Problem Set 12

Data Science 602: Data Analysis and Machine Learning

Spring 2022

In the Google shared drive (/602/data), the file enron.txt is a subset of the Enron Corpus, a collection of over 500,000 emails from senior management of Enron Corporation leading to its collapse in 2001[1]. The subset comprises the text of about 15,000 emails available through the TensorFlow Data Set (TFDS) source aeslc (annotated Enron Subject Line Corpus).

Using this dataset, construct a neural net that will generate 50 random characters, beginning with the sequence `The`, that are generated from the distribution of text in the file.

This exercise can be replicated using any of the following sources in the texts and documentation:

- **Raschka** - Character-level language modeling in TensorFlow, pages 600-613
- **Géron** - Generating Shakespearean Text Using a Character RNN, pages 526-534
- **TensorFlow documentation** Text Generation with an RNN https://www.tensorflow.org/text/tutorials/text_generation

Adjust the temperature (α in Raschka) to avoid repeating text. Using a GPU runtime to fit the model is advised, which may still require several hours to train.

[1]The data was made public in a subsequent investigation by the Federal Energy Regulatory Commission, and the original source is available through the Library of Congress as a 700 MB ZIP file at https://www.loc.gov/item/2018487913/