Von der Theorie zur Praxis: Datenaufbereitung für das Training großer Sprachmodelle

Hammam Abdelwahab | 30.10.2023
Lennard Helmer | 30.10.2023

Von der Theorie zur Praxis: Datenaufbereitung für Sprachmodelle

Agenda

- Part 01: Theorie und Beispiele
- Part 02: Praxisbeispiel im Rahmen MLOps Open GPT-X

Part 01

Datenaufbereitung für große Sprachmodelle Theorie

Datenaufbereitung für Sprachmodelle

Datenquellen

- In der heutigen Technologielandschaft wird großer Wert auf große Sprachmodelle gelegt.
- Große Sprachmodelle werden in der Regel mit einer Mischung aus:
  - Gecrawlte Webdaten.
  - Kuratierte qualitative hochwertige Korpora. [1]

Datenaufbereitung für Sprachmodelle

Beispiele für große Datensätze

Common Crawl	PubMed Central	Github	Redpajama
OSCAR	DM Math	Starcoder	Wikipedia
C4	Arxiv	FreeLaw	RefinedWeb

Datenaufbereitung für Sprachmodelle

Beispiele für große Datensätze

Common Crawl	PubMed Central	Github	Redpajama
OSCAR	DM Math	Starcoder	Wikipedia
C4	Arxiv	FreeLaw	RefinedWeb

Poor pre-processing leads to poor data quality

Datenaufbereitung für Sprachmodelle

Statistische Analyse von Qualitätskennzahlen

- Unzureichende Vorbereitung führt zu schlechter Qualität.

- Untersuchung der Verteilung einer Qualitätskennzahl. [2]

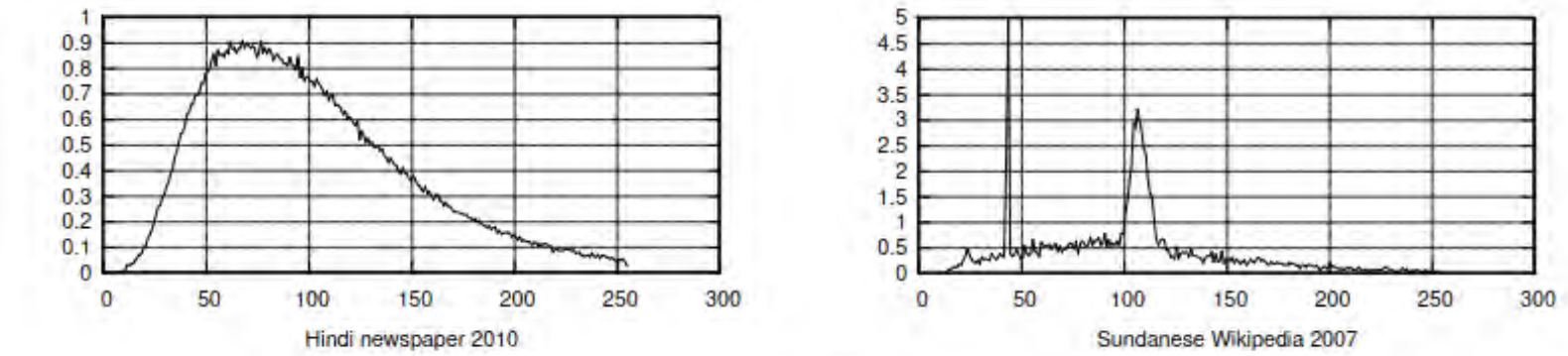


Figure 1: Sentence length distribution for two corpora (percentage for number of characters)

Part 01

Theorie: Beispiele für Datenaufbereitung und Pipelines bei großen Sprachmodelle

Datenaufbereitung für Sprachmodelle

Beispiele für Datenaufbereitung bei großen Sprachmodelle

Model	Year	Processing Steps
Megatron [3]	2019	- Remove documents < 128 characters - Remove new lines - **Deduplication** (Local Sensitivity Hashing)
GPT3 [4]	2019	- Quality Classifier (Logistic Regression) - **Fuzzy deduplication**
BLOOM [5]	2022	- **Deduplication** - Remove web domain < 20 MB
Llama [6]	2023	- CCNET Pipeline (for C4) - Keep largest 28 websites (StackExchange)
Falcon [7]	2023	- **Deduplication** - URL Filtering for adult content removal

Datenaufbereitung für Sprachmodelle

Beispiele für Data Quality Piplines bei großen Sprachmodelle

- CCNET Pipeline – 2017. [8]

