

2025년 2학기 제10회 정보보안SW융합전공 프로젝트 발표회

다중 문서 포맷을 지원하는 개인정보 비식별화 도구

De-Identification tool supporting multiple document formats

CONTENTS

01 문제제기

02 플로우차트

03 모듈설명

04 시연영상

05 기대효과

06 향후발전 및 문제점

팀원소개



박윤지
DOC 모듈
XLS 모듈



백서진
WEB
NER
HWP 모듈



정보남
XML 모듈
OCR 모듈
PDF 모듈



함근희
WEB
PPT 모듈

01. 문제제기 문제 인식

- 실무 문서에는 성명, 주민등록번호, 연락처 등 다양한 개인정보가 포함됨
- 관리 부실 또는 보안 사고 발생 시 대규모 개인정보 유출로 직결
- ➔ 개인정보보호법 제3조, 제17조에 따른 개인정보의 적절한 보호·처리 조치가 필수

개인정보위, 1.6TB 소송문서 유출 법무법인 '로고스'에 과징금 5억 철회

발행일 : 2025-11-21 11:33

< □ 가



보안뉴스 25.11.21

01. 문제제기

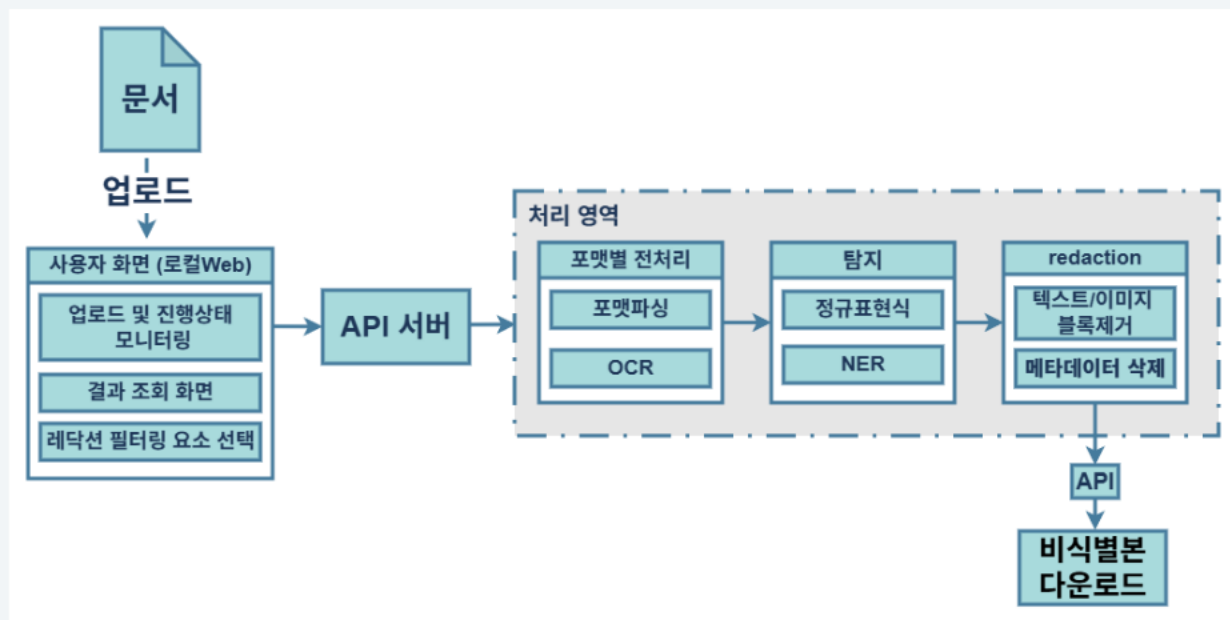
- 문제제기

- 기존 비식별화 도구는 특정 문서 포맷 또는 영어권 사용자 중심으로 제공되는 경우가 많음
- OLE, OOXML, PDF 등 다양한 문서 포맷을 하나의 절차로 일괄 처리할 수 있는 공개된 비식별화 기술은 부족함

