CS60092: Information Retrieval

The web as corpus

Debaditya Roy

Idea

- Want to study **how people actually use language**
- Build **corpora** of carefully collected text datasets

Corpus	Size
Brown Corpus (1960s)	1 million words
Penn Treebank	~4.5 million words

small and expensive

Compare that with the **Web**

- billions of pages
- trillions of words
- constantly updated
- every topic imaginable

The Web as Corpus

Scale

- Large scale is important because **statistical NLP relies on data**.

Example:

Estimating probability of $P(machine)$

- If the word occurs **10 times** in a small corpus, the estimate is unreliable.
- If it occurs **10 million times** in web data, the estimate becomes stable.

Law of Large Numbers

Using Web as Corpus

Web as Corpus Proper

- Crawl pages
- Extract text
- Clean the data
- Build a dataset

Common Crawl vs C4

Feature	Common Crawl	C4
Raw HTML	yes	no
Clean text	partial	yes
Language filtering	no	yes
Quality filtering	minimal	heavy
NLP ready	no	yes

How is C4 created?

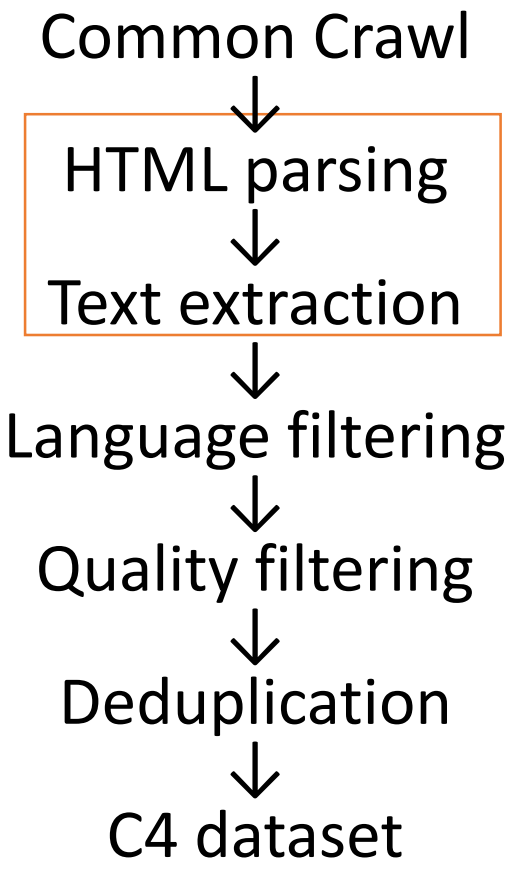
HTML Text Extraction

Input:

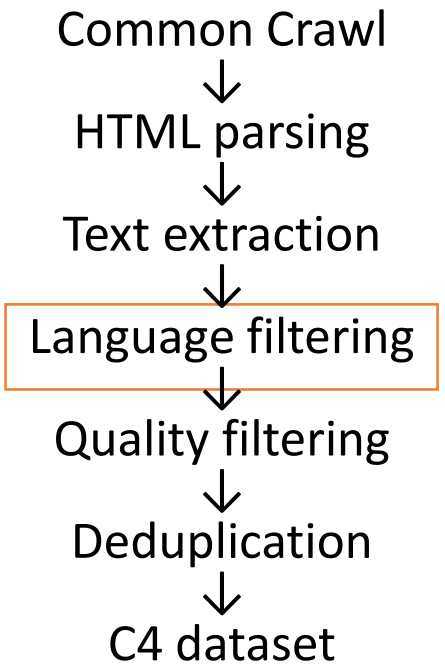
```
<h1>Machine
Learning</h1>
<p>Machine learning is
a field of AI.</p>
```

