Topic 1: Pretraining Data

1. FineWeb: Hugging Face's 15T-token web text dataset built from Common Crawl using advanced filtering and deduplication: https://arxiv.org/abs/2406.17557, https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1
2. RedPajama: Open 1.2T-token dataset replicating Meta's LLaMA data recipe across diverse sources: https://www.together.ai/blog/redpajama
3. RedPajama-Data v2: 30T-token dataset with multi-language coverage and document-level quality metrics: https://www.together.ai/blog/redpajama-data-v2
4. Dolma: AI2's 3T-token English corpus emphasizing transparency and diverse data mix: https://allenai.org/blog/dolma-3-trillion-tokens-open-llm-corpus-9a0ff4b8da64
5. Dolma: https://huggingface.co/papers/2402.00159
6. Dolma paper: https://arxiv.org/pdf/2402.00159
7. RefinedWeb: Falcon's web dataset (600B tokens) built with aggressive filtering and deduplication: https://arxiv.org/abs/2306.01116
8. C4: Google's 750 GB "clean" web dataset powering T5, filtered from Common Crawl: https://arxiv.org/abs/1910.10683
9. A Pretrainer's Guide to Training Data: https://arxiv.org/abs/2305.13169
10. Creating, Curating, and Cleaning Data for LLMs: Talk on large-scale data collection, filtering, and deduplication workflows: https://www.youtube.com/watch?v=HEGaei7k0zE
11. Common Crawl: https://commoncrawl.org/
12. Tensorflow C4: https://www.tensorflow.org/datasets/catalog/c4
13. HuggingFace C4: https://huggingface.co/datasets/allenai/c4