- 프로젝트 목표

- 문서 포맷별 내부 구조(OLE, OOXML, PDF)를 분석하여, 다중 문서 포맷을 통합적으로 처리할 수 있는 비식별화 시스템을 설계
- 레이아웃을 유지하면서 재현 가능한 비식별화 파이프라인을 오픈소스 형태로 제시

02. 플로우차트 제안 기법 플로우 (Flow chart)



1. 로컬 웹에 문서 업로드
2. 포맷 판별
3. 텍스트 추출
4. 정규화 & 오프셋 매핑
5. 개인정보 탐지 (정규식 + 유효성검증 + NER)
6. 레이아웃 보존 기반 마스킹
7. 비식별 문서 출력

03. 모듈 설명

포맷별 추출 방식

- XML 문서

OOXML/HWPX는 XML 구조를 기반으로 텍스트를 정규화하여 추출

포맷	핵심파일	설명
.DOCX	word/document.xml, word/charts/*, word/embeddings/*.xlsx, word/media/*	<w:t> 치환 + 차트(<a:t>/<c:v>) 치환, 임베디드 XLSX 별도 처리, media는 이미지 OCR 가능
.PPTX	ppt/slides/*, ppt/charts/*, ppt/embeddings/*.xlsx	슬라이드 <a:t> 치환 + 차트 텍스트 치환, 임베디드 XLSX 처리 (notesSlides 미처리)
.XLSX	xl/sharedStrings.xml, xl/worksheets/*, xl/charts/*	SST <t> + 시트 <v>/<t> + 차트 텍스트 치환 (숨김시 트 제외 없음)
.HWPX	Contents/*, Charts/*, Images/*, BinData/*, Preview/*	본문/차트 치환, Preview 제거, Images/BinData(이미 지) OCR 가능, BinData(PK) 임베디드 XLSX 처리

03. 모듈 설명

포맷별 추출 방식

- OLE 문서

OLE 기반 문서는 포맷별 내부 스트림 구조에 따라 개별 추출 로직 적용

포맷	주요스트림/ 구조	텍스트 추출 방식 핵심
.DOC	WordDocument, 0/1Table, Fib, PlcPcd	Character Position(CP) 기반 텍스트 조각 복원
.PPT	PowerPoint Document, TextCharsAtom / TextBytesAtom	슬라이드 단위 텍스트 Atom 순회 후 Unicode 복원
.XLS	Workbook, SST(Shared String Table), CONTINUE records	SharedStringTable 참조, 분할 문자열 병합
.HWP	BodyText, DocInfo	문단 블록 단위 압축 해제 및 텍스트 디코딩

03. 모듈 설명

OCR 작동방식

- OCR 처리 흐름
 - 문서 내 이미지 또는 텍스트 레이어를 판별하여 OCR 적용 여부 결정
 - 페이지 또는 포함 이미지를 OCR 분석을 위한 이미지 형태로 변환
 - EasyOCR로 이미지 내 텍스트 블록(문자열, 좌표, 신뢰도) 추출
 - 추출된 블록은 Qwen2.5-VL 기반 LLM을 사용해 후처리
 - OCR로 얻은 텍스트+좌표를 기반으로 민감정보를 판별 후 bbox 위치에 바로 마스킹을 적용해 레딕션 처리

전번 01012341234
전화번호 010-2128-4331
이메일: genuheerrr@naver.com
카드: 6556 2082 1728 1454
운전면허번호 11-23-237875-80
집번 023132345



전번 [REDACTED]
전화번호 [REDACTED]
이메일: [REDACTED]
카드: [REDACTED]
운전면허번호 [REDACTED]
집번 [REDACTED]

