Exercise: Research Pre-Training Datasets

Lesson Downloads

When it comes to training language models, selecting the right pre-training dataset is important. In this exercise, we will explore the options available for choosing a pre-training dataset, focusing on four key sources:

- CommonCrawl,
- Github,
- Wikipedia, and
- the Gutenberg project.

These sources provide a wide range of data, making them valuable resources for training language models. If you were tasked with pre-training an LLM, how would you use these datasets and how would you pre-process them? Are there other sources you would use?

In this exercise, you will construct a fictional pre-training dataset for a fictional task. The goal is to get you thinking about how to construct a pre-training dataset for your own task.

Step 1: Evaluate the available pre-training datasets

Begin by examining the four sources mentioned in the introduction - CommonCrawl, Github, Wikipedia, and the Gutenberg project. Assess the size, quality, and relevance of the data provided by each source for training language models.

CommonCrawl

Read about CommonCrawl on its websi[illegible]org/

Question 1 of 5

What is the size of the CommonC[illegible]ataset

- Less than 10 TB
- Between 10TB and 100TB
- Between 100TB and 1PB
- Greater than 1PB

Submit

Question 2 of 5

How could one best describe the d[illegible]CommonCrawl dataset

- Highly curated and structured
- Semi-structured and clean
- Unstructured and noisy

Submit

Github

Read about the Github dataset on its web[illegible]https://www.githubarchive.org/

Question 3 of 5

The Github dataset contains both public and private repositories.

- False
- True

Submit

Wikipedia
Read about the Wikipedia da… website: [wikim… …nloads](…).

Question 4 of 5
What formats are the Wikipedia …
- XML
- JSON
- SQL
- All of the above

Submit

Gutenberg Project
Read about the Gutenberg Projec… website: [https://www.g… …berg.org/](…)

Question 5 of 5
How many books are in the Gu… …erg …
- At most 10,000
- 10,000 - 100,000
- 100,000 to 1 million
- More than 1 million

Submit

Step 2. Select the appropriate datase…
Based on the evaluation, choose the da… …requirements of pre-training a Language Model (LLM). Consider factors such … …ersity of data, domain-specific relevance, and the specific language model objectives.

For your use case, rank the datasets in order of preference. For example, if you were training a language model to generate code, you might rank the datasets as follows:
1. Github
2. Wikipedia
3. CommonCrawl
4. Gutenberg project

Explain your reasoning for the ranking. For example, you might say that GitHub is the best dataset because it contains a large amount of code, and the code is structured and clean. You might say that Wikipedia is the second-best dataset because it contains a large amount of text, including some code. You might say that CommonCrawl is the third-best dataset because it contains a large amount of text, but the text is unstructured and noisy. You might say that the Gutenberg project is the worst dataset because it contains text that is not relevant to the task.

Dataset Ranking
Rank the 4 datasets in order or preference for training your large language model.

Enter your response here, there's no right or wrong answer

Submit

Explain your ranking
In a few sentences explain why your chose your ranking. Why was the first dataset you chose so important?

Enter your response here, there's no right or wrong answer

Submit

Step 3. Pre-process the selected datasets
Depending on the nature of the chosen datasets, pre-processing may be required. This step involves cleaning the data, removing irrelevant or noisy content, standardizing formats, and ensuring consistency across the dataset. Discuss how you would pre-process the datasets based on what you have observed.

Pre-process the selected datasets
What sort of pre-processing would you conduct on these datasets? It may be useful to look at some of the data yourself to understand what sorts of concerns could emerge.

Enter your response here, there's no right or wrong answer

Submit

Step 4. Augment with additional sources
Consider whether there are other relevant sources that can be used to augment the selected datasets. These sources could include domain-specific corpora, specialized text collections, or other publicly available text data that aligns with your language model's objectives, such as better representation and diversity.

Augment your choices
Think about your use case and consider what additional sources your will need.

Enter your response here, there's no right or wrong answer

Submit

Exercise End

Great work! We've done some investigation to see what datasets we'll use to train our model. This very important aspect of the work should not be overlooked. After all, the foundation of a foundation model is its data!