**Data Pipeline:** Federated Learning with Pretrained Text DNNs

Arjun Singh

Joel Stremmel

11-17-19

**Data Streams:**

Our data is accessible via the [tff.simulation.datasets](https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets) module in the [Tensorflow Federated API](https://github.com/tensorflow/federated). The [tff.simulation.datasets.stackoverflow.load\_data()](https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets/stackoverflow/load_data) method loads a mapping of clients (Stack Overflow user IDs) to examples (their posts and post metadata). The data is not brought into memory until training starts and is accessible in model-ready batches through the [tf.data.dataset](https://www.tensorflow.org/api_docs/python/tf/data/Dataset) module.

**Data Size and Attributes:**

The data contains the full body text of all Stack Overflow questions and answers along with metadata. The API pointer is updated quarterly. The metadata includes:

* Creation date
* Question title
* Question tags
* Question score
* Type (Question or Answer)

The data is split into train, validation, and test sets with:

* Train: 342,477 distinct users and 135,818,730 examples.
* Validation: 38,758 distinct users and 16,491,230 examples.
* Test: 204,088 distinct users and 16,586,035 examples.

**Data Location:**

The data is hosted by Kaggle and made available through the [tff.simulation.datasets](https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets) module in the [Tensorflow API](https://github.com/tensorflow/tensorflow). Stack Overflow owns the data and has released the data under the [CC BY-SA 3.0](https://creativecommons.org/licenses/by-sa/3.0/) license.

**Access Software:**

The [Tensorflow Python API](https://github.com/tensorflow/tensorflow) provides access to the data. Using Python, we will able to load, explore, and construct models using the Stack Overflow dataset. The [tf.data.dataset](https://www.tensorflow.org/api_docs/python/tf/data/Dataset) module makes loading and training efficient by generating data samples in batches rather than requiring pulling the entire dataset into memory. This will enable analysis and model development using standard Python tools and the Google Colab notebook environment.