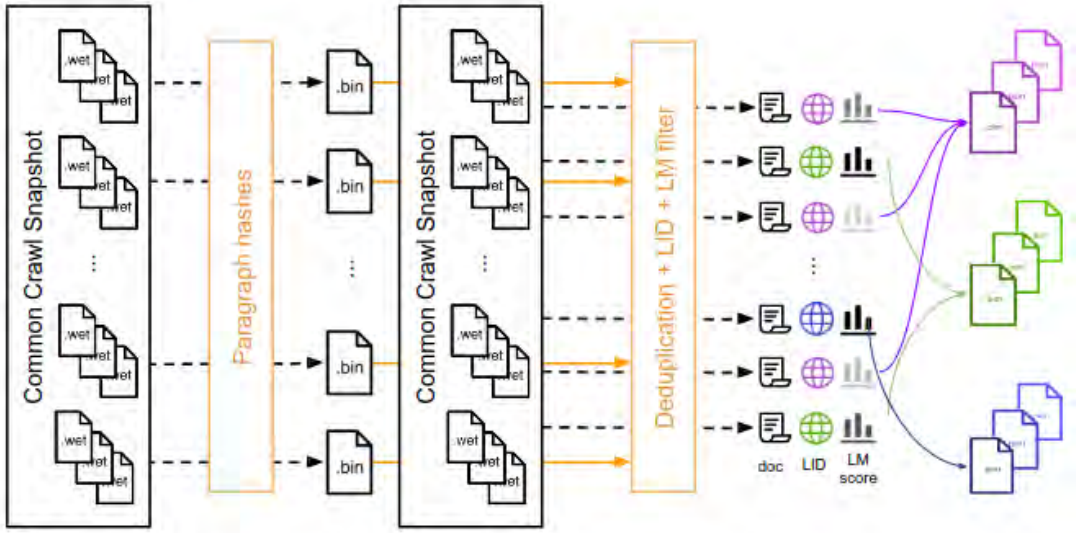


Figure 1: We show the whole pipeline for downloading and processing one snapshot of Common Crawl. First we download all the wet files and compute the paragraph hashes that we group and save into binary files. Then we process every document of the wet files independently: we deduplicate the paragraph using the binary files, we do a language identification and compute language model perplexity score. Finally, we regroup the documents into json files by language and perplexity score. The steps of the pipeline indicated with dashed arrows are parallelisable.

Datenaufbereitung für Sprachmodelle

Beispiele für Data Quality Piplines bei großen Sprachmodelle

- Ungoliant Pipeline – 2020. [9]

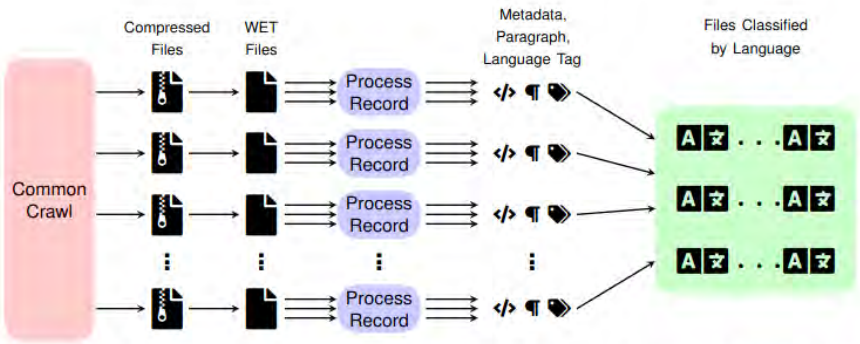


Figure 1: Scheme of the Ungoliant pipeline. The red square represents CommonCrawl content hosting, where the compressed shards are fetched. The *Process Shard* steps hold shard processing, paragraph creation and merging (see Figure 2), and are internally parallelized.

