WordScape: a Pipeline to extract multilingual, visually rich Documents with Layout Annotations from Web Crawl Data

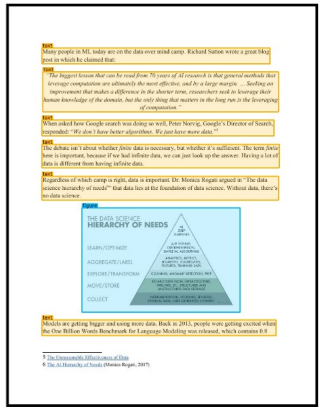
Maurice Weber[1], Carlo Siebenschuh[2], Rory M. Butler[2], Anton Alexandrov[1,3], Valdemar R. Thanner[1], Georgios Tsolakis[1], Haris Jabbar[4], Ian Foster[2,5], Bo Li[6], Rick Stevens[2,5], Ce Zhang[1]

[1]ETH Zürich, [2]University of Chicago, [3]INSAIT, Sofia University, [4]TU Darmstadt, [5]Argonne National Laboratory, [6]UIUC

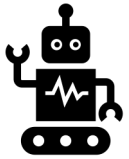
NeurIPS Datasets and Benchmarks, 2023

https://github.com/DS3Lab/WordScape

Visually rich Document Understanding



Q: *"What is the Mind vs. Data dispute about?"*



A: *" The Mind vs. Data dispute in AI encapsulates the debate about the nature of artificial intelligence. It essentially revolves around two contrasting … "*

Setting

- Abundance of semi-structured data in visual documents (PDFs, MS Word, …)
- Easily understood by humans; difficult for automated data processing engines

Challenges

- Inherently multimodal problem, combining text, visual features and layout.
- Diversity in how Information is represented across Industries, cultures, languages.
- Vast Amounts of high-quality data needed.

Contributions

- We present **WordScape**[1], a pipeline which enables the creation of large training datasets for document understanding.

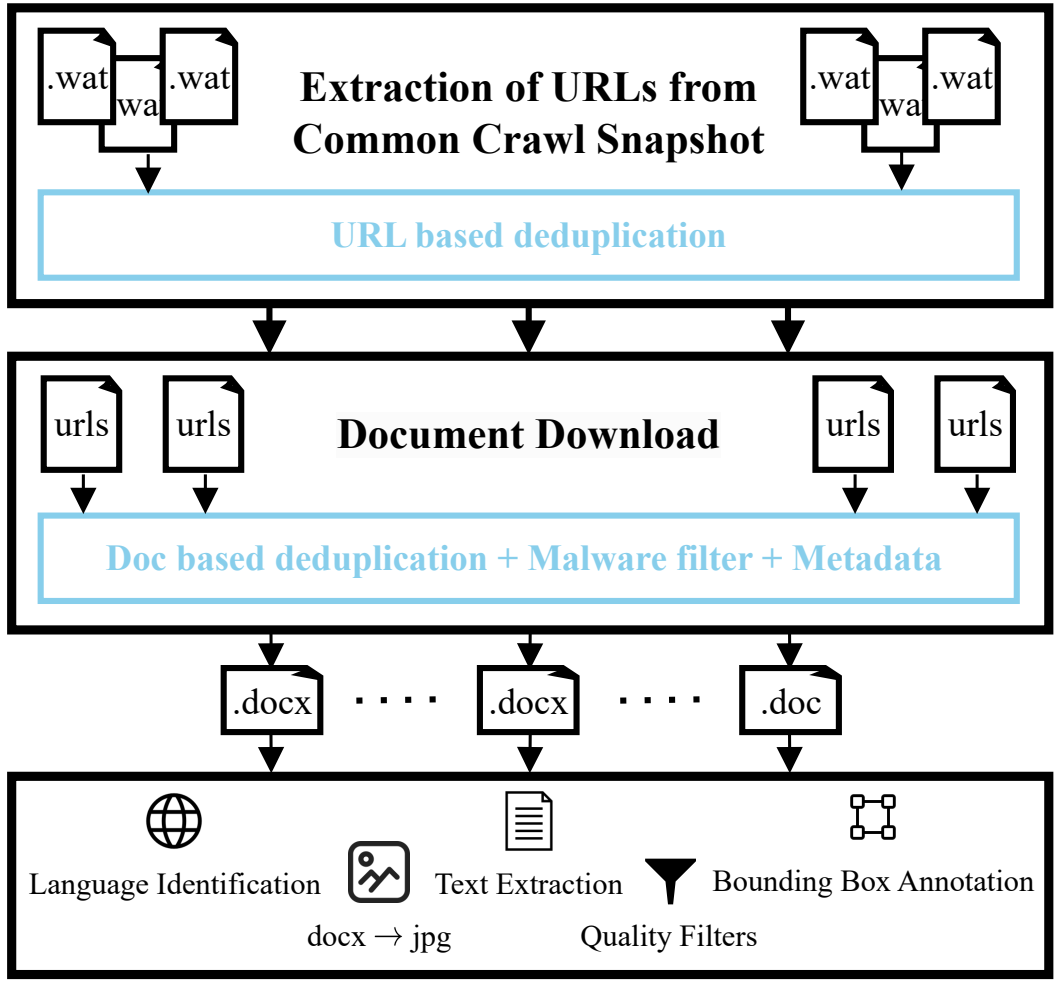
- **Datasets** created with WordScape consist of
  (1) Rendered document **page images**,
  (2) **Text** extracted in reading order,
  (3) **Bounding Box annotations** for semantic entities.

