CMU 11-667 Homework 1

HanlinXu

Date: 07/12/2024

Question 1.1 (5 points)

1.2.A Knowledge Cutoff

C4: The "knowledge cutoff" for C4 is typically determined by the latest crawl date from Common Crawl, and a widely used version (like in T5) is based on 2019 data.
RedPajama-Data-v1: The initial release of RedPajama was in April 2023

1.2.B Data Source

C4: Sourced from the Common Crawl corpus which consists primarily of web pages, including news articles, blogs, and more.
RedPajama-Data-v1: RedPajama is a clean-room, fully open-source implementation of the LLaMa dataset. Commoncrawl, C4, GitHub, Wikipedia, Gutenberg and Books3, ArXiv, Stackexchange.

1.3.C Open-sourced Language Model

C4: T5 (Text-to-Text Transfer Transformer)
RedPajama-Data-v1: LLaMA

1.4.D License and Copyright Protections

C4: The dataset is released under the Common Crawl license. Documents within the dataset inherit diverse copyright protections depending on their original source. For example, News articles might have strict copyright. Blog posts may have permissive terms.
RedPajama-Data-v1: It aggregates data with varying protections.

1.5.E Task Performance Challenges

C4: A language model trained solely on C4 might struggle with tasks requiring domain-specific knowledge, like biomedical or legal text processing, due to a lack of specialized data.
RedPajama-Data-v1: Tasks requiring real-time or highly dynamic information

Question 1.2 (3 points)

1.2.A Example

1. API-Only Access (No Public Weights): OpenAI's GPT-4
2. Public Weights, Non-Public Pre-training Dataset: Meta's LLaMA 2
3. Public Weights and Pre-training Dataset: RedPajama-INCITE-Base-7B

1.2.B Compare and Contrast

Transparency: RedPajama offers the most transparency, detailing data sources and processing steps. LLaMA 2 provides general information without specifics, while GPT-4 offers minimal details.
Data Accessibility: RedPajama's datasets are publicly accessible, LLaMA 2's datasets are not publicly available, and GPT-4's datasets are undisclosed.

1.2.C Use Case for Models Trained on Publicly Accessible Data

Scenario: Developing educational tools that require transparency and reproducibility.
Transparency: Models like RedPajama, with publicly accessible data, allow educators to understand and explain the data sources and processing methods, fostering trust and facilitating learning.
Reproducibility: Public datasets enable educators and students to replicate experiments.
Ethical Considerations: Using models with transparent data sources ensures adherence to ethical standards.

Question 2.1 (6 points)

2.1.A

6368

2.1.B HTML (), code

HTML tags (e.g., `<code>`): `html_to_text` treats them as regular text.

```
<p>This is a paragraph with <code>&lt;div&gt;</code>.</p>
→
This is a paragraph with `<div>`.
```

Code blocks: `html_to_text` ensures the code is preserved in its textual form and formatted.

```
<code>if (x > 0) { return x; }</code>
→
`if (x > 0) { return x; }`
```

2.1.C headers (e.g. h1, h2, etc.), embedded images (e.g. the img tag), and tables

Headers: extracts only the plain text without hierarchical level.

```
<h1>Main Title</h1>
<h2>Subtitle</h2>
→
Main Title
Subtitle
```

Embedded Images: extracts the alt attribute of images as text.

```
<img src="example.jpg" alt="Example Image">
<img src="no-alt.jpg">
→
Example Image
[Image]
```

Tables: extracts table content row by row, separating cells with tabs.

```
<table>
<tr><th>Header1</th><th>Header2</th></tr>
<tr><td>Data1</td><td>Data2</td></tr>
</table>
→
Header1     Header2
Data1       Data2
```

2.1.D Common Crawl team html to text

The acronym WET stands for "WARC Encapsulated Text". As many tasks only require textual information, the Common Crawl dataset provides WET files that only contain extracted plaintext.

Question??

Question 2.2 (8 points)

2.2.A high-quality documents

3297

2.2.B Identify two low-quality documents and evaluate filter approaches

Low-Quality Examples: repetitive or boilerplate content, large portions of broken text or invalid characters.
heuristic-based filte: Heuristics like detecting repeated phrases or overuse of certain keywords (e.g., "Privacy Policy", "Terms") could work. However, this approach may incorrectly remove some valid documents containing similar legal language. Detecting an excessive percentage of invalid characters or non-alphanumeric symbols could flag problematic documents.
Classifier: Building a classifier requires labeled training data and computational resources, making it more complex than heuristics.

2.2.C Impact of Filtering and Cleaning on Non-English Text

Filtering Impact: Checking for punctuation or a high percentage of alphanumeric characters can disproportionately exclude non-English documents that use non-Latin scripts. These scripts often lack the same punctuation density as English.
Cleaning Impact: Removing paragraphs without punctuation or containing long sequences of alphanumeric characters may unintentionally exclude valid text in certain languages (e.g., Japanese, which may not always use punctuation in the same way).

2.2.D Domain-Specific Cleaning and Filtering

Domain: Social Media Data
Remove Retweets or Duplicated Content: For example, in tweets, remove text starting with "RT @" to avoid redundant retweets in the dataset.
Spam Detection: Use keyword-based filtering to exclude tweets containing an excessive number of hashtags, links, or promotional phrases (e.g., "Buy now!", "Click here").

2.2.E A Survey on Data Selection for Language Models

See Figure 1

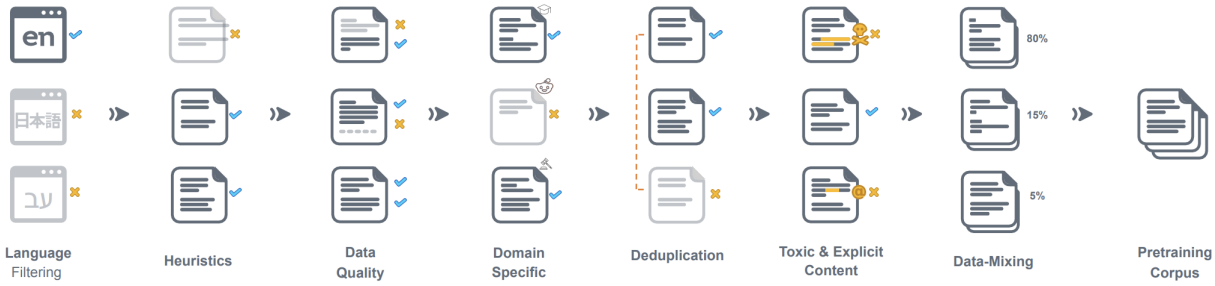

Figure 1: An overview of the data filtering pipeline for pretraining

Question 2.3 (6 points)

2.3.A How long in seconds does your code take to process the entire datase

Generating train split: 2924 examples [09:52, 4.94 examples/s]

2.3.B How to make the code significantly faster

Parallel Processing for WARC File Parsing
Batch Processing

2.3.C Advantage of Packing Over Padding

Packing refers to concatenating shorter sequences together into a single batch up to the maximum sequence length, while padding fills shorter sequences with a no-op token until they reach the maximum sequence length.
Advantages of Packing:
1. Better Utilization of Sequence Length
2. Higher Training Efficiency
3. Reduced Memory Usage
4. Improved Generalization