データ前処理: ②重複排除

重複データは言語モデルの多様性を低下させ、性能に影響を与える

そのため、コーパスの重複を、文レベル/文書レベル/データセットレベルなど様々な粒度で除去する必要がある

- 文レベル

  繰り返しの単語やフレーズを含む低品質な文は、言語モデルで反復パターンを引き起こす可能性があるため除去する

- 文書レベル

  単語やn-gramの重複率に基づいて類似した内容の文書を削除する

- データセットレベル

  データセットが汚染されないよう、学習セットと評価セットの重複を防ぐ

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