データ前処理: ①品質フィルタリング

機械学習ベースで品質を測定する方法と、ルールベースの方法がある
前者は口語なども削除されてコーパスの多様性が低下する可能性がある

- 言語ベース

  学習に不要な言語のテキストをフィルタリングする

- メトリックベース

  生成されたテキストに関する評価指標 (perplexity: テキストの流暢さ など) に基づいて不自然な文章を検出する

- 統計ベース

  句読点や文長などの統計的特徴を利用して低品質なデータをフィルタリングする

- キーワードベース

  HTMLタグや定型文、不快な言葉などのノイズや不要な要素を特定して削除する

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