Code

The code can be found at https://github.com/PosterStevens/20Spring (https://github.com/PosterStevens/20Spring)

Code Architecture

models

Models are defined and stored in models folder.

- sequence2vector.py defines keras implementation of word2vector and motified word2vector
- recommendation_system.py defines one keras implementation of our recommendation model

dataset

function to load tables, modify tables, join tables and transform tables are stored in dataset folder

- public
  - join_dataset.py defines functions to join tables
  - load_dataset.py defines functions to load tables
  - select_data.py defines functions to subset tables
  - manager.py defines class to map ID into sparse label
  - 5_gram_sku2vec_v2.h5: pretrained SKU2vectors model
- private
  - user_map.csv: contains three columns: original ID, combined-attribute ID, label
  - sku_map.pkl: dictionary to map sku ID into label
  - user_encoder.pkl: sklearn.preprocessing.LabelEncoder object to map user combined-attribute ID into label

data

data, temp_data, output data and preprocessing file are stored in this folder

- data_preprocessing.cpp: defines functions to convert sorted table into n-gram data
- JD_click_data.csv: click request data of users
- JD_user_data.csv: user information
- JD_sku_data.csv: Stock Keeping Unit(SKU) information
- JD_order_data.csv: order data of users
- recommendation_X.npy X for training our recommendation model
- recommendation_Y.npy Y for training our recommendation model

root

main script includes preprocessing script, word2vector pretraining script and model training script are stored in this folder

- data_preprocessing.bat is a preprocessing batch file to convert sorted tables in data folder into n-gram data.
- main_sku.ipynb preprocesses n-gram data, then trains sku2vector model
- main_recommendation_system_training.ipynb trains recommendation model
- main_CNN_RNN_baseline.ipynb trains baseline RNN & CNN based model