LREC-COLING 2024

CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages

Thuat Nguyen[1], Chien Van Nguyen[1], Viet Dac Lai[1], Hieu Man[1], Nghia Trung Ngo[1], Franck Dernoncourt[2], Ryan A. Rossi[2], Thien Huu Nguyen[1]

[1]University of Oregon, [2]Adobe Research

Adobe

UNIVERSITY OF OREGON

I. Introduction: CulturaX

- Important factors behind LLM developments:
  - Model sizes
  - **High-quality training datasets**
- Lack of transparency for training data in recent SOTA LLMs models
- Lack of open-source, readily usable dataset to effectively train **multilingual LLMs**

UNIVERSITY OF OREGON

I. Introduction: CulturaX

- **Largest** open-source multilingual dataset.
- High-quality **6.3 trillion tokens** for **167 languages**:
  - Meticulously clean and deduplicate
  - Tailored for Large Language Model (LLM) development
- **Fully released** to the public
  - Available on: https://huggingface.co/datasets/uonlp/CulturaX

UNIVERSITY OF OREGON

II. Dataset Creation

- Combination of mC4 and OSCAR datasets
  - Latest iteration of mC4 (version 3.1.0)
  - All available OSCAR corpora: (from 20.19 to 23.01)
- Then the merged data is extensively cleaned and deduplicated at the document level to produce the **highest quality** data.
- Dataset size: **27 TB** from **13.5B documents** with **6.3 trillion tokens**

II. Dataset Creation

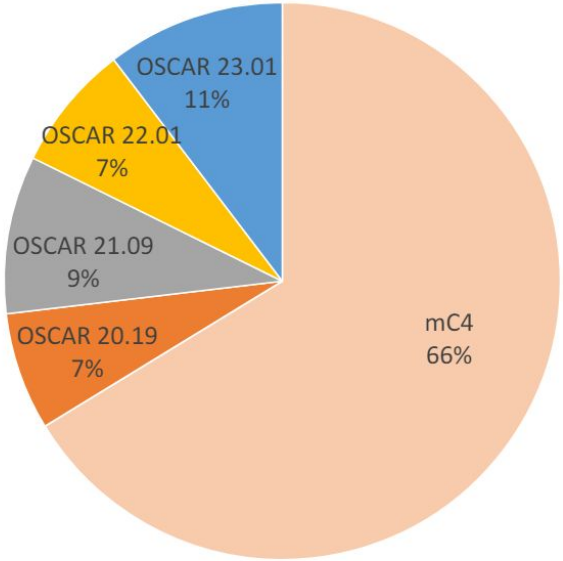

Figure 1: Distribution of document counts from mC4 and OSCAR in our initial dataset.

UNIVERSITY OF OREGON

III. Data Cleaning Pipeline Overview

- Comprehensive pipeline for highest quality data
- Five key steps:

1. Language identification
2. URL-based filtering
3. Metric-based cleaning
4. Document refinement
5. Data deduplication

3.2. Language Identification & URL-based Filtering

Language Identification:

- Re-predict languages for mC4 using **FastText**
- Remove documents with differing predictions

URL-based Filtering:

- Leverage **UT1 blacklist** to eliminate **toxic/harmful sources**

3.3. Metric-based Cleaning

Utilize distributions of dataset metrics to identify and filter outliers

Key metrics used:

- Number of words, character/word repetition ratios, stopword ratio
- Language identification confidence, Perplexity score, document length
- ……

Threshold selection using Interquartile Range (IQR) method

- Compute Q1 and Q3 percentiles for each metric and language
- Select thresholds based on percentiles (Q1=10, Q3=90)

3.4. Document Refinement

Clean retained documents to improve quality

Remove noisy/irrelevant portions:

- Short lines at the end of documents (footer details, unhelpful information)
- Lines containing JavaScript (JS) keywords

Preserve documents with > 2 JS lines (likely coding tutorials)

Require at least two different types of JS keywords to avoid removing helpful content

3.5. Data Deduplication

Two deduplication methods:

1. MinHash deduplication for each language's dataset
    - MinHashLSH method with 5-grams and 0.8 Jaccard similarity threshold
2. URL-based deduplication
    - Remove documents with identical URLs
    - Handle updated articles bypassing near-deduplication

Deduplication performed independently for each language

IV. Statistics & Analysis

Code	Language	#Documents (M)					Filtering Rate (%)	#Tokens	
		Initial	URL Filtering	Metric Filtering	MinHash Dedup	URL Dedup		(B)	(%)
en	English	5783.24	5766.08	3586.85	3308.30	3241.07	43.96	2846.97	45.13
ru	Russian	1431.35	1429.05	922.34	845.64	799.31	44.16	737.20	11.69
es	Spanish	844.48	842.75	530.01	479.65	450.94	46.60	373.85	5.93
de	German	863.18	861.46	515.83	447.06	420.02	51.34	357.03	5.66
fr	French	711.64	709.48	439.69	387.37	363.75	48.89	319.33	5.06
zh	Chinese	444.37	444.03	258.35	222.37	218.62	50.80	227.06	3.60
it	Italian	406.87	406.04	254.72	226.42	211.31	48.06	165.45	2.62
pt	Portuguese	347.47	346.76	217.21	200.11	190.29	45.24	136.94	2.17
pl	Polish	270.12	269.73	170.86	151.71	142.17	47.37	117.27	1.86
ja	Japanese	247.67	247.19	137.88	114.64	111.19	55.11	107.87	1.71
vi	Vietnamese	182.88	182.72	118.67	108.77	102.41	44.00	98.45	1.56
Total (42 languages)		**13397.79**	**13366.17**	**8254.28**	**7471.48**	**7181.40**	**46.40**	**6267.99**	**99.37**
Total (167 languages)		**13506.76**	**13474.94**	**8308.74**	**7521.23**	**7228.91**	**46.48**	**6308.42**	**100.00**

UNIVERSITY OF OREGON

V. Conclusion

- CulturaX: A large-scale, high-quality multilingual dataset with **6.3 trillion tokens** across **167 languages**
- Openly accessible to the public
  - https://huggingface.co/datasets/uonlp/CulturaX
- Potential impact of CulturaX:
  - Promotes transparency and democratization of LLM technology
  - Supports development of LLMs for low-resource languages
  - Advances state-of-the-art in multilingual NLP

UNIVERSITY OF OREGON