- : CommonCrawl compressed shard.
- : Uncompressed shard, containing records.
- </>: Record Metadata
- : Language identification
- ¶: Paragraph, composed of sentences identified as

Datenaufbereitung für Sprachmodelle

Beispiele für Data Quality Piplines bei großen Sprachmodelle

- Ungoliant Pipeline – 2020. [9]

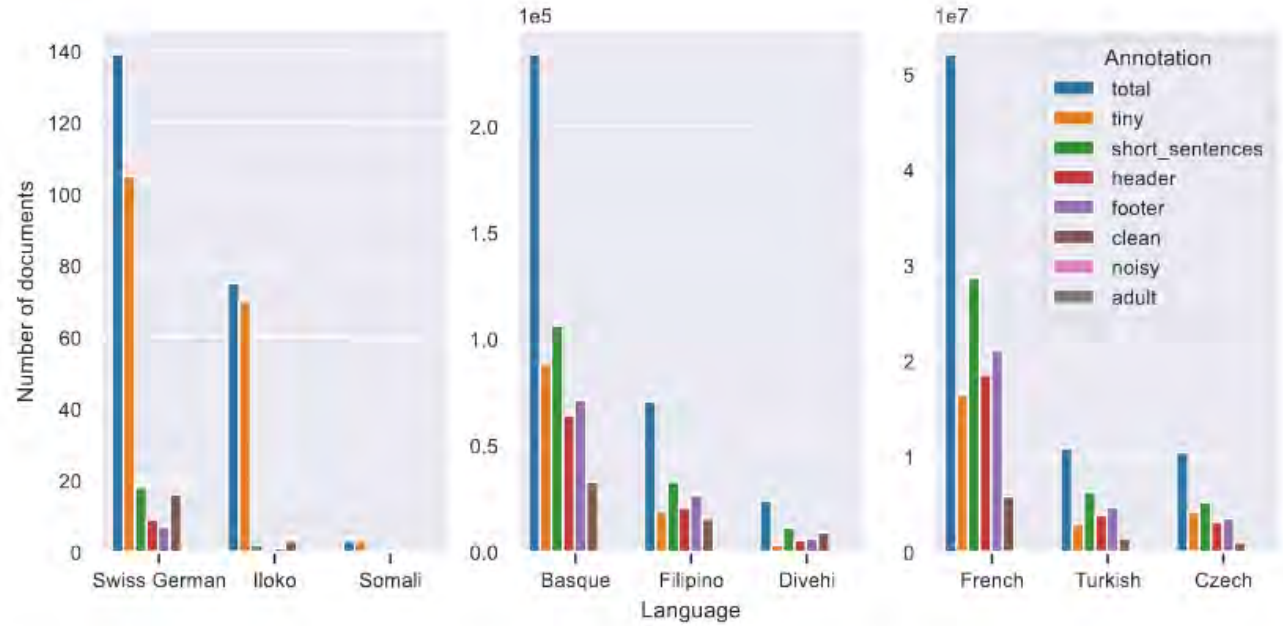


Figure 3: Annotation count in selected low, mid and high resource languages (scales are adapted to corpus size)

Datenaufbereitung für Sprachmodelle

Zusammenfassung

- Große Sprachmodelle basieren meist auf im Internet gecrawlten und kuratierten Daten.
- Statistische Analyse von Qualitätsmetriken für die Bewertung der Datenqualität.
- Die Deduplizierung Ansätze werden regelmäßig eingesetzt, um eine hohe Qualität zu erzielen.
- Open-Source-Datenpipelines, die die Erzeugung hochwertiger Daten aus dem Internet unterstützen
- Bewertung der Datenqualität unter dem Gesichtspunkt der Einhaltung von Vorschriften in Forschungspublikationen

Datenaufbereitung für Sprachmodelle

Referenzen

- [1] Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., ... & Launay, J. (2023). The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

- [2] Eckart, T., Quasthoff, U., & Goldhahn, D. (2012, May). The Influence of Corpus Quality on Statistical Measurements on Language Resources. In *LREC* (pp. 2318-2321).
- [3] Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., & Catanzaro, B. (2019). Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.
- [4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.
- [5] Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... & Manica, M. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Datenaufbereitung für Sprachmodelle

Referenzen

- [6] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

- [7] Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., ... & Launay, J. (2023). The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

- [8] Wenzek, G., Lachaux, M. A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., & Grave, E. (2019). CCNet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.
- [9] Abadji, J., Suárez, P. J. O., Romary, L., & Sagot, B. (2021, July). Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In *CMLC 2021-9th Workshop on Challenges in the Management of Large Corpora*.

