24COP509: Deduplication

Module leader: Dr Georgina Cosma, Department of Computer Science, Loughborough University
g.cosma@lboro.ac.uk

Schedule

- Research and Preparation (1 hour)
- Group Presentations (20 minutes per group)
- Final Wrap-up (5-10 minutes)

Group Research Questions

Group 1: Foundational Problem Analysis

Your task involves explaining why duplicate data presents a fundamental challenge in language model development.

1. What particular challenges does duplicate data present for language model training?
2. Analyse the paper's key example of the 61-word repeated sequence. What does this reveal about current data collection practices?
3. In what ways does duplicate data affect model evaluation and benchmarking?
4. What are the broader implications of train-test overlap for model development?

Group 2: Technical Implementation - Exact Matching

Your task requires explaining the technical approach to identifying exact duplicates in large datasets.

1. How does the suffix array methodology identify duplicate content?
2. For what reasons did the researchers select a 50-token threshold for matching?
3. Which computational challenges arise when implementing exact substring matching at scale?
4. How does this method address the trade-off between precision and computational efficiency?

Group 3: Advanced Deduplication Methods

Your task involves explaining the approximate matching approach and its advantages.

1. What advantages does approximate matching offer compared to exact substring matching?
2. How does the MinHash algorithm estimate document similarity?
3. For what reasons is the Jaccard Index particularly suitable for this application?
4. Which technical considerations influenced the selection of parameters for approximate matching?

Group 4: Performance Analysis

Your task requires analysing the empirical results of deduplication on model performance.

1. How do the researchers quantify the impact of deduplication on model performance?
2. What evidence suggests that deduplication improves model quality?
3. In what ways do different deduplication strategies affect training efficiency?
4. Which metrics are most relevant for evaluating the success of deduplication?

Group 5: Research Implications

Your task involves analysing the broader implications for AI development.

1. How might these findings influence future approaches to dataset curation?
2. What are the scalability implications for larger language models?
3. How do these results relate to broader questions of model reliability and evaluation?
4. Which research questions remain unaddressed in this domain?

Presentation Requirements

Each group should prepare a 10-minute presentation comprising:

- A clear explanation of the assigned topic
- Technical analysis supported by evidence from the paper
- Visual aids (graphs, diagrams, or key figures from the paper)
- Discussion of implications for AI development
- Connections to broader themes in machine learning