Automated Sentence Completion Models Development

Guidelines to Run Models

After downloading the zip file, you need to unzip it, and there will be two folders inside: `NGram Models` and `LSTM Models`.

Running N-Gram Models

- In the `NGram Models` folder, there are two `.ipynb` files (each is a model) and ten `.txt` files.
- There are two options to run the code in `john_ngram_model.ipynb`:
  - First option is uploading `john_ngram_model.ipynb` and ten `.txt` files, then running it on Google Colab
  - Second option is running `john_ngram_model.ipynb` in your IDE (e.g. Visual Studio Code, PyCharm, etc.). Please make sure you have the `.txt` files placed within the same folder with the `john_ngram_model.ipynb` file.
- The code in `emma_ngram_model.ipynb` can be run in your IDE:
  - By default, the code is set up to print three suggestions for the next word following the segment “she was not,” as well as one sentence and the calculated perplexity.
  - By changing the value on line 148, you are able to change which ebook the model trains on. Uncomment lines 156-163 to train on every available ebook from the NLTK Gutenberg corpus.

Running LSTM Models (WARNING: It take a couple hours of time to train each LSTM model)

- In the `LSTM Models` folder, there are three `.ipynb` files (each is a LSTM model) and two `.txt` files. The following are applied to run all three models.
- You need upload one of three `.ipynb` files as a notebook on Kaggle. Then after opening the notebook, on menu bar under the notebook name, there is a `Settings` option. Choose `Settings > Accelerator > GPU P1000`. Make sure you have at least 10-hour usage for the GPU.
- You also need to upload the two `.txt` files as input on Kaggle:
  - Go to the `Input` tab on the right-hand side of the kernel interface.
  - Right of the `Add Input` button there is an `Upload` button
  - Click the `Upload > New Dataset` you can drag and drop the two `.txt` files, or you can browse and select them. Name your dataset as `lstm-model-data`.
  - Wait for the file to upload. Once it's done, it will be listed in the "Datasets" tab.
- After uploading the text files, you can now run all the cells.