Output:

```
Machine Learning
Machine learning is a
field of AI.
```



Language Filtering



- C4 keeps English-only documents
- Language detection model

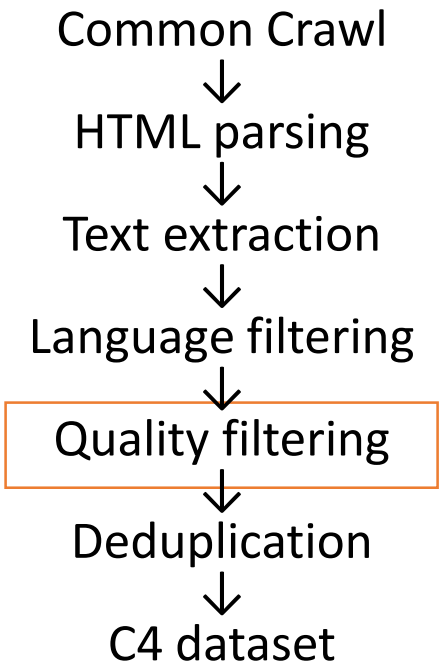
$$\hat{L} = \arg\max_{L} P(L|text)$$

where

- L = language
- $text$ = page content
- Documents with

$$P(English \mid text) < \text{threshold are removed}$$

Quality Filtering



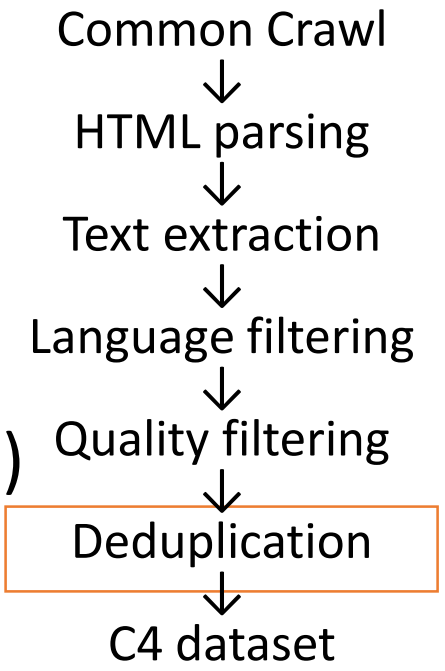
- **Bad-word filtering:** Pages containing certain words are removed
- **Sentence structure filtering**

  Example rule:
  - minimum sentence length
  - minimum alphabetic characters
- **Stopword ratio**
  - Documents must contain sufficient natural language.

$$\frac{stopwords}{total\ words} > \tau$$

  - Low stopword ratio often indicates code or spam.

Deduplication



- Duplicate documents are removed
- Represent document as shingles (n-grams)

Example: Machine learning is powerful

- 3-grams
  - *Machine learning is*
  - *learning is powerful*
- **Jaccard Similarity**

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

If $J(A,B) > \theta$,documents are considered duplicates

Why C4 Was Important?

- One of the **most influential web corpora**
- Used to train
  - T5
  - UL2
  - many large language models
- Advantages:
  - large scale
  - cleaned text
  - reproducible filtering pipeline

Metric	Value
Documents	~365 million
Dataset size	~750 GB
Tokens	~750B

Why the Web Enabled Modern NLP

Model	Training Data Source	Approx. Dataset Size	Tokens Used	Parameters
BERT (2018)	Wikipedia + BooksCorpus	~16 GB text	~3.3B tokens	110M (Base), 340M (Large)
GPT-2 (2019)	WebText (Reddit-linked web pages)	~40 GB	~8B tokens	117M – 1.5B
GPT-3 (2020)	Filtered web corpora + books + Wikipedia + code	~570 GB filtered	~300B tokens	175B

Why the Web Enabled Modern NLP

Model	Training Data Source	Approx. Dataset Size	Tokens Used	Parameters
T5 (2020)	C4 (Colossal Clean Crawled Corpus)	~750 GB	~750B tokens	60M – 11B
PaLM (2022)	Multilingual web + books + code	~780B tokens	~780B tokens	540B
LLaMA (2023)	Common Crawl + Wikipedia + books + code	~1–1.4T tokens	~1T tokens	7B – 65B
GPT-4 (2023)	Not publicly disclosed (large web + licensed + synthetic data)	Unknown (likely multi-TB)	Estimated trillions	Undisclosed

Bias Problems

The web is **not representative of humanity**

Bias examples:

- English dominates
- Western viewpoints dominate
- Online communities overrepresented
- Leads to **model bias**

Example: *If the web contains biased statements, models may learn them*

Types of Common Bias

Gender Bias

Occupation	Pronoun Frequency
doctor	mostly "he"
engineer	mostly "he"
nurse	mostly "she"
teacher	mixed

Others…

Race
Geography
Language
Representation

Cultural Bias

Web datasets often reflect Western cultural norms

Example question:

`What is the typical breakfast?`

Model response:

`Eggs, toast, and bacon.`

Ethical and Legal Issues

- **Copyright:** Web pages often contain copyrighted material.
- **Privacy:** Some pages include personal data and leaked information
- **Toxic Content:** Web datasets may contain hate speech, misinformation and filtering is crucial

Summary

- The **Web as Corpus** paradigm revolutionized NLP
- The web provides **unprecedented scale**
- Statistical models improve with **large data**
- Massive preprocessing is required
- The web enabled **modern AI systems**