- We show that pretraining on WordScape can significantly **reduce the need for human annotated labels** on four benchmark tasks.

- We release **9.4M URLs** to documents together with the pipeline code, enabling the creation of datasets with up to **40M pages**.

https://github.com/DS3Lab/WordScape

Pipeline Overview



1) URL Extraction

- Extract URLs from CommonCrawl pointing to word documents
- Deduplication based on the URLs

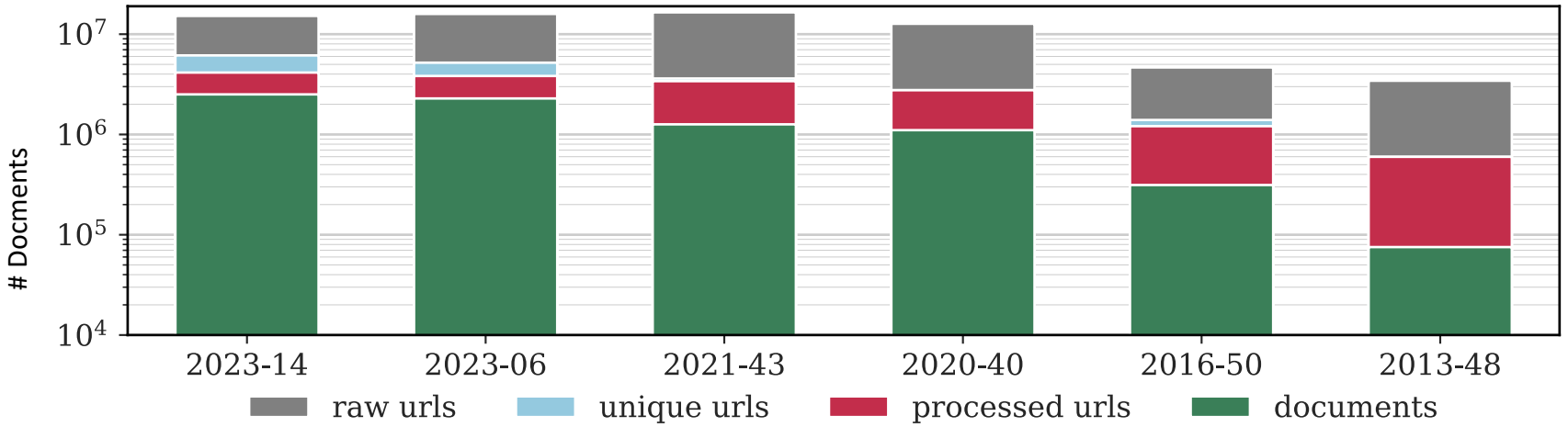
2) Document Download

- Distributed download on a SLURM cluster
- Document Deduplication, Malware Filter, Metadata

3) Document Annotation

- Annotate Layout using the OXML format
- Render document pages
- Extract Text, Identify Language, Quality Filters

