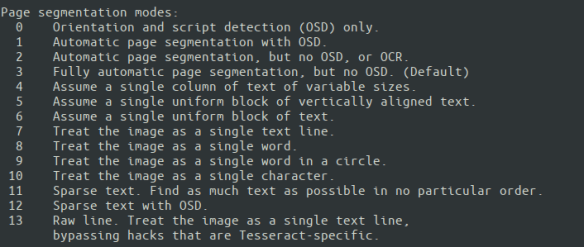
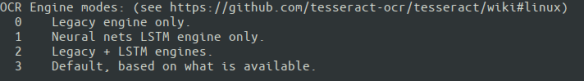
1. Trích xuất văn bản có trong hình ảnh bằng Tesseract-OCR ( tài liệu: [Trang chủ · UB-Mannheim/tesseract Wiki · GitHub](https://github.com/UB-Mannheim/tesseract/wiki)) các bước thực hiện là:

* Cài đặt thư viện bằng câu lệnh pip install pytesseract.
* Khai báo thư viện
* Import dữ liệu hình ảnh
* Câu lệnh chính để thực thi là

custom\_oem\_psm\_config = r'--oem 3 --psm 6'

pytesseract.image\_to\_string(image, config= , lang = ‘ ')

* Với tham số lang là ngôn ngữ đích(danh sách hơn 70 ngôn ngữ: [Languages/Scripts supported in different versions of Tesseract | tessdoc](https://tesseract-ocr.github.io/tessdoc/Data-Files-in-different-versions.html))
* Số lượng tham số có thể điều chỉnh gồm 2 tham số(psm và ocr) trong config
* PSM gồm
* OCR gồm
* Tài liệu : [Tesseract OCR best practices - Ivan's Software Engineering BlogIvan's Software Engineering Blog](https://ai-facets.org/tesseract-ocr-best-practices/)

1. Lưu lại văn bản mới trích xuất ra.
2. Từ văn bản mới trích xuất ta thu được thông tin có cấu trúc như: tên công ty, địa chỉ, ngày bản, sô hóa đơn, tổng cộng…
3. Lọc dữ liệu cần thiết trong văn bản có cấu trúc bằng SpaCy (tài liệu: [spacy · PyPI](https://pypi.org/project/spacy/)) các bước thực hiện là:

* Chuyển đổi văn bản thành dữ liệu với mỗi hàng tương ứng với một dòng của thanh ghi.
* Phân tích cấu trúc của một câu điển hình
* Trích thông tin cần thiết.