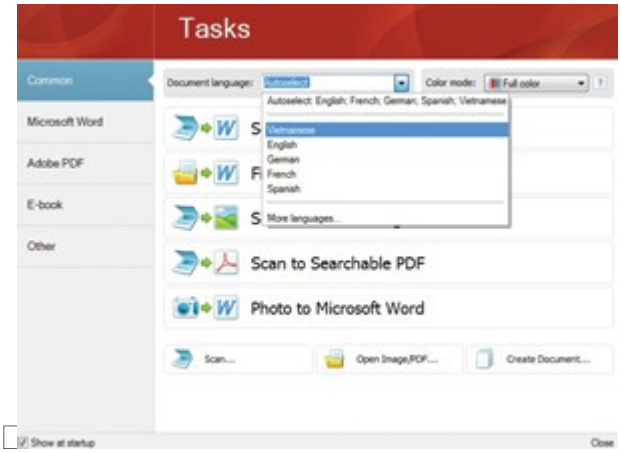
Số hóa tài liệu không cần Internet

Trước đây, qua bài viết "Số hóa tài liệu tiếng Việt" (ID: A1006_124), chuyên mục Làm thế nào đã có dịp chia sẻ với độc giả một số trang web và phần mềm dùng công nghệ nhận dạng ký tự quang học hay nhận dạng chữ in, chữ đánh máy (OCR - Optical Character Recognition)

Cho phép trích xuất và chuyển đổi tài liệu tiếng Việt dạng ảnh (ảnh từ máy quét, máy ảnh, tập tin PDF dạng ảnh…) thành các tài liệu có thể biên tập (dạng tập tin văn bản – text ví dụ Microsoft Word…).



Hình 1: Giao diện ABBYY FineReader Professional 11 khá trực quan và dễ dùng

Nhìn chung, ưu điểm của các trang web số hóa tài liệu tiếng Việt là sự thuận tiện, người dùng có thể dùng bất kỳ máy tính nào để truy cập dịch vụ số hóa, đăng nhập tài khoản và sử dụng. Tuy nhiên, nếu bạn là chuyên viên soạn thảo hợp đồng, nhân viên văn thư, hay công việc đòi hỏi phải thường xuyên chuyển sách báo, văn bản, biểu mẫu tiếng Việt in trên giấy thành tài liệu lưu trữ có thể chỉnh sửa được trên máy tính thì việc số hóa tài liệu tiếng Việt trên trang web đòi hỏi bạn phải luôn luôn kết nối Internet. Vì vậy, nếu đường truyền Internet trục trặc thì công việc số hóa tài liệu của bạn cũng bị ảnh hưởng.

Giải pháp cho việc số hóa tài liệu tiếng Việt không cần Internet là sử dụng phần mềm cài đặt trên máy tính. Bài viết "Số hóa tài liệu tiếng Việt" trước đây từng giới thiệu VietOCR, một chương trình nguồn mở Java/.NET, hỗ trợ nhận dạng tài liệu tiếng Việt ở dạng ảnh PDF, TIFF, JPEG, GIF, PNG, và BMP (xem thêm http://vietocr.sourceforge.net)

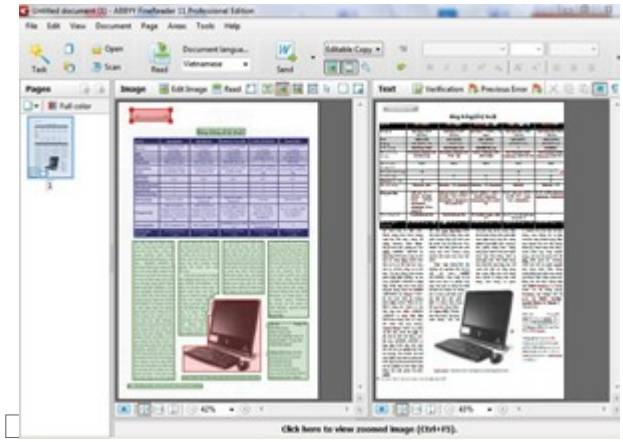
Trong bài viết này, chúng tôi giới thiệu phần mềm số hóa tài liệu tiếng Việt ABBYY FineReader Professional 11. ABBYY FineReader Professional 11 dùng công nghệ nhận dạng tài liệu ADRT (Adaptive Document Recognition Technology) của hãng ABBYY – Nga, có khả năng nhận dạng các cấu trúc logic, cách dàn trang cũng như các định dạng khác nhau trong tài liệu nhiều trang, ví dụ: Mục lục, đầu trang, chân trang, chú thích bảng, chú thích ảnh

ABBYY FineReader Professional 11 hỗ trợ nhiều kiểu định dạng tập tin đầu vào như BMP, PCX, DCX, JPEG, JPEG 2000, JBIG2, PNG, TIFF, PDF, XPS, DjVu, GIF,

WDP và nhiều kiểu định dạng tập tin đầu ra như DOC, DOCX, XLS, XLSX, PPTX, RTF, PDF, PDF/A, HTML, CSV, TXT, ODT, DjVu, EPUB, FB2. Hiện phần mềm ABBYY FineReader Professional 11 có thể nhận dạng tài liệu của 189 ngôn ngữ, trong đó có tiếng Việt. Bạn có thể tải về dùng thử ABBYY FineReader Professional 11 tại http://finereader.abbyy.com/professional. Phiên bản dùng thử 15 ngày cho phép số hóa 50 trang tài liệu và mỗi lần số hóa 1 trang đầu tiên trong danh sách.