Document URLs



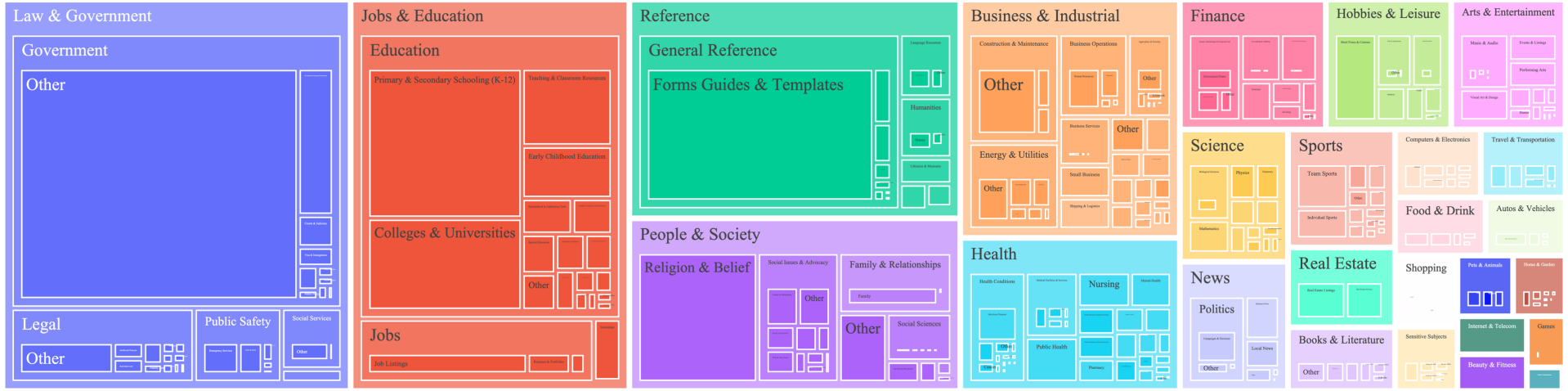
- We release WordScape as a set of **9.4M URLs to Word documents**, together with checksums and the pipeline code on Github[1].
- The URLs were **deduplicated** using the file checksums and return valid responses[2].
- The URLs cover **6 CommonCrawl** Snapshots between 2013 and 2023.

[1] https://github.com/DS3Lab/WordScape

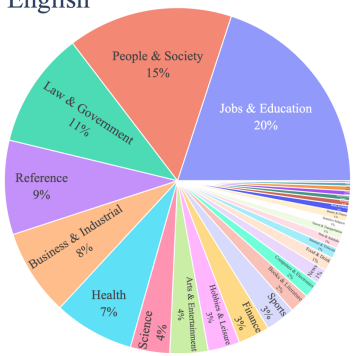
[2] as of June 6th, 2023.

