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

藤井巧朗, 柴田幸輝, 山口篤季, 十河泰弘 「日本語Tokenizerの違いは下流タスク性能に影響を与えるか？」『言語処理学会 第29回年次大会 発表論文集』 2023: 1480-1485.

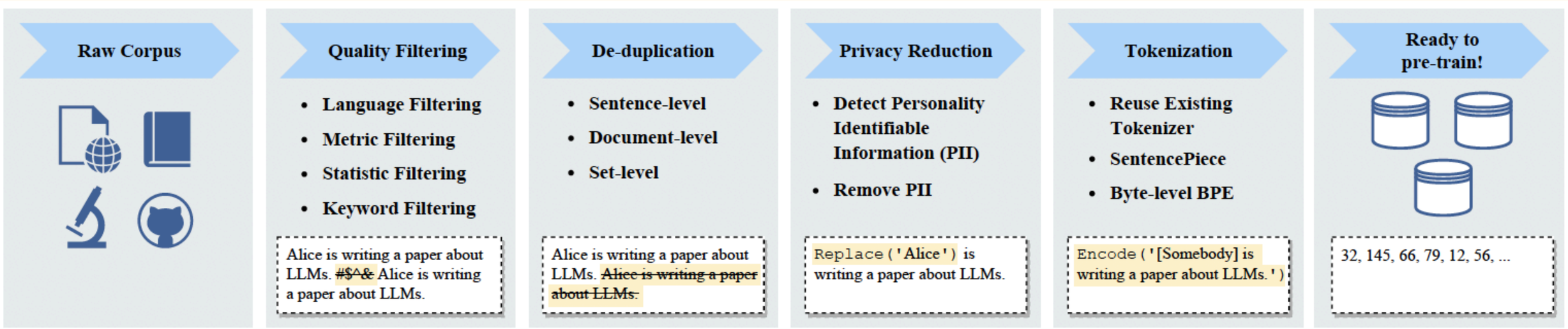

Fig. 6: An illustration of a typical data preprocessing pipeline for pre-training large language models.