Sử dụng dễ dàng

Trước tiên, bạn tải về phần mềm ABBYY FineReader Professional 11 và cài đặt vào máy tính. Để minh họa bài viết, chúng tôi chuẩn bị sẵn tập tin đầu vào bằng cách dùng máy quét (scan) HP LaserJet 100 Color MFP M175a quét một trang trong Tạp chí Thế Giới Vi Tính với độ phân giải 300dpi, ảnh giai sắc xám (grayscale), độ sâu màu 24 bit, định dạng JPG. Tài liệu đầu vào có định dạng bảng, chữ in đậm, chữ hoa, chữ thường, chia cột, ảnh, chú thích ảnh, chữ chân trang. Sau đó, chúng tôi chạy chương trình ABBYY FineReader Professional 11.



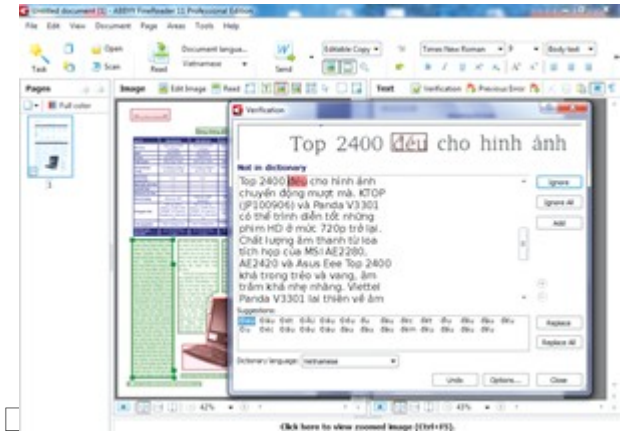
Hình 2: Bên trái là tài liệu đầu vào, bên phải là tài liệu đầu ra sau khi chương trình tự động nhận dạng và phân tích.

Giao diện ABBYY FineReader Professional 11 khá trực quan và dễ dùng. Ngay màn hình đầu tiên, bạn có thể tùy chọn ngôn ngữ của tài liệu cần số hóa hoặc để chế độ chương trình tự động nhận dạng ngôn ngữ (autoselect). Tiếp theo chọn phương thức số hóa tài liệu: Kiểu tập tin đầu vào và kiểu tập tin đầu ra. ABBYY FineReader Professional 11 cung cấp 5 chế độ số hóa tài liệu: Thường dùng (common), Microsoft Word, Adobe PDF, E-book, các chế độ khác (Other). Trong mỗi chế độ lại có nhiều phương thức số hóa tài liệu, chẳng hạn trực tiếp từ máy quét sang tập tin Microsoft Word, từ tập tin (PDF/hình ảnh) sang Word, từ ảnh trong máy ảnh sang Word.

Sau khi bạn chọn tập tin cần số hóa, ABBYY FineReader Professional 11 sẽ tự động phân tích và kết xuất tài liệu sang kiểu định dạng tập tin đầu ra mà bạn đã chọn. Bạn không cần phải thực hiện thao tác mở ứng dụng (chẳng hạn Microsoft Word), mà ABBYY FineReader Professional 11, sau khi phân tích, nhận dạng tài liệu, sẽ tự động mở tài liệu đã số hóa bằng ứng dụng mà bạn đã chọn trước đó.

ABBYY FineReader Professional 11 cho phép bạn tự xác định các vùng hình ảnh, bảng, chữ của tài liệu cần số hóa, nhằm giúp chương trình nhận dạng tài liệu chính xác hơn. Các thao tác thực hiện khá đơn giản và trực quan: Chọn công cụ và kéo thả trên đối tượng (ảnh, bảng, chữ) cần xác định. Ngoài ra, ABBYY FineReader Professional 11 còn hỗ trợ người dùng kiểm tra các lỗi nhận dạng mà chương trình nghi ngờ, bạn nhấn chọn Verification.

Cửa sổ kiểm tra, sửa lỗi nhận dạng được thiết kế trực quan: Phía trên là nội dung tài liệu gốc, phía dưới là nội dung tài liệu đã nhận dạng. Bạn có thể sửa lỗi trực tiếp, nhập lại từ bị nhận dạng sai, hay chọn từ mà ABBYY FineReader Professional 11 đề nghị trong ô Suggestions. Nhấn Confirm hay Replace để sửa, nhấn Ignore để bỏ qua.



Hình 3: Cửa sổ kiểm tra, sửa lỗi nhận dạng được thiết kế trực quan.

Kết quả nhận dạng tài liệu tiếng Việt của ABBYY FineReader Professional 11 khá tốt. Tài liệu sau khi nhận dạng có ít lỗi về nhận dạng chữ. Bố cục và định dạng sau khi nhận dạng thể hiện đầy đủ các bảng, chữ in đậm, chữ hoa, chữ thường, chia cột, ảnh, chú thích ảnh, chữ chân trang như tài liệu đầu vào. Điều này, giúp người dùng tiết kiệm nhiều thời gian và công sức trong việc số hóa tài liệu tiếng Việt.

Hiện nay, nhu cầu số hóa tài liệu tiếng Việt để lưu trữ và có thể chỉnh sửa trên máy tính khá lớn. Tuy nhiên, việc chọn giải pháp số hóa tài liệu tiếng Việt phù hợp, hiệu quả, đáp ứng tốt yêu cầu sử dụng của người dùng cá nhân không phải là điều dễ dàng. Hy vọng, qua loạt bài viết về các dịch vụ và phần mềm số hóa tài liệu tiếng Việt, bạn đọc sẽ chọn được cho mình giải pháp phù hợp. Bạn có thể xem thêm một số bài viết liên quan đến số hóa tài liệu, dữ liệu tại http://www.pcworld.com.vn/T1223993, http://www.pcworld.com.vn/T1226045.