Part 02

Datenaufbereitung für große Sprachmodelle Praxisbeispiel

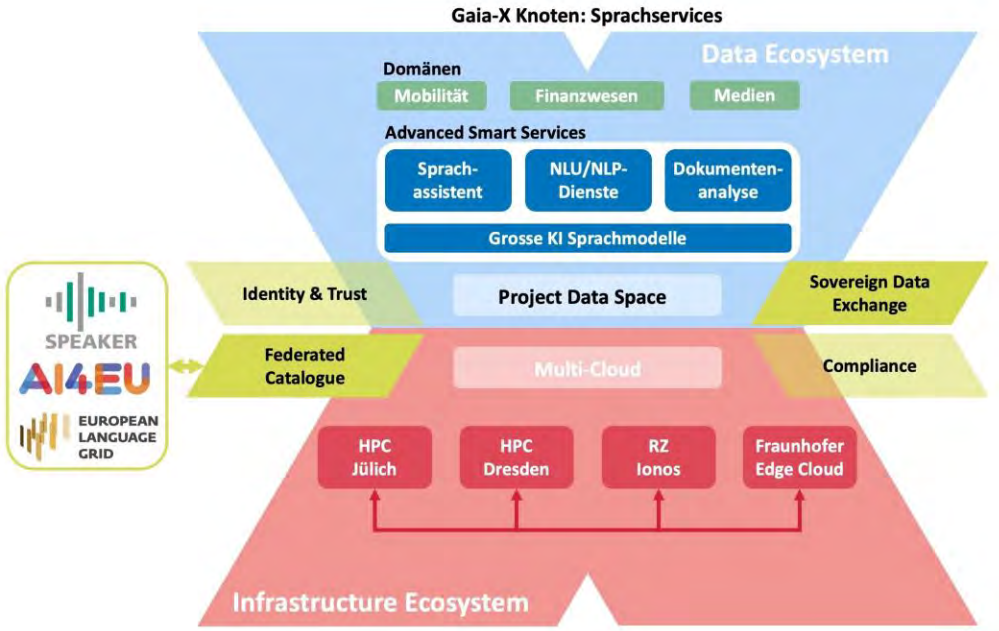
Open GPT-X

Europäisches Sprachmodelle

Ziele

OpenGPT-X erstellt und trainiert große KI-Sprachmodelle, um innovative Sprachanwendungsdienste für die europäische Wirtschaft voranzutreiben.

- Europäische Datenschutzstandards
- Fokussierung auf europäische Sprachen
- Gegengewicht zu chinesischen und amerikanischen Modellen



Unsere Rolle

MLOps für große Sprachmodelle

Aktuell: Unterstützung bei der Datenaufbereitung

Das Training der Sprachmodelle für europäische Sprachen bedarf einer großen Menge aufbereiteter Daten.

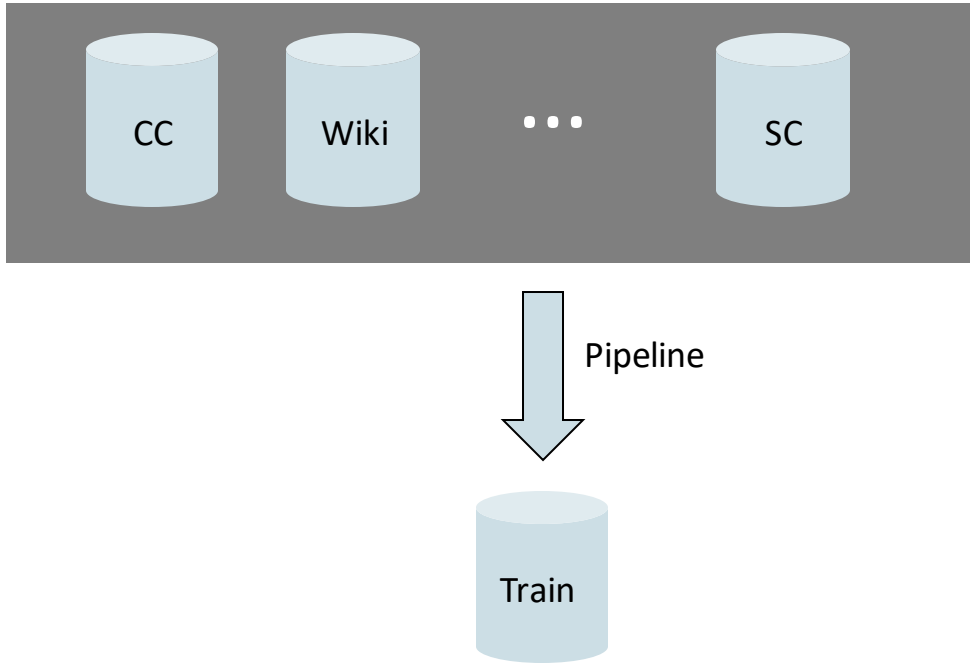
- Unterstützung bei der Implementierung der Datenpipeline
- Durchführung der Aufbereitungsprozesse
- Bereitstellung der Daten

Ziel: MLOps für große Sprachmodelle greifbar machen

Aufbau einer Dateninfrastruktur
Datenquellen

Datenquellen

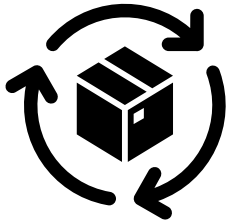
- Common Crawl
- Wikipedia
- OSCAR
- C4
- Pile
- Redpajama
- Github
- Star Code
- weitere kuratierte Datensätze...



