DS-GA 1004 Big Data Final Project – Script Explanations and execution steps

Devansh Singh, ds6137@nyu.edu

Link to GitHub repository : https://github.com/nyu-big-data/final-project-nyu-big-data-project-ds6137

1. **data_subsample_split.py** : This script is used for subsampling from the complete dataset based on the fraction that has been passed in the arguments. It first considers only those users having at least 10 interactions and then subsamples users. It ultimately creates the final training, testing and validation datasets. The script stores these datasets in parquet format in the same directory as itself i.e. the directory where this python script is present and executed from. On Dumbo, it will store it in the users directory of the person who is executing this script. It stores these datasets as 'train_data_final.parquet', 'val_data_final.parquet' and 'test_data_final.parquet'.

   Usage: $ spark-submit data_subsample_split.py 0.05
   Here "0.05" is the size of fraction that we want to subsample.

   $ spark-submit data_subsample_split.py fraction_size

2. **training_testing_hyper.py** : This script is used for hyper parameter tuning on validation set ('val_data_final.parquet'), training the model and saving the best tuned ALS model.

   Usage:
   ```
   $ spark-submit training_testing_hyper.py path/training_data_final.parquet path/val_data_final.parquet path/model_name_to_be_stored
   ```

   I have used it on Dumbo as:
   ```
   $ spark-submit training_testing_hyper.py hdfs:/user/ds6137/train_data_final.parquet hdfs:/user/ds6137/val_data_final.parquet hdfs:/user/ds6137/als_model_tuned
   ```

3. **testing_for_tuned.py** : This script is used for evaluating the best tuned ALS model on the test dataset ('test_data_final.parquet'). It takes the location of the saved best tuned model and test dataset in its arguments.

   Usage:
   ```
   $ spark-submit testing_for_tuned.py path/best_tuned_model path/ test_data_final.parquet
   ```

   I have used it on Dumbo as:
   ```
   $ spark-submit testing_for_tuned.py hdfs:/user/ds6137/als_model_tuned hdfs:/user/ds6137/test_data_final.parquet
   ```

4. **extension.py** : This script is used for implementing the Fast search extension using the Annoy library. Before executing this script, annoy needs to be installed. To install, simply do **pip install --user annoy**.

   Usage:
   ```
   $ spark-submit extension.py path/ test_data_final.parquet path/best_tuned_model
   ```

   I have used it on Dumbo as:
   ```
   $ spark-submit extension.py hdfs:/user/ds6137/test_data_final.parquet hdfs:/user/ds6137/als_model_tuned
   ```

5. **training_normal.py :** This is the initial script that was created for training the model without any hyper parameter tuning. Default parameter values are used in the ALS model. It trains and saves the model.

   Usage:

   ```
   $ spark-submit training_normal.py path/training_data_final.parquet path/model_name_to_be_stored
   ```

   I have used it on Dumbo as:

   ```
   $ spark-submit training_normal.py hdfs:/user/ds6137/train_data_final.parquet hdfs:/user/ds6137/als_model_normal
   ```

6. **testing_normal.py :** This is the initial script that was created for evaluating the model on both validation and test datasets.

   Usage:

   ```
   $ spark-submit testing_normal.py path/als_model_normal path/val_data_final.parquet path/test_data_final.parquet
   ```

   I have used it on dumbo as:

   ```
   $ spark-submit testing_normal.py hdfs:/user/ds6137/als_model_normal hdfs:/user/ds6137/val_data_final.parquet hdfs:/user/ds6137/test_data_final.parquet
   ```