

*Tôi có thói quen dịch sách từ các file PDF. Điều khó nhiều công cụ hỗ trợ dịch như OmegaT lại không hỗ trợ file có đuôi là PDF. Điều đó, gây mất thời gian và bất tiện vì thế tôi cố gắng tra cứu trên mạng cách chuyển file PDF sang *.doc hoặc *.txt nhưng không hiệu quả. Kết quả google trả về thường là phần mềm Online hoặc phần mềm Windows. Cuối cùng, cũng có cách để giải quyết trên Linux. Có hai cách để thực hiện:*

Cách 1: Tách tất cả các chữ từ file PDFs (bao gồm cả chữ trong hình)

Đây là hướng dẫn sẽ giải thích cách để tách tất cả các chữ từ file PDF bằng cách sử dụng phối hợp Ghostscript và công cụ mã lệnh OCR gọi là tesseract-ocr. Đầu tiên cần chuyển PDF sang file hình ảnh riêng lẻ (TIFF) sau đó chúng ta có thể dùng OCR- quét chúng trở lại. Chúng ta cần Ghostscript để làm điều này. Trước tiên, bạn cần chắc chắn đã cài nó lên hệ thống.

```
sudo apt-get install ghostscript
```

Khi đó, chúng ta có thể dùng Ghostscript để thực sự chuyển PDF hãy sử dụng ps

```
gs -dNOPAUSE -sDEVICE=tiffg4 -r600x600 -dBATCH -sPAPERSIZE=a4  
-sOutputFile=Output_File_Name.tif Name_of_PDF.pdf
```

Chúng ta cần đổi ở đây là tên của file PDF và tên file đầu ra tương ứng với "Name_of_PDF.pdf" và "Output_File_Name.tif" ở dòng lệnh trên.

Khi kết thúc ra có 1 file TIFF có kích thước lớn cái mà sẽ dùng OCR (Optical Character Recognition) để quét. Chúng ta sử dụng "tesseract-ocr". Nhưng đầu tiên chúng ta cần cài.

```
sudo apt-get install tesseract-ocr tesseract-ocr-eng
```

Gói "tesseract-ocr-eng" để hỗ trợ ghi nhận ngôn ngữ tiếng anh và yêu cầu cần có để tesseract-ocr có thể làm việc. Để làm việc với ngôn ngữ khác bạn có thể thay thế ví dụ "tesseract-ocr-deu" hỗ trợ tiếng Đức.

Kết thúc, là chuyển file TIFF sang file TXT bao gồm tất cả các chữ, thông thường cả hình ảnh trong file gốc PDF

```
tesseract Output_File_Name.tif Name_of_TXT -l eng
```

Ở đây, "Output_File_Name.tif" là tên file gốc mà bạn đặt ở trên và Name_of_TXT là tên đầu ra của file có đuôi là *.txt. Nếu PDF không phải là tiếng anh, thì đặt giá trị "-l eng" bằng giá trị khác tương ứng với ngôn ngữ chuẩn trong file pdf được chương trình hỗ trợ.

Chú ý: Chất lượng của chữ được tác ra từ hình ảnh trong PDF được tốt hay không phụ thuộc vào nguồn gốc của hình ảnh trong PDF.

Cách 2: Chiết dữ liệu từ file PDF:

Nhược điểm của cách trên là cả chữ trong file PDF cũng được tách ra. Bạn có thể dùng công cụ dưới đây để tách riêng chúng.

Đầu tiên là cài đặt công cụ

Trong Ubuntu

```
sudo apt-get install poppler-utils
```

Trong Fedora:

```
sudo yum install poppler-utils
```

Với những bản phân phối khác, có thể tìm công cụ poppler-utils trong kho phần mềm tương ứng.

Từ dòng lệnh, để tách tất cả hình ảnh từ file "pdffile.pdf" và đặt nó vào đường dẫn

/home/<username>/pdfimages/ thì dùng dòng lệnh:

```
pdfimages -j pdffile.pdf ~/pdfimages/
```

File JPEG sẽ được lưu trữ cùng phần mở rộng PPM trừ phi bạn dùng thông số đặc biệt “-j”

Để tách tất cả các chữ thực sự và đặt vào file có tên giống với tên của file PDF bạn có thể dùng dòng lệnh dưới đây:

```
pdftotext pdffile.pdf
```