03. 모듈 설명 정규식 & 유효성 검증

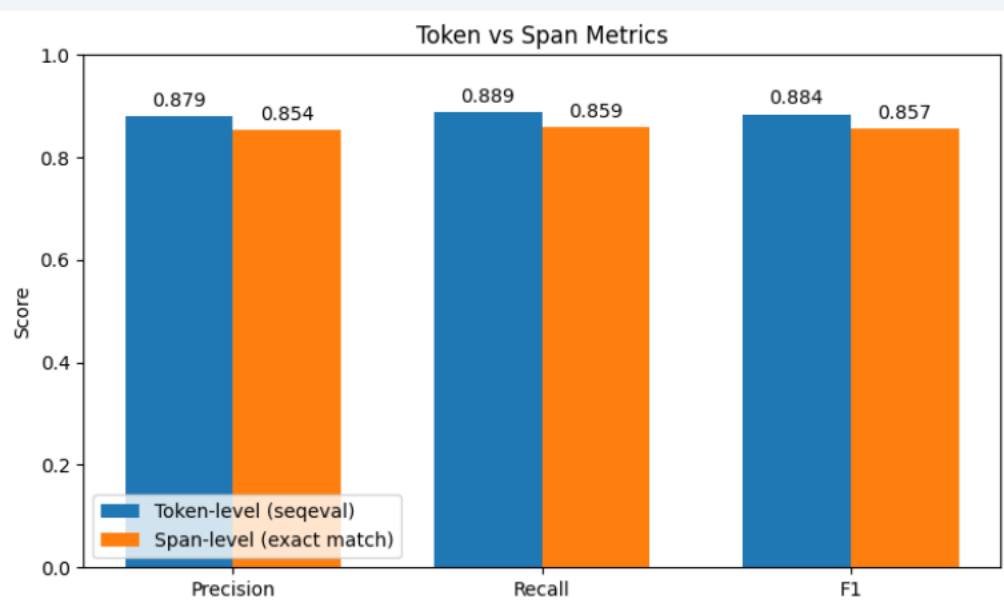
- 정규식 탐지 종류

- 주민등록번호 (신/구버전 지원)
- 외국인 번호 (신/구버전 지원)
- 운전면허번호 (신/구버전 지원)
- 여권번호
- 전화번호 (010 및 지역번호(031 등)지원)
- 이메일 주소
- 카드번호

- 유효성 검증

- 주민등록번호 - 체크섬, 생년월일 유효성, 등록 구분 코드 검사 (1/2/3/4)
- 외국인 번호 - 외국인 번호용 체크섬, 생년월일 유효성, 등록 구분 코드 검사 (5/6)
- 운전면허번호 - 발급년도 범위 검증 (1960~)
- 전화번호 - 길이 및 시작 규칙 확인 & 하이픈 규칙 확인
- 이메일 주소 - 허용 문자 & 도메인 형식
- 카드번호 - IIN 규칙 + Luhn 알고리즘

03. 모듈 설명 NER 탐지방식



- KLUE-RoBERTa 기반 문맥 탐지

- 탐지 대상: 이름(PS) / 지명·주소(LC) / 기관(OG)
- 방법: KLUE-BERT 토큰 분류 + BIO 시퀀스 라벨링 → 엔터티 경계 복원
- 흐름: 정규화 → 토큰나이징 → 예측 → 엔터티 합치기
- 토큰 기준(seqeval): 정밀도 0.879 · 재현율 0.889 · F1 0.884
- 엔터티(스팬) 기준: 정밀도 0.854 · 재현율 0.859 · F1 0.857
- 라벨별 F1(토큰 기준): PS 0.922 · LC 0.805 · OG 0.831

03. 모듈 설명 비식별화 처리

- 일반문서 (XML/ OLE)
 - 문서 파싱 단계에서 텍스트 + 원본 오프셋 (위치정보) 확보
 - 정규화된 문자열 내 탐지 위치를 원본 문서 좌표로 역매핑
 - 탐지 문자열을 동일 길이 '*'로 치환
 - 표/줄바꿈/스타일 등 문서 레이아웃 유지
- PDF 문서
 - PyMuPDF 기반 텍스트 블록 추출 및 레dak션
 - 텍스트 레이어 기반 PDF 우선 지원
 - 각 텍스트 조각의 BBox(x,y,w,h) 영역 확보
 - 해당 영역을  형태로 마스킹
 - 텍스트 숨김이 아니라 실제 삭제/대체

03. 모듈 설명 웹UI/UX

문서뷰어

연락처 정보

유대전화

010-4728-3395

REGEX:전화번호

이메일

jiwoohan@naver.com

REGEX:이메일

주소 정보

자택

서울특별시 용산구 이태원로 55길 12

NER:LC

직장

서울특별시 중구 세종대로 110

NER:LC

금융 정보

신용카드번호

6556-2082-1728-1454

REGEX:카드번호

계좌번호

하나은행

NER:OG

123-456789-01011

신분증 정보

운전면허번호

11-23-204027-80

REGEX:운전면허번호

여권번호

M87654321

REGEX:여권번호

소속 기관

근무처

LG전자(주)

