

Data Privacy를 위한 데이터 익명화와 비식별화

IDEA LAB SEMINAR

이관호

Contents

- **문제상황**
 - 개인정보 침해 사례들
- **해결방안**
 - K-익명화
 - 비식별화
- **Method**

문제 상황

- **LLM에 개인의 사생활 정보가 유출되는 사건**
 - 24년 3월 20일, ChatGPT 유료 이용자들의 개인정보가 유출되는 사건이 발생
 - 한국인 687명의 개인정보(이름, 이메일, 신용카드번호)가 다른 이용자에게 노출됨
- **실시간 데이터 처리 시스템에서 발생할 수 있는 개인정보 유출**
 - 온라인 주식 거래 시스템
 - 수만 건의 주식 주문, 가격변동, 거래 내역이 실시간으로 스트리밍 되고 있음
→ 거래자의 패턴 추정, 고액 투자자 식별, 위치 기반 추적 가능
 - 환자 모니터링 시스템
 - 환자들의 혈압, 심박수, 혈당 등 민감한 생체 정보가 실시간으로 병원 서버로 전송되고 있음
→ 전송 중 해킹 당할 시 환자들의 질병, 위치, 상태가 외부에 노출됨

해결 방안

- 데이터 익명화

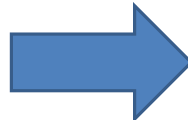
- K-익명화

- 공격자가 나를 특정할 수 없게 나를 포함한 최소 k명의 사람 중 하나로 보이게 하는 것 → 소분류를 대분류로 바꾸는 작업
 - 예시: 원본 데이터: 성북구 사는 사람, 종로구 사는 사람, 관악구 사는 사람
익명화 데이터: 서울 사는 사람, 서울 사는 사람, 서울 사는 사람 → 3-익명화

- 데이터 비식별화

- 가명처리
 - 데이터 삭제
 - 데이터 마스킹

저는 노량진에서 거주 중인
32세 홍길동입니다. 거주 중인
빌라 앞에 차량번호 23가
5555차량이 입구를 막았어요.



저는 동작구에서 거주 중인
40대 홍**입니다. 거주 중인
빌라 앞에 차량번호
*****차량이 입구를 막았어요.

