

# データ前処理: ③プライバシー保護

- 学習前コーパスから個人を特定できる情報 (Personally Identifiable Information) を除去する必要がある
- 直接的には、キーワードスポッティングのようなルールベースの手法を採用することで効果的に名前/住所/電話番号等を除去可能
- プライバシー攻撃におけるLLMの脆弱性は、事前学習コーパスに重複したPIIデータが存在することに起因している
  - ➔ 重複排除によりリスクをある程度軽減できる

# データ前処理: ④トークン化

テキストを個々のトークン (最小の構成単位) に分ける処理

分け方は単語ベースや文字ベースなどがある

- バイト・ペア・エンコーディング (BPE) トークン化

文字ごとに分割 → 文字ペアの出現頻度が高いものから結合リストに追加 → 結合リストの出現頻度の高い文字ペアから順に結合してトークナイズ

- WordPiece トークン化

文字ごとに分割 → 文字ペアの頻度が高く個々のパーツの出現頻度が低いものから結合して辞書に追加 → 先頭の文字から順に辞書にある最長のサブワードにより分割してトークナイズ

- ユニグラム トークン化

文字ごとに分割 → 出現した全文字パターンを辞書に追加してUnigram言語モデルで計算し、失った時の損失が低い語彙を削除 → 全文字パターンのうちトークンの出現確率が高い語彙の組み合わせを選択してトークナイズ