DATA QUALITY, LANGUAGE COVERAGE

Common Crawl ~ 9 PiB
↳ only TEXT (HTML or PDF)

Domain level sampling → "BUDGET" for DOMAIN
↳ GRAPH-BASED RANKING

→ Below rank ~ 8M → SPAM

IN TIME...

- List based → page rank → harmonic centality (NOW)
  + NEW & REVISITS OF URL
  → NEW
  REPRESENTETIVE ? → Duplicate
  ↳ Depth of site
  Text quality → Regional coverage

DOWNLOADING Common Crawl

→ DOWNLOAD-PATH

cc-downloader download-paths CC-MAIN-2024-46 wet
path To Folder

Formats

1) WARC - web page captures
2) ARC - older, before 2012
3) WAT - metadata and links
4) WET - plain text extracted From HTML
5) Index → CDX format or Parquet
6) Webgraphs & RANKS
7) Crawl metrics

WARC → • Freezes interent traffic between a client
- Succesful | Not succesful collected seperatedly
- robots.txt
- Shuffled Files
- 1 MiB max content → otherwise TRUNCATED

Metadata → Replicability of data
- WARC & HTTP headers
- robots.txt, 404

Applications

- Before only Wik. & Books were used
- But CamemBERT peformed well on CC data !!
  - Bias
  - Code
  - Not representative
  - Unbalanced domain & Languag

Text extraction

- jusText
- TrafilaturaSe → Different or
- Resiliparse
- Readability

Heuristic Filtering
- Lenght
- symbol-To-word-ratio
- Remove lists → Discourage BLOCKLIST of WORDS
- Require alphabetic caracter
- Stop words lists
- Repetitions

Deduplication
- xxhash64 (not GPU!! and FAST)
- MinHash (fuzzy and GPU)

Quality Filter (Perplexty Filtering)
- KenLM trained on Wikipedia
- KenLM trained on adult content } n-gram models
- Quality classifiers
- Education classifiers

LangID project → Language identification
Web Language Project