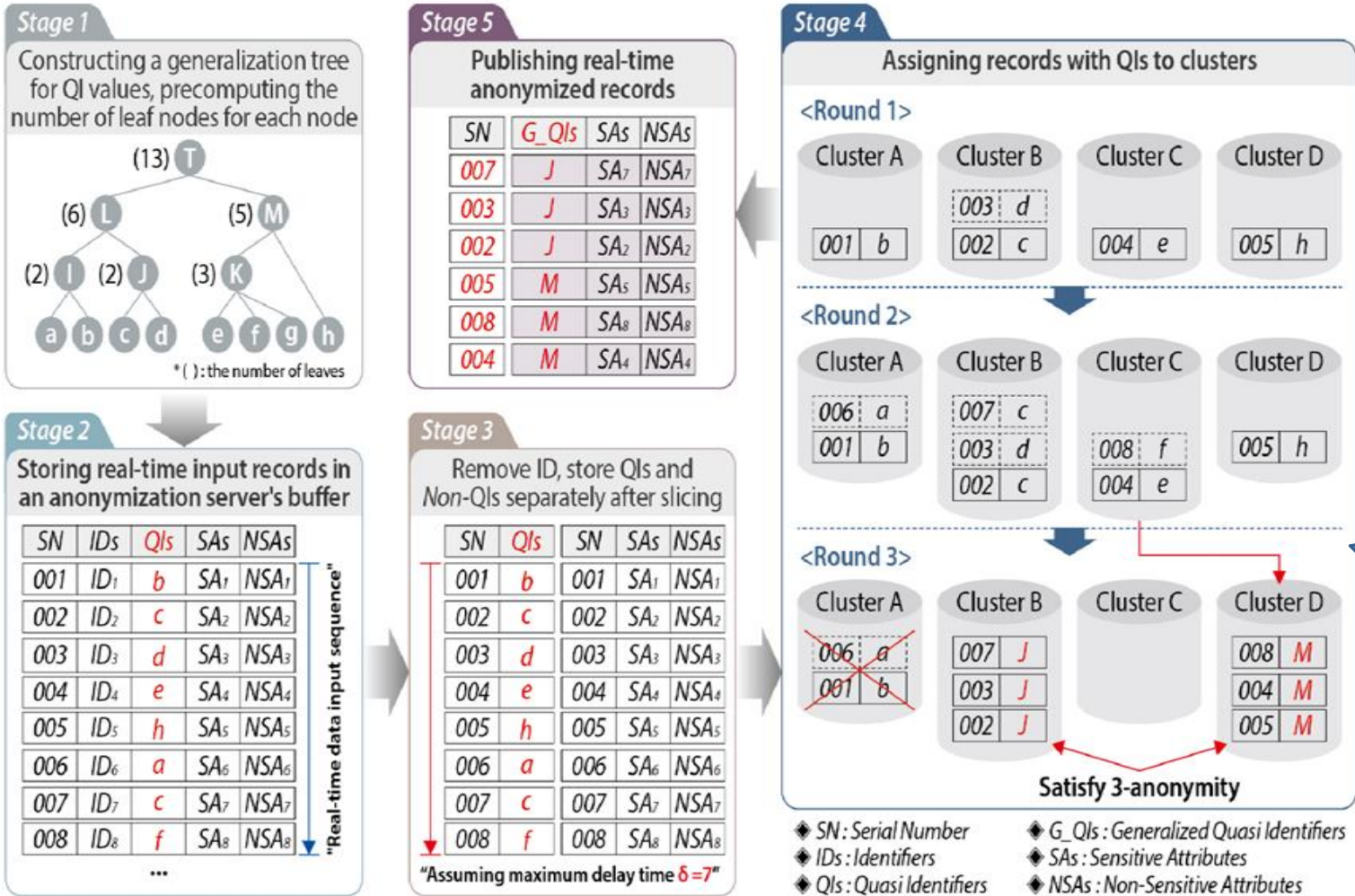
Methods

- 실시간 데이터 익명화 알고리즘
- 텍스트 개인정보 탐지 및 비식별화 기술
- 멀티모달 개인정보 탐지 및 비식별화 기술

익명화 알고리즘 - SUHDSA

• 알고리즘 예시

- GOAL: 트리를 이용해 덜 일반화된 말로 데이터를 바꾸는 작업(10살 → 10대 → 청소년)



기본적으로 각 클러스터에 하나씩 배정

정보 손실이 적은 클러스터로 병합

1. 크기 $\geq k \rightarrow$ 익명화
2. 크기 $< k \rightarrow$ 삭제

익명화 알고리즘 - SUHDSA

- 필요한 Measure

- 1. 클러스터의 정보 손실(Information Loss)에 대한 measure

i: Cluster에 속한 데이터의 row, j: j번째 column

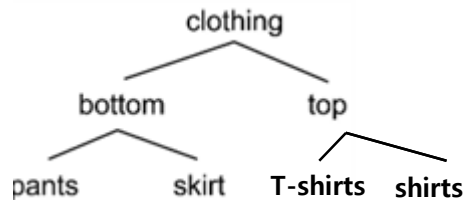
→ R_{ij} : i번째 데이터의 j번째 컬럼의 값, R'_j : j 번째 컬럼 값들의 공통 조상(대분류) D_j : j 번째 컬럼의 유일값 개수

→ **Nbrleaves(a)**: a의 잎의 개수

직관적 해석: 더 일반화(뭉뚱그려 말할)수록 정보의 손실이 커짐

$$IL(C) = \sum_{i \in C} \sum_{j=1}^n \frac{\text{nbrleaves}(R'_j) - \text{nbrleaves}(R_{ij})}{|D_j|}$$

- 예시: Cluster C1이 [skirts, pants]을 가지고 있는 경우, C2 는 [pants, T-shirts]을 갖고 있는 경우(i=2, j=1)



$$IL(C1) = (2-0)/2 + (2-0)/2 = 4$$

$$IL(C2) = (7-0)/2 + (7-0)/2 = 7$$

- 2. 실시간으로 들어오는 데이터와 클러스터의 거리 measure

$$CAILDDistance(R, C) = IL(C \cup \{R\}) + (IL(C \cup \{R\}) - IL(C)) \times \log_2(|C|)$$

텍스트 데이터 개인정보 탐지 및 비식별화 기술

- 전체 개요:

- 문장 내에 개인정보에 관한 단어를 인식 → 인식한 개인 정보 비식별화

인식 단계

A : 계약금을 먼저 보내셔야 다른 분이 방을 안 보죠.

A : 계약금을 먼저 보내셔야 다른 분이 방을 안 보죠.

A : 우리 4499-391-879344로 보내주시고, 올 수 있는 날짜 말씀하세요.

A : {우리 4499-391-879344: QT_ACCOUNT_NUMBER}로 보내주시고, 올 수 있는 날짜 말씀하세요.

B : 제가 지금 진주 살고 있어서 바로 가기는 어려워요.

B : 제가 지금 {진주: LC_ADDRESS} 살고 있어서 바로 가기는 어려워요.

A : 이 방은 옥수역이랑 진짜 3분 거리에서 다들 보겠다는 방이에요.

A : 이 방은 {옥수역: LC_PLACE}이랑 진짜 3분 거리에서 다들 보겠다는 방이에요.

B : 오늘 오후에는 꼭 계약금 보내드릴게요.

B : 오늘 오후에는 꼭 계약금 보내드릴게요.

A : 그럼 보내고 나서 010@14삼2삼12로 문자 하시든지 아니면 해피홈 부동산 1522-2383로 전화하세요.

A : 그럼 보내고 나서 {010@14삼2삼12: QT_MOBILE}로 문자 하시든지 아니면 {해피홈부동산: OG_WORKPLACE} {1522-2383: QT_PHONE}로 전화하세요.