Insbesondere Daten aus dem Web sind in großen Mengen verfügbar

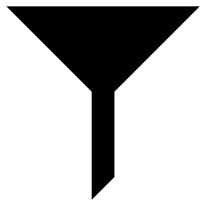
Aufbau einer Datenpipeline

Die Bestandteile



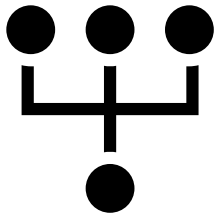
Konvertieren

- Vielzahl an Quellen
- Vielzahl an Formaten
- Große Datenmengen



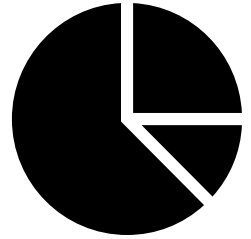
Filtern

- „Gute“ Texte für das Training verwenden
- Metadatenberechnung



Deduplizieren

- Hohe Hardware Anforderungen
- Potentiell große Auswirkungen auf Modellqualität



Sampling

- Ausgewogene Verhältnisse
- Schwierigkeit: genügend Daten je Sprache

Deduplizierung

Warum ist Deduplizierung notwendig?

Vorteile der Deduplizierung beim Training großer Sprachmodelle

- Die Deduplizierung des Trainingsdatensatzes reduziert die Rate der Ausgabe von gespeicherten Trainingsdaten um den Faktor 10
- Das Training von Modellen auf deduplizierten Datensätzen ist effizienter, da die Datensätze bis zu 19% kleiner sind.
- In einigen Fällen reduziert Deduplizierung die Perplexity um bis zu 10 %.

Lee, K., Ippolito, D., et al., Deduplicating Training Data Makes Language Models Better. ArXiv 2022.

Perplexity ist ein Qualitätsmerkmal für Sprachmodelle. Im Kern wird die „Verwirrung“ des Modells durch unbekannte Daten gemessen. Je geringer die Perplexity, umso besser war das Training des Modells.

Deduplizierung
Ausgangslage

Herausforderungen

- Deduplizierung ist Hardware hungrig
- Verschiedene Verfahren
- Partitionierung der Daten

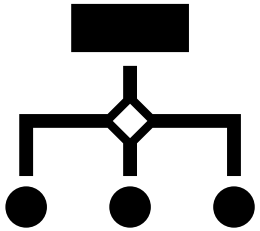
Verwendung der richtigen Hardware

- Zugriff auf große HPC Cluster
- Server, bereitgestellt durch Ionos
- Fraunhofer interne Rechencluster

Deduplizierung
Herangehensweise

Unser Ansatz

- Verwendung des MinHash Algorithmus
- Verwendung von Spark zur effizienten Parallelisierung
- Partitionierung der Daten nach Sprachen



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Spark

Durchführung weiterhin sehr Ressourcen intensiv. Parallel untersuchen wir Alternativen

Aktuelle Forschung

Themen

- Vergleich verschiedener Deduplizierungsverfahren und Herangehensweisen bei der Partitionierung
- Pipelinestruktur in SLURM basierten Systemen
- Veröffentlichung von (Zwischen-) Ergebnissen

Algorithmen
MinHash
ExactHash
SimHash
Bloom Filters
Suffix Array

Ausblick
2024

LLMOps : Was bedeutet das?

- Trainingspipeline
- Automatisierte Modellevaluation
- Vertrauenswürdigkeit im Betrieb
- Monitoring
- Drift Detection
- Continous Retraining
- …

Die Entwicklung und der Betrieb großer Sprachmodelle ist komplex und bedarf des Einsatzes guter Strukturen

Kontakt

Hammam Abdelwahab
hammam.abdelwahab@iais.fraunhofer.de

Lennard Helmer
lennard.helmer@iais.fraunhofer.de

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Schloss Birlinghoven 1
53757 Sankt Augustin

www.iais.fraunhofer.de