Topic Distribution

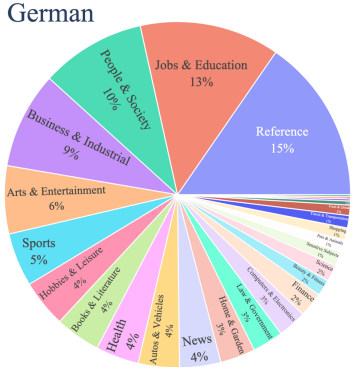
Overall



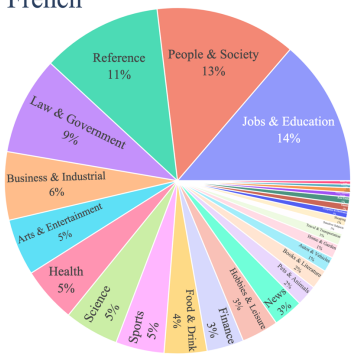
English



German



French



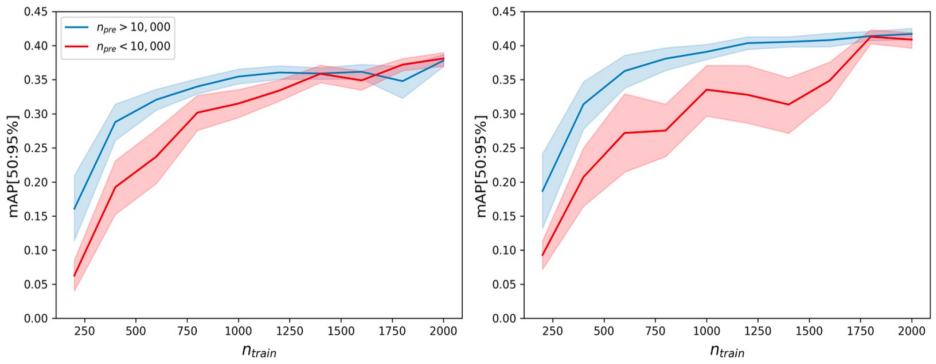
- **Hierarchical topic modelling** using Google Cloud NLP API[1] on a subset of 25K documents from the 2022-49 CommonCrawl Snapshot.
- Overall most frequent Top-level category is **Law & Government (21.9%)**, followed by Jobs & Education (17.5%)
- Significant differences **between languages** (e.g., en vs. de)

[1] https://cloud.google.com/natural-language

Benchmark Evaluations

- **Object Detection** evaluation on Layout Analysis Benchmarks and a custom dataset with scientific documents.
- Pretraining on WordScape, can significantly **reduce the number of high-quality finetuning samples** needed.
- Pretraining on WordScape yields better results than pretraining on PubLayNet[1], for Layout analysis on DocLayNet[2].

Layout Detection mAP@IoU[0.5:0.95] on the custom scientific papers dataset. YOLOv8 (left), and DETR (right).



Layout Detection mAP@IoU[0.5:0.95] on DocLayNet

	$N_f = 1k$	$N_f = 5k$	$N_f = 20k$	$N_f = 69k$
Random Initialization	0.299	0.553	0.727	0.753
PubLayNet (200k)	0.467	0.659	0.720	0.745
WordScape (200k)	**0.508**	**0.679**	**0.734**	**0.755**

Table Detection mAP@IoU[0.5:0.95] on ICDAR 2019 cTDaR

	$N_f = 75$	$N_f = 150$	$N_f = 300$	$N_f = 600$
$N_p = 0$	0.869 ± 0.008	0.888 ± 0.011	0.949 ± 0.006	0.974 ± 0.003
$N_p = 1.25k$	0.906 ± 0.012	0.912 ± 0.011	0.951 ± 0.005	0.972 ± 0.003
$N_p = 2.5k$	0.914 ± 0.009	0.929 ± 0.008	0.960 ± 0.004	0.974 ± 0.003
$N_p = 5k$	$\mathbf{0.924} \pm 0.007$	0.924 ± 0.011	0.956 ± 0.005	0.974 ± 0.003
$N_p = 10k$	0.919 ± 0.006	$\mathbf{0.931} \pm 0.010$	$\mathbf{0.961} \pm 0.005$	$\mathbf{0.975} \pm 0.003$

Text Detection F1@IoU 0.5 on FUNSD

	$N_f = 25$	$N_f = 50$	$N_f = 100$	$N_f = 149$
$N_p = 0$	0.621	0.690	0.723	0.772
$N_p = 10k$	0.840	0.840	0.823	0.861
$N_p = 50k$	0.868	**0.870**	**0.857**	0.869
$N_p = 100k$	**0.872**	0.869	0.850	**0.882**

[1]Zhong et al., *Publaynet: largest dataset ever for document layout analysis.* ICDAR, 2019.

[2]Pfitzmann et al. *Doclaynet: A large human-annotated dataset for document-layout segmentation.* ACM SIGKDD, 2022.

Conclusion

- We proposed a pipeline to create datasets consisting multilingual, diverse, visually rich **documents** with **layout annotations**.
- Scales to **millions of pages**, fusing text, visual and layout annotations.
- Dataset is **published as a list of 9.4M URLs** together with the pipeline code available on Github[1].
- Main **limitation** is the reliability of bounding box annotations for certain semantic entities, as they rely to some extent on the assumption that formatting correlates with user intent.
- Exploring **multimodal models** trained on WordScape, making full use of text, image and layout modalities, is fruitful ground for future research.

[1] https://github.com/DS3Lab/WordScape