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/1bjveok8XMBFZ9TM9KmnnnbL-7fqmWiFQ> (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:*

1. pip3 install -r requirements.txt
2. Also, install GPT-2 Simple as -

pip3 install gpt-2-simple

OR

pip install gpt-2-simple

1. Place the csv file you want encoded in the repo. (eg ROCStories.csv)
2. Then, edit Trainer.py: ( in gpt-2-keyword-generation/ )
   1. such that csv\_path points to your csv file.
   2. Also, edit all other fields in encode\_keyword(), as needed. (Documentation in Trainer.py)

*Go to gpt-2-keyword-generation/*

1. 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*](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*

1. **Option 1 - Train via GPU (Google Colab) - Very Fast - But, can only train upto 12 hrs.**
   1. Take the encoded txt and upload it to the base folder of your drive.
   2. 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:*

* 1. 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

* 1. 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)
  2. Now, run python3 loadmodel.py (in gpt-2-keyword-generation/example) to generate example text.

1. **Option 2 - Train on your own computer - Very slow**
   1. Edit gentxt.py (in gpt-2-keyword-generation/example) to point to file on which model will be trained
   2. Run python3 gentxt.py (saves model in checkpoint folder in current directory)
   3. Go to step 10