NER:OG

부서

T

NER:OG

개발팀

직급

책임연구원

작성일자

2024년 12월 28일

작성자

한지우

NER:PS

15

매칭 결과

정규식

총 8 · OK 8 · FAIL 0

All OK FAIL

텍스트 검색

전화번호 3

phone_mobile

010-4728-3395

OK

phone_city

02-516-1717

OK

phone_city

02-984-5124

OK

이메일 1

email

jiwoohan@naver.com

OK

카드번호 1

card

6556-2082-1728-1454

OK

운전면허번호 1

drivers_license

11-23-204027-80

OK

평가 리포트

리포트 다운로드 JSON

인감정보 비율

40.69%

166자 / 408자

중 항목 (Unique)

15

Rate: 15 | Overlap: 0%

정규식: 0 NER: 0

정규식 검증

OK 8

FAIL 0

FAIL 원인

-

valid=false 항목만 집계

NER 평균 Score

0.59

선택된 라벨만 반영

정책

정규식 규칙

주민등록번호

외국인등록번호

이메일

전화번호

전화번호

카드번호

여권번호

운전면허번호

NER 라벨

PS - 2

LC - 2

OG - 3

처리 시간

Step	Time
extract	13ms
match	7ms
ner	546ms
redact	0ms
total	593ms

유형별 건수

Type	Count
전화번호	3
주민등록번호	1
이메일	1
카드번호	1
운전면허번호	1
여권번호	1

라벨별 건수

Label	Count
OG	3
PS	2
LC	2

04. 시연

05. 검증

- 데이터셋: 총 10
- 포맷: HWP(X),
- 모듈: 정규표현

구분
하이픈 (미)포함
NER
기타

매칭 결과 NER			
총 40 · 평균 score 0.61			
Label	Text	Score	Pos
PS	함근희	0.65	0-3
OG	진관초등학교	0.62	48-54
OG	신도중학교	0.62	55-60
OG	신도고등학교	0.61	61-67
LC	서울특별시 은평구 북한산로2 1026동 1302호	0.52	225-252
PS	박윤지	0.57	253-256
OG	은평초등학교	0.61	300-306
OG	신사중학교	0.52	307-312
OG	선정고등	0.46	313-317
OG	중부대학교	0.60	320-325

4개, 기타 10개)
테스트

형태 예시
4-5678, 01012345678
선정고등학교 → 선정고등
메일 , 여권번호

06. 성과 요약

내부망 일괄 처리로 리스크·비용 축소

다중 포맷을 한 파이프라인에서
로컬 처리
(업로드→정규화→탐지→치환)
하므로 외부 이전 없이
일괄 비식별화·다운스트림
자동화가 가능하다.

다중 포맷 일괄 처리 및 레이아웃 보존

일반 문서는 원본 오프셋 역매핑
+ 동일 길이 치환,
PDF는 bbox 마스킹으로
표/줄바꿈 등 레이아웃을 유지해
검토·공유 단계의
재작업을 줄인다.

정확도·신뢰성

- 실사용 기반 NER 및 OCR 오탐/미탐 피드백
- 재학습/평가셋 확장
- 배포

3단계의 반복으로 정확도와 신뢰성
지속 개선.

07. 기대효과 및 향후 발전

기대효과

- 도입 비용 부담 최소화
- 개인정보 처리 리스크 감소
- 개인정보 보호 규정 대응 역량 강화

문제점

- 일부 포맷에서는 차트 미지원
- hwp파일의 경우, 명세 설명 부족으로 제대로 된 차트 지원이 아님
- OCR 모듈이 이미지 속 특정 부분을 민감 정보로 인식 불가

향후 연구

- 차트 완벽 지원할 수 있도록 수정
- 텍스트 박스 부분 레딕션 추가
- OCR 모듈 성능 업그레이드

THANK YOU
감사합니다.