비식별화
마스킹, 가명화

텍스트 데이터 개인정보 탐지 및 비식별화 기술

- 개인 정보 단어 탐지 방법

- NER(Named Entity Recognition)

- 텍스트 내에서 인명, 지명, 기관명과 같은 특정한 정보를 가진 단어나 구를 식별하고 분류하는 Task
 - 예시1 : 올해 [삼성전자|OG_COMPANY]에 입사했습니다. [정보항목_태그]
 - 예시2 : 저는 [동작구|LC_ADRESS]에 있는 [A회사|OR_COMPANY]에 다닙니다.

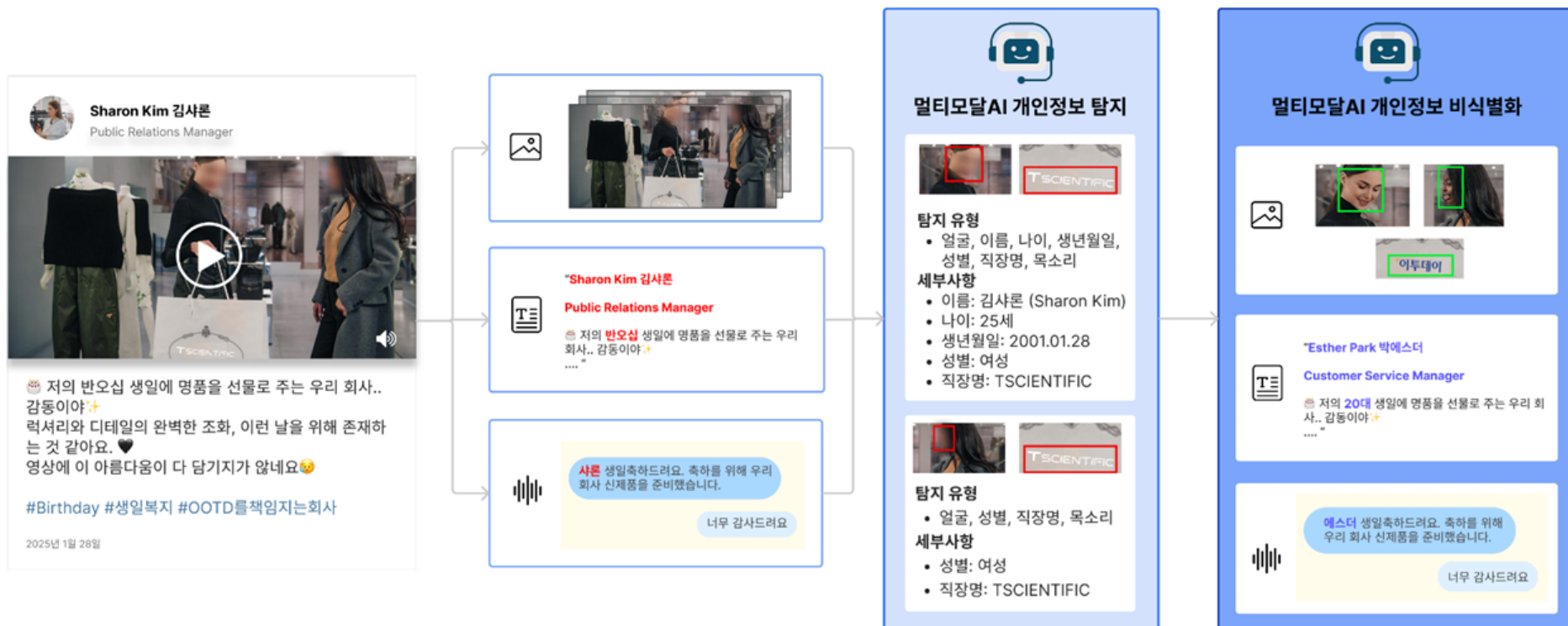
- 이런 NER Task에 특화된 언어 모델을 Fine-Tuning하여 사용

- 33개 개인정보에 대해 태그되어있는 데이터셋을 만들고 이를 학습 시킴
 - 33개 개인정보 예시: 닉네임, 운전면허, 차량번호, 금융정보, 회사 부서, 직책 등등
 - 예시1: [우리 4499-391-879344|QT_ACCOUNT_NUMBER] 로 8000원 보내주세요.
 - 예시2: 저는 [김민수|PS_NAME] 이고 [남자|CV_SEX] 입니다

멀티모달데이터 개인정보 탐지 및 비식별화 기술

• 전체 개요:

- 사진, 비디오, 오디오에서 Text 추출 → 개인 정보 단어 탐지 → 비식별화



*얼굴은 단순히 얼굴 인식 모델을 사용한 후 Blur처리 하는 것으로 보여짐



Thank You!

NER TASK에서의 학습 데이터

- BIO 태깅을 통한 데이터셋 만들기

Token	Tag
연락처는	O
010	B-QT_PHONE
4321	I-QT_PHONE
9988	I-QT_PHONE
이메일은	O
sujin.kim+dev@example.com	B-QT_EMAIL
계좌는	O
1002-334-991234	B-QT_ACCOUNT_NUMBER
에요	O
.	O