PDF, DLP, FONT 삼각 관계 과제

BoB10 디지털포렌식 이예진

1) 내용 요약

DLP는 TEXT기반 탐지, 차단을 하는 기능이다. 이때, PDF 파일 중 개인정보가 담겨 있는 경우 이 파일을 외부 저장장치로 옮길 시 이를 탐지가 필요하다. 하지만 이때 Font의 별로 OTF와 TTF로 속성에 따라 해당 글씨가 TEXT로 탐지되기도 하고 IMAGE로 탐지되기도 한다.

이때 해결방안으로 기존 폰트만 사용하고, 새로 설치한 속성값을 변경한다. 그리고 이미 프로세싱이 가능한 DLP로 바꾼다가 있다.

2) 직접 테스트 결과

- 주요 기능

① 이동식 드라이브(Removable Disk)감지 : get_drive()

win32api와 win32file 패키지를 이용하여 드라이브의 속성을 확인하고 그 속성이 이동식 드라이브, 즉 Removable Disk이면 리스트에 그 경로를 저장하여 리턴해준다.

② 파일 및 디렉토리의 이동, 생성, 수정, 삭제 행위 감시 : watchdog

파이썬의 패키지 중 watchdog을 이용하여 파일과 디렉토리가 이동, 생성, 수정, 삭제와 같은 이벤트가 발생하면 이를 탐지하였다. 해당 이벤트가 발생할 때, 파일의 확장자가 pdf일 경우 개인정보를 포함하였는지 확인하였다.

③ pdf파일 내 개인정보 유무 확인 및 경고 메시지 생성 : pdf_extract()

앞서 watchdog에서 발견한 pdf파일의 경로를 받아 해당 pdf를 pdf2image를 이용하여 모두 이미지로 변경하였다. 그 후 해당 이미지에서 OCR을 이용하여 모든 텍스트를 추출하는 과정을 거친다. 텍스트 중 주민등록번호가 있는지 확인하기위해 아래와 같은 정규표현식을 사용해 주었다.

```
\d{2}([0]\d|[1][0-2])([0][1-9]|[1-2]\d|[3][0-1])[-]*[1-4]\d{6}
```

₩d{2} : 맨앞 정수 2자리(생년)은 어떤 정수값이 와도 무관하다.

[0]₩d|[1][0-2] : 첫자리가 0인 경우, 뒤에 어떤 정수가 와도 괜찮다. 첫자리가 1인 경우 뒷자리는 0,1,2만 올 수 있다.

([0][1-9]|[1-2]₩d|[3][0-1]) : 생일은 첫자리가 0이면 뒷자리가 0이 될 경우 0일이 되기 때문에 0 다음에는 1~9만 올 수 있다.

[-]* : -가 0개 또는 1개 이다.

[1-4] : 주민번호 뒷자리 첫번째 숫자는 1~4만 갖는다

₩d{6} : 주민번호 첫자리를 제외한 숫자는 총 6자리이다.

해당 위치에서 감지된 주민번호가 있을 경우 경고 메시지를 출력한다.

- 테스트 결과

3) 대응 방안

- 해당 폰트의 속성을 파이썬 스크립트를 이용해서 수정해준다.
- Pytesseract의 특성상 이미지 해상도에 따라 오탐과 위탐이 발생할 수 있으므로 머신러닝을 통해 최적의 dpi를 설정해 준다.