Tesseract로 파이썬 OCR 구현하기

참고 링크 : https://www.youdad.kr/implementing-python-ocr-with-tesseract/

Tesseract OCR 활용 심화(정확도 향상 관련) : https://yunwoong.tistory.com/72

(내용 요약해서 노션에 올린거라 안봐도 무관함)
티스토리 정리해둔것 : https://cherish-days02.tistory.com/330 (PS : 43Nzc1Mz)

Tesseract -> 광학 문자 인식 엔진 (프리웨어)

Google이 개발을 후원하고, Linux, Windows 및 MAC OS 에서도 사용 가능함

많은 언어 및 스크립트에 대한 모델들이 추가되어 총 116개 언어 제공중

설치 과정은 맨 위 참고 링크에 나와 있으므로 생략함

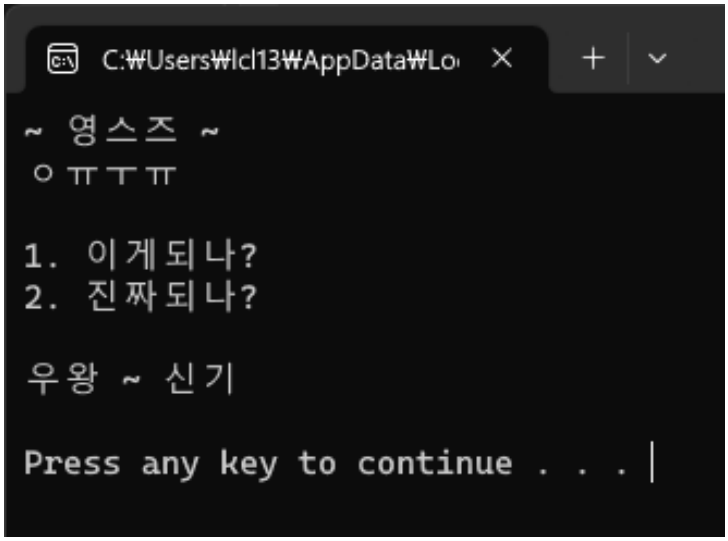
~ 코드 부분 ~

참고 영상 : https://www.youtube.com/watch?v=L8q-KCbXybc

```
pip install pytesseract
pip install Pillow
```

```
//app.py
import pytesseract
from PIL import Image

//경로는 각자 수정하기
pytesseract.pytesseract.tesseract_cmd = r'E:\Tesseract-OCR\tesseract.exe'

file_path = "D:\abc.png"

a = Image.open('D:\abc.png')

result = pytesseract.image_to_string(a, lang='kor')
print(result)
```

참고 영상에서는 데이터셋 있으면 텐서플로우로 정확도 올릴 수 있다고 함.. 이건 첨부터 구현해야하는거라 좀 생각해봐야 할듯…



Tesseract가 인식한 내용

~ 영수증 ~

1. 이게되나?

2. 진짜되나?

우왕 ~ 신기

원본 내용

문제점

- 경로에 파일이 있어야 제대로 열리는 것 같음
  ex) visual studio의 경우 repos 안의 해당 프로젝트 폴더 내에 파일이 존재해야 열림
- 손글씨가 아닌 컴퓨터 폰트임에도 불구하고 낮은 정확도
  (여러 폰트나 이미지, 파일 등으로 정확도를 학습시킬 필요가 있음)
- 파일명에 한글이 들어갔을 경우 읽어오기 불가능
  (이건 utf-8 인코딩 추가해주면 해결될 것으로 보임)

장점

- 기존에 진행했던 주제와는 다르게 성과를 직접 확인 가능해보임
  (정확도 향상, 프로그램으로 작동하다보니 결과와 오류 발생시 왜 발생했는지를 확인 가능)
- 굳이 말하자면 인공지능 분야인만큼 물어볼 만한 사람이 많을듯.. GPT한테라도..
- 확장성이 좋아 보임. 지금 당장은 문자인식만 했지만 여기에 가명처리를 추가한다던가 기타 프로그램에 부수적으로 넣던가 할 때의 확장성

생각해 볼 것들

- 개인정보 마스킹이나 가명정보 처리 과정을 어떻게 처리할건지, 어떤 시나리오를 예상할 수 있을지
- PNG 뿐만 아니라 PDF 파일도 OCR 처리가 가능함 → 이 장점을 최대한 살리기
- 프로그램 처리 과정
  ex : 신분증 처리 프로그램의 경우
  (프로그램 A) OCR 인식 결과를 암호화하여 프로그램 B로 전송
  (프로그램 B) A로부터 받은 결과를 복호화하여 (혹은 복호화하지 않고도) DB와 결과를 대조하여 결과를 프로그램 C로 전송
  (프로그램 C) 대조 성공/실패 결과를 출력
- **단순 OCR 인식만 할 경우 신분증 진위여부는 어떻게 구별해 낼 건지(사실 이게 가장 핵심이라고 생각됨…** 진위여부 구별을 넣을건지 아님 가명처리만 할건지도 생각해봐야될 듯?**)**
  근데 가명처리만 하면 뭔가… 뽀대가 안 산다..

읽어볼 링크

Selvy OCR의 신분증 인식 관련 홍보 자료
https://ocr.selvasai.com/idcard/idcard_intro.html

useB.의 OCR 프로세스 관련 홍보 자료
https://useb.co.kr/solution/ocr

신분증 진위확인에 대한 OCR 솔루션 관련 자료
https://blog.naver.com/inzisoft/222356106818
→ 프로세스 같은 내용 제안서 작성할 때 참고하면 좋을듯

tilkoblet의 신분증 이미지 텍스트 추출(OCR API) 관련 자료
https://tilko.net/Help/Api/POST-api-apiVersion-Tilko-OCR-License

이미지 합성을 이용하여 맞춤형 OCR 엔진 만들기(꼭 읽어볼것)
https://medium.com/@aimap.marker/이미지-합성을-이용해-맞춤형-ocr-엔진-만들

기-632602e59571
→ 19년도 글이긴 한데 작성자 이메일 주소도 나와있어서 모르는 거 생기면 연락해볼 수 있을것 같음
→ 우리가 진행할 직접적인 PBL 내용일 것 같아서 강추..ㅎ

컴트루 테크놀로지 블로그 글 : http://www.comtrue.com/comtrue/ai_08/
→ 제안서 작성 시 해당 기술의 필요성에 대해 강조할 때 참고하기 좋을 듯

딥러닝을 활용한 한글문장 OCR 프로젝트(이것도 읽어보는것 추천)
https://medium.com/@sunwoopark/딥러닝을-활용한-한글문장-ocr-프로젝트-hclt-2019-bb9d17622412