Week 1: Dataset Creation and Cleansing

Task Description:

For Week 1, your task is to first create a **raw database** of any creative type. This can include:

- Poems
- Short stories (beware of their length)
- Recipes
- Jokes (but prepare yourself for potentially embarrassing outputs from your GPT-2 model 😅), etc.

Save the dataset in a JSON file. Here's an example structure for your raw dataset:


```
[
  {
    "poem": "I am taken with the hot animal\nof my skin, grateful to swing my limbs\n\nand have them move as I intend, though\nmy knee, though my shoulder, though something\nis torn or tearing. Today, a dozen squid, dead\n\non the harbor beach: one mostly buried,\none with skin empty as a shell and hollow\n\nfeeling, and, though the tentacles look soft,\nI do not touch them. I imagine they\nwere startled to find themselves in the sun.\n\nI imagine the tide simply went out\nwithout them. I imagine they cannot\n\nfeel the black flies charting the raised hills\nof their eyes. I write my name in the sand:\nDonika Kelly\n. I watch eighteen seagulls\n\nskim the sandbar and lift low in the sky.\nI pick up a pebble that looks like a green egg.\n\nTo the ditch lily I say\nI am in love\n.\nTo the Jeep parked haphazardly on the narrow\nstreet\nI am in love\n. To the roses, white\n\npetals rimmed brown, to the yellow lined\npavement, to the house trimmed in gold\nI am\n\nin love\n. I shout with the rough calculus\nof walking. Just let me find my way back,\nlet me move like a tide come in."
  },
  {
    "poem": "It’s neither red\nnor sweet.\nIt doesn’t melt\nor turn over,\nbreak or harden,\nso it can’t feel\npain,\nyearning,\nregret.\n\nIt doesn’t have\na tip to spin on,\nit isn’t even\nshapely—\njust a thick clutch\nof muscle,\nlopsided,\nmute. Still,\nI feel it inside\nits cage sounding\na dull tattoo:\nI want, I want—\n\nbut I can’t open it:\nthere’s no key.\nI can’t wear it\non my sleeve,\nor tell you from\nthe bottom of it\nhow I feel. Here,\nit’s all yours, now—\nbut you’ll have\nto take me,\ntoo."
  }.....
]
```

You can also include additional metadata for each entry, such as:

- **Tags**
- **Author**
- **Type** (e.g., sonnet, prose, etc.)

Dataset Size:

Make sure the dataset is appropriately sized:

- **Too small**: Risk of underfitting during training.
- **Too large**: Prolonged training time (though it may yield better results).

Dataset Cleansing:

Once the raw dataset is ready, the next step is to create a **cleansed.txt** file. This involves:

1. Removing unnecessary elements, such as special characters (`@~{[<`).
2. Eliminating empty or duplicate entries.
3. Tailoring the cleansing process to your chosen category (e.g., preserving line breaks and punctuation for poems).

Below is a sample Python code snippet for cleansing:

```
import json

# Load raw data from JSON
with open("poems_data.json", "r") as f:
    data = json.load(f)

# Cleansing process
# Example: Remove entries with empty poems or filter based on
specific criteria
data = [poem for poem in data if len(poem["poem"].split())]

# Print the number of entries after cleansing
print(len(data))

# Check the first few cleansed entries
print(data[:5])
```

Next Steps:

Once you've completed this part, you can proceed to:

1. **Data Modeling**: Structuring the data for training.
2. **Dataset Creation**: Preparing the dataset.

Good luck, and happy dataset creation!

I will release the next steps of tokenization and Dataset creation soon.