Analytics Vidhya

Training Data Curation

Instructor

Sourab Mangrulkar

Machine Learning Engineer at

Creator of PEFT

Training Data Curation

Process of collecting and aggregating the training data

Training Data Curation

Mixture of filtered web data and curated high quality corpora

- Wikipedia
- Stackoverflow
- Github
- Social media conversations
- Books
- Technical papers

Principles

- Massive Scale

Prinicple: Massive Scale

300 Billion Tokens

Prinicple: Massive Scale

2 Trillion Tokens

Prinicple: Massive Scale

Falcon LLM

3500 Billion Tokens

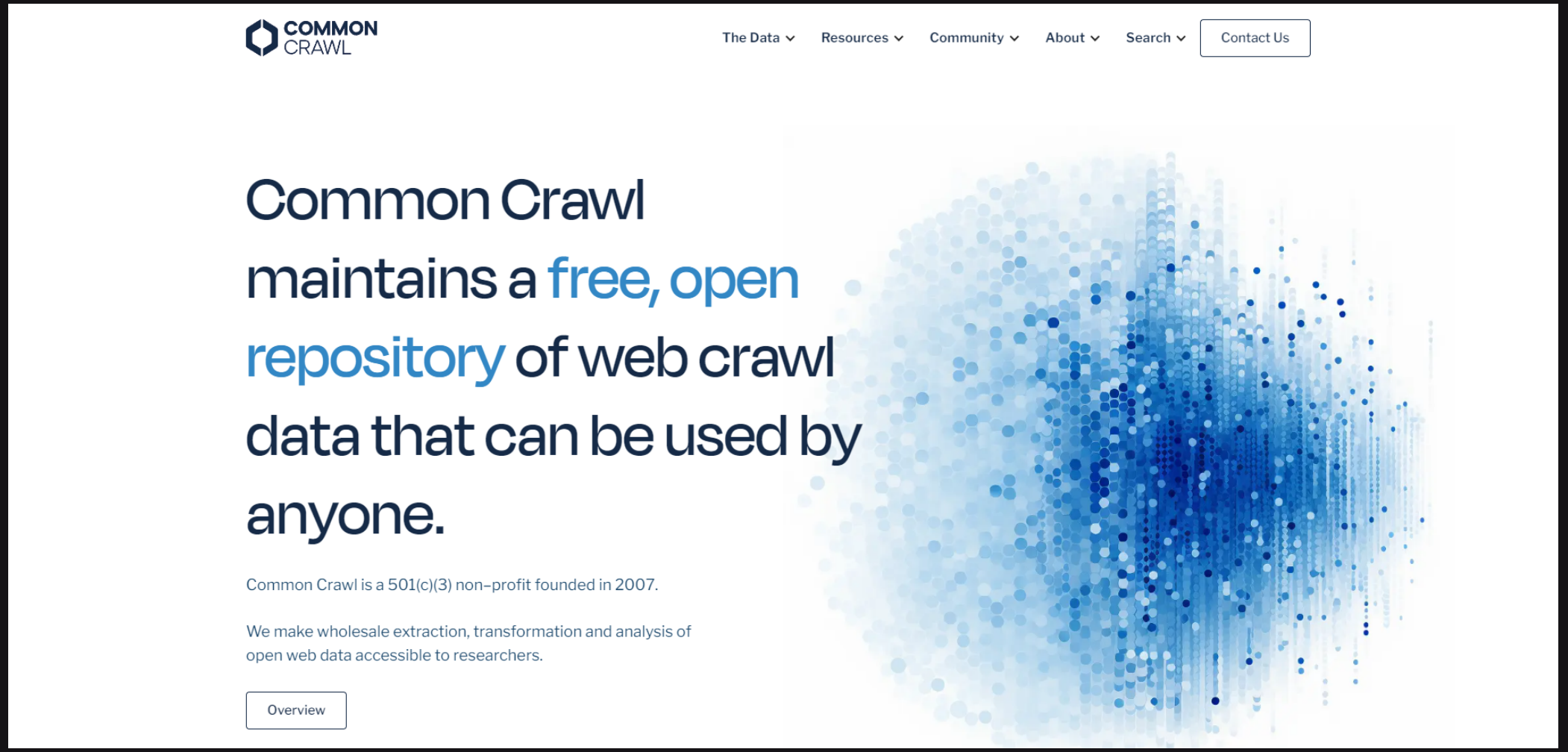
Principles

- Massive Scale
- Diversity
- High Quality

Open Source Datasets

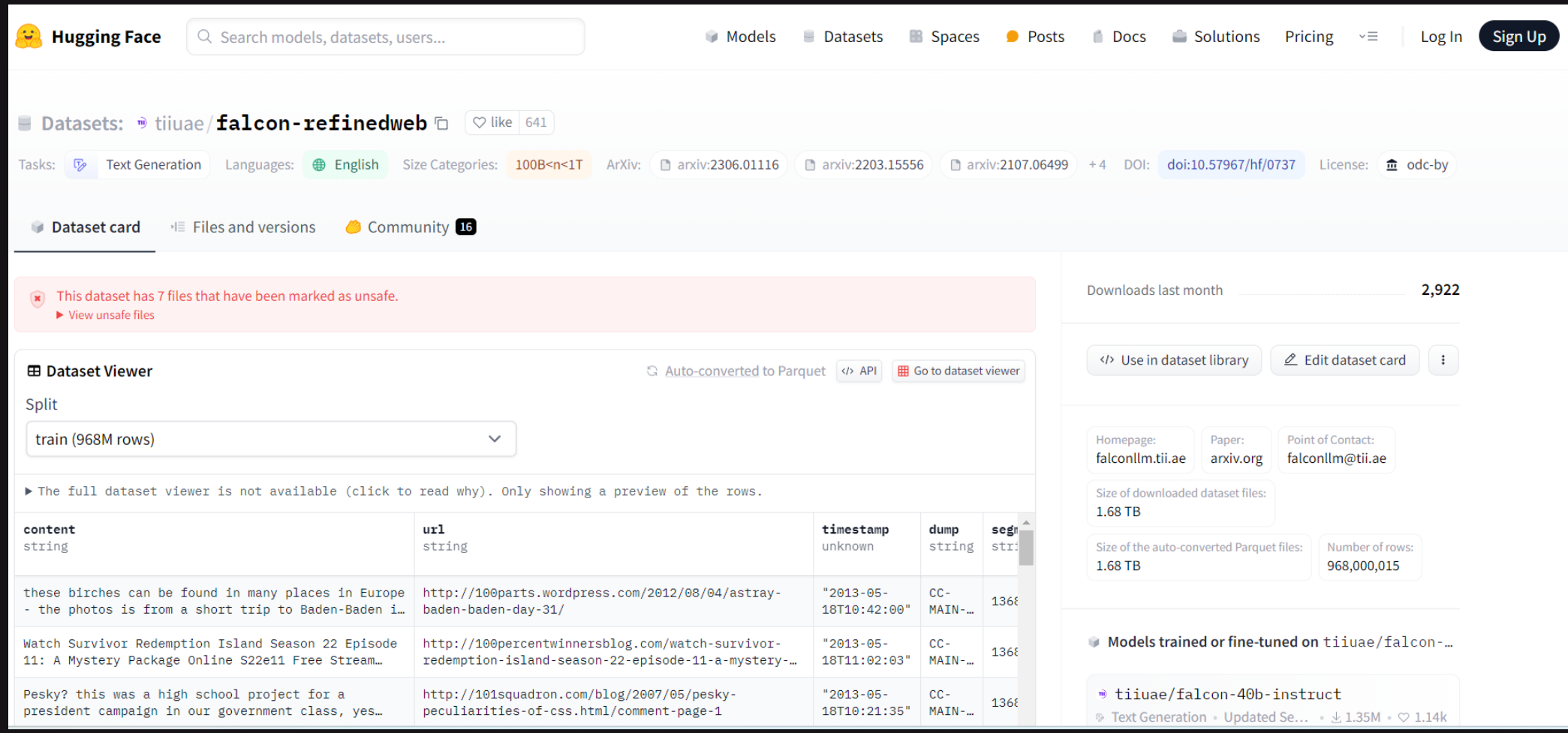
Common Crawl

Link: https://commoncrawl.org/



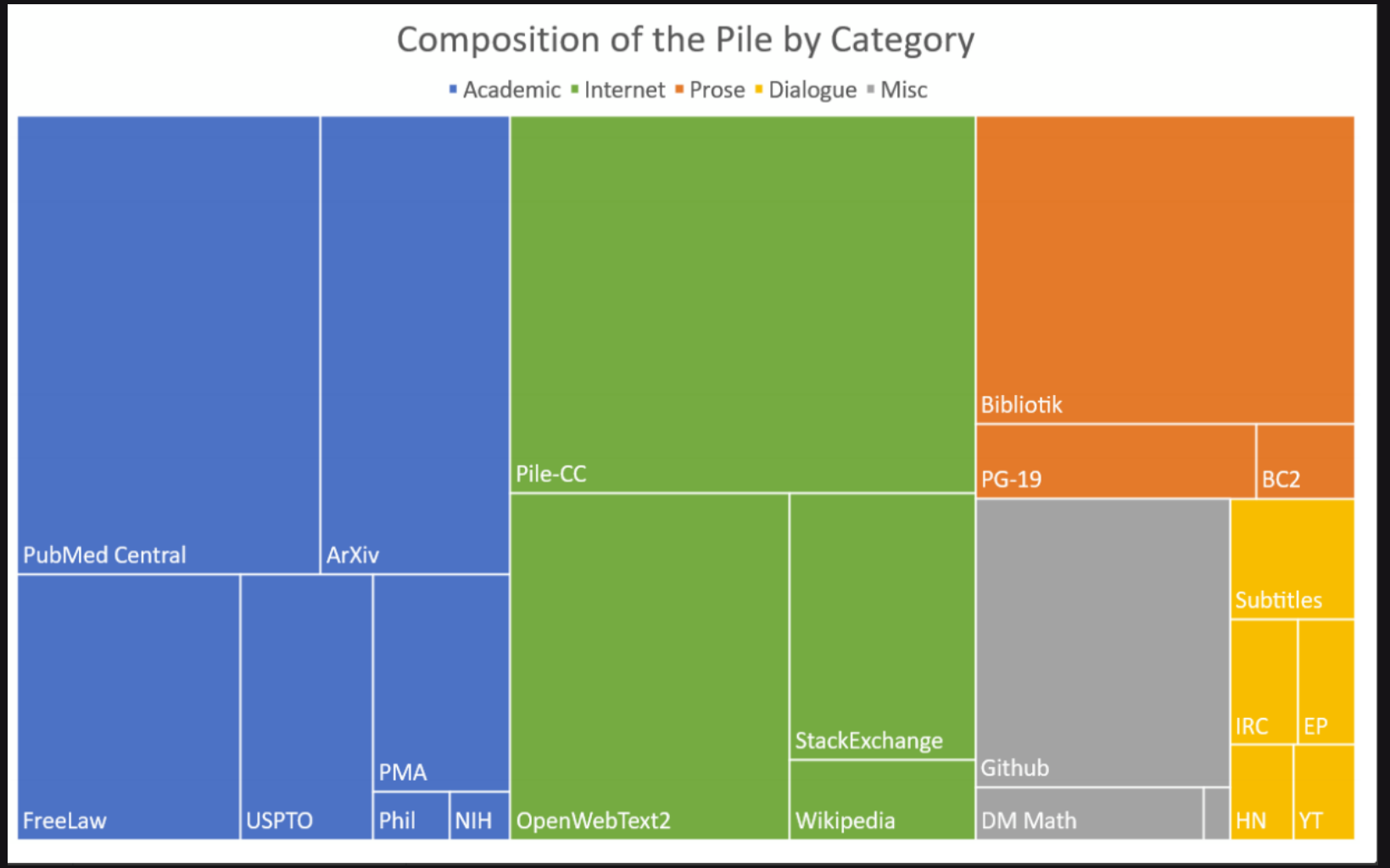
Refined Web Dataset

Link: https://huggingface.co/datasets/tiiuae/falcon-refinedweb

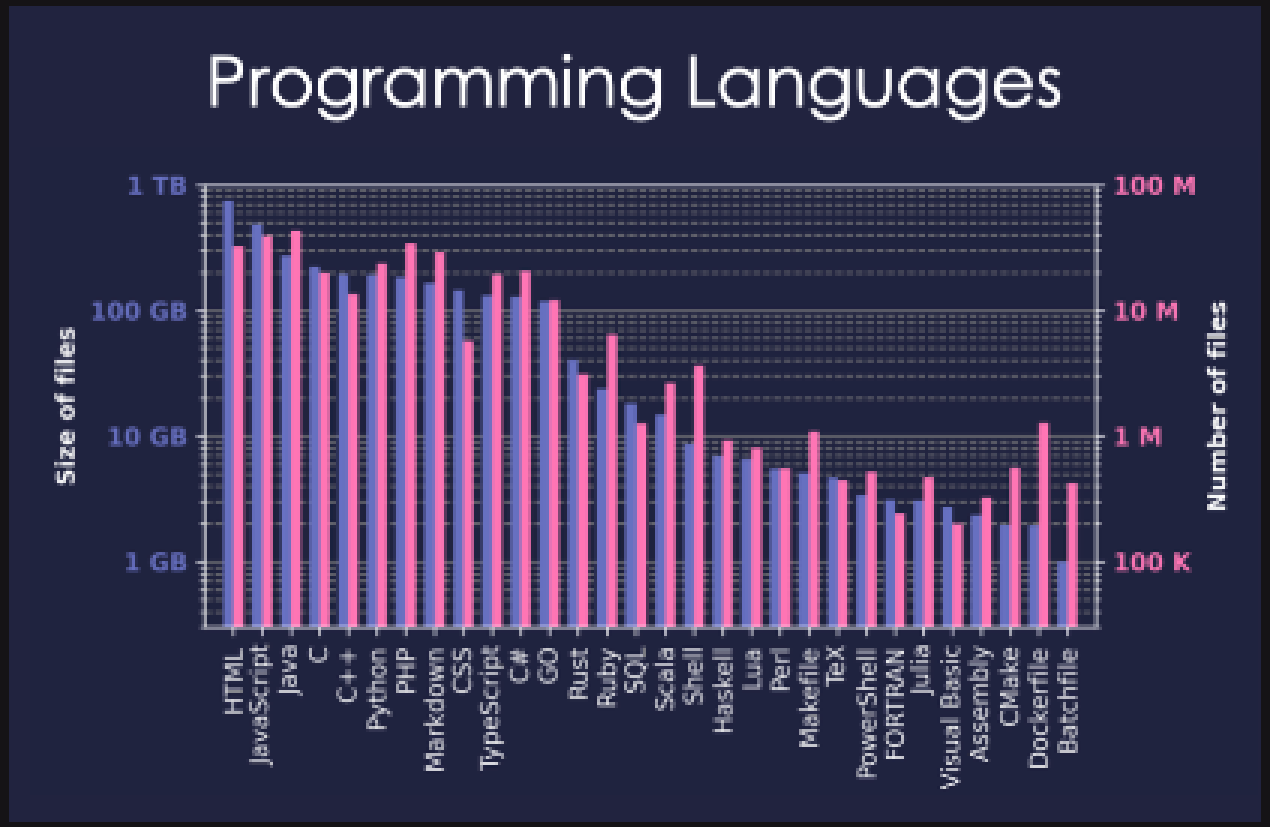


Analytics Vidhya

Pile



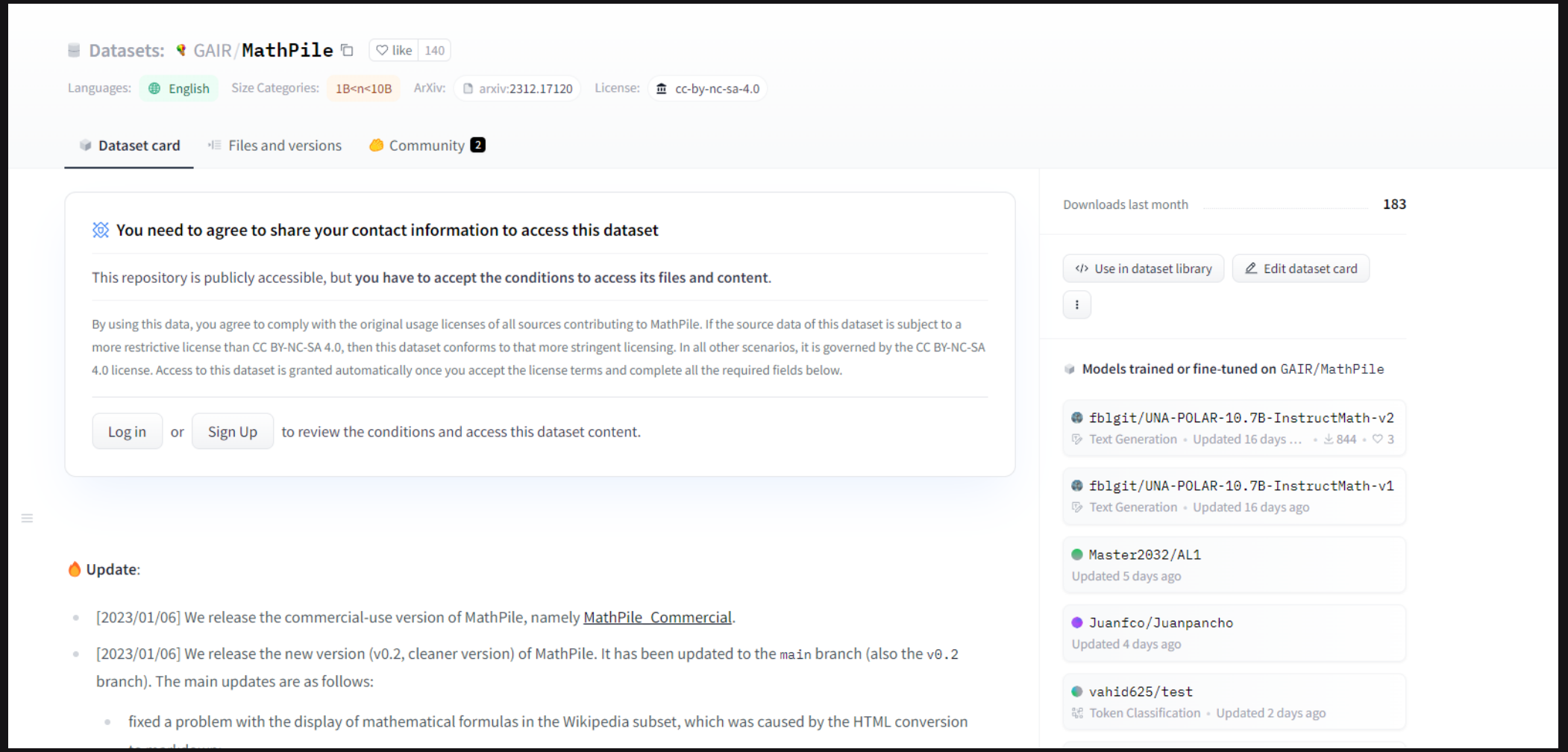
The Stack



Analytics Vidhya

Math Pile

Link: https://huggingface.co/datasets/GAIR/MathPile



Analytics Vidhya

Steps involved in Training Data Curation

- Estimate training data size using scaling laws
- Focus on high-quality training data

Thank You

Analytics Vidhya