Movie Recommender

Files Included

- api.py: The back-end web application
- dbhelper.py: The helper class for accessing database
- euclidean.py: Some helper functions for calculating euclidean distance metrics
- pearson.py: Some helper functions for calculating pearson corelation coefficient
- knn.py: Calculate all the recommendation result using kNN algorithm and store them into database. It will also evaluate the prediction result by utilizing RMSE method.
- svdRecommender.py: Calculate all the recommendation result using SVD algorithms and store them into database. It will also evaluate the prediction result by utilizing RMSE method.
- frontend/: The web pages
- dbmisc/: Some helper scripts for manipulating database, spliting dataset
- data/: The original dataset

Toolchains and Required Packages

Python