Documentation for Story Generation Project

References

1. https://www.analyticsvidhya.com/blog/2018/03/text-generation-using-python-nlp/ (Basic example - training your own network - character based)
2. https://github.com/minimaxir/textgenrnn (Pre-trained character based model)
3. https://github.com/minimaxir/gpt-2-simple (Wrapper to train GPT 2 on your own data)
4. https://github.com/minimaxir/gpt-2-keyword-generation (Allows generation of test based on given keywords using gpt-2)
5. https://github.com/openai/gpt-2 (Official GPT 2 Repo)
6. https://talktotransformer.com/ (Test GPT 2 here!)
7. https://colab.research.google.com/drive/18OVj5BNCyRVUM0yG7DpFXlCXQZI6eaVs (Google Colab Notebook for training GPT-2-Simple-Keyword-Generation) - **Open in Google Chrome**
8. http://cs.rochester.edu/nlp/rocstories/ (5-sentence story dataset - 50000 strong)
9. https://docs.google.com/document/d/19SnginuamjEYaGPiokrAlPKt-m23BFBt6KUA7YCzeBU/edit?usp=sharing (latest version of this documentation)

Steps

1. The dataset for training must be in csv format. (Cannot be trained on other file formats?)
2. Clone the repo - https://github.com/prmehta24/gpt-2-keyword-generation
   Go to repo in terminal/cmd:
3. `pip3 install -r requirements.txt`
4. Also, install GPT-2 Simple as -
   `pip3 install gpt-2-simple`
   OR
   `pip install gpt-2-simple`
5. Place the csv file you want encoded in the repo. (eg ROCStories.csv)
6. Then, edit Trainer.py: (in gpt-2-keyword-generation/)
   a. such that csv_path points to your csv file.
   b. Also, edit all other fields in encode_keyword(), as needed. (Documentation in Trainer.py)

   Go to gpt-2-keyword-generation/
7. Now, run “python3 Trainer.py” to encode the csv dataset and get an encoded txt file (eg encodedStories.txt)
   Optional: Go to https://github.com/minimaxir/gpt-2-keyword-generation/blob/master/keyword_encode.py (Line 11) to understand the meaning of ~,@, ^, etc. in the encoded txt file

8. **Option 1 - Train via GPU (Google Colab) - Very Fast - But, can only train upto 12 hrs.**
    a. Take the encoded txt and upload it to the base folder of your drive.
    b. Open Reference 7 and follow the instructions there to train and play with your model.
    If you want to play with the model on your own computer:
    c. After downloading file.zip from Reference 7(Step 5), you can extract it. Use the extracted checkpoint folder and place in gpt-2-keyword-generation/example (if already present, replace the existing checkpoint folder)
    OR
    d. Download gpt2_weights.zip from the base folder of your drive. Unzip it. Use the extracted checkpoint folder and place in gpt-2-keyword-generation/example (if already present, replace the existing checkpoint folder)
    e. Now, run python3 loadmodel.py (in gpt-2-keyword-generation/example) to generate example text.
9. **Option 2 - Train on your own computer - Very slow**
    a. Edit gentxt.py (in gpt-2-keyword-generation/example) to point to file on which model will be trained
    b. Run python3 gentxt.py (saves model in checkpoint folder in current directory)
    c. Go to step 10