COMP4901B Large Language Model

Lam Yeung Kong Sunny 20867162 ykslam@connect.ust.hk

Assignment 1

1. Task 1

- Commend used:
  - cd data_preprocess
  - python homework.py --fname data.warc --output cleaned_test.txt --dfname topic_dataset.json --num_records 2000
- Cleaning/heuristic/English detection logic:

clean_text:

1. Change all newline syntax to \n
2. Drop paragraphs that contain more than 100 alphanumeric characters with no whitespace between them using regex
3. Drop paragraphs that do not contain punctuation using regex
4. Get clean paragraphs
5. Join them with \n.
6. Return clean paragraphs

heuristic_quality_filter:

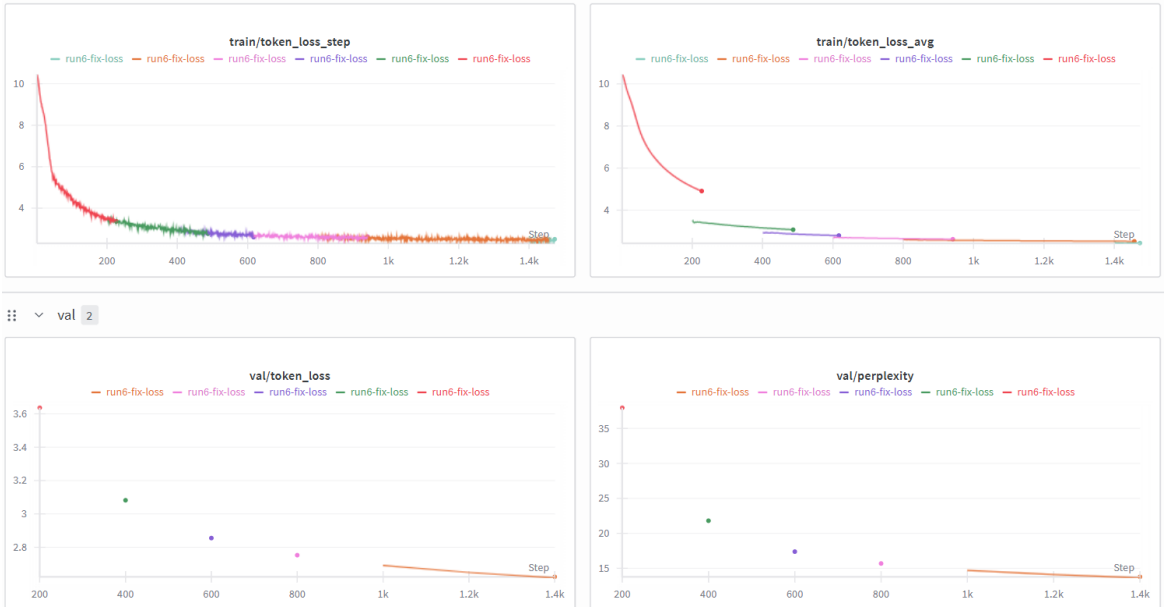
1. Split the text, where the gaps between them are spaces, to be tokens.
2. Return false if tokens contain bad word(s)
3. Return false if the text contains no punctuation
4. Return false if the text is purely whitespace.
5. Return true if 80% or more of its characters are alphanumeric, punctuation, or whitespace, return false otherwise

is_english_text:

1. Use detect_langs api to see if the text is English
2. If estimated probability is high enough, first condition is passed.
3. Use regex to find all English letters in the text in a list.
4. Find total letters using char.isalpha()
5. If len(English_letters) / len(total_letters) >= 0.9 and first condition is true, return true. Otherwise, return false

```
569 passed out of 2000 records processed.
Cleaned documents saved to: cleaned_test.txt
139 deduplicated out of 219 records processed.
```

2. Task 2
   Commend used: In run_babylm.sh
   - Training setup
     - GPU
     - effective batch size : 512
     - Micro batch size: 32
     - sequence length: 256
     - LR: 0.001
     - Warmup ratio: 0.1
     - Steps: 1475 = 1 epoch



3. Task 3

- Commend used: In colab_demo.ipynb

  ! cd /content/drive/MyDrive/COMP4901B-Homework1/COMP4901B-LLMs-main/assignment1/llama_training && python run_llama.py --pretrained-model-path run6-fix-loss-pretrain-1-0.001.pt --option generate

  from huggingface_hub import hf_hub_download

repo_id = "yuzhen17/llama2-42M-babylm"

filename = "llama2-42M-babylm.pt"

local_download_path = "/content/drive/MyDrive/COMP4901B-Homework1/COMP4901B-LLMs-main/assignment1/llama_training"

hf_hub_download(
    repo_id=repo_id,
    filename=filename,
    local_dir=local_download_path
)

! cd /content/drive/MyDrive/COMP4901B-Homework1/COMP4901B-LLMs-main/assignment1/llama_training && python run_llama.py --generated_sentence_low_temp_out generated-sentence-temp-0_providedModel --generated_sentence_high_temp_out generated-sentence-temp-1_providedModel.txt --option generate

- My Trained model:

  Temperature is 0.0

  White Bird is a 2023 American war drama movie starring Jackie Hutton, Jackie Hutton, Jackie Hutton, Jackie Hutton, Jackie Hutton, Jackie Hutton, Jackie Hutton, Jackie Hutton, Jackie Hutton, Jackie Hutton, Jackie Hutton, Jackie Hutton, Jackie Hutton, Jackie Hutton, Jackie Hutton,

  ---

  Wrote generated sentence to generated-sentence-temp-0.txt.

  load model from run6-fix-loss-pretrain-1-0.001.pt

  Temperature is 0.5

  White Bird is a 2023 American war drama movie starring Marlon McClay, Joe McClay, Ronald McClay, Ronald McClay, Ronald McClay, David, Ronald

McClay, Ronald McClay, Ronald McClay, Shawn McClay, Ronald McClay, Michael McClay, Ronald McClay, Ronald Mc

---

Wrote generated sentence to generated-sentence-temp-1.txt.

- Provided Trained Model:

  Temperature is 0.0

  White Bird is a 2023 American war drama movie starring Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt, Diana Hunt,

  ---

  Wrote generated sentence to generated-sentence-temp-0_providedModel.

  load model from llama2-42M-babylm.pt

  Temperature is 0.5

  White Bird is a 2023 American war drama movie starring George D. Dixon, Richard Dixon, Tiffany, George Deal, David L. M. Dixon, Richard Dixon, Chris Dixon, George A. C. Dixon, Richard Dixon, David H. Clark, Bill H. Bush, David H. Bush, John H. Bush, David H. Bush, Peter H. Bush

  ---

  Wrote generated sentence to generated-sentence-temp-1_providedModel.txt.

  Explanation:

  The temperature = 0.5 will be better, as you can in the results for both models, the outputs with temp=0.5 are more diverse and less cohesive. On the other hand, the outputs with temp=0 are less diverse and more cohesive. In a normal writing, we typically want to have more diversity, like objects from examples or authors, which will cause the writing to be less weird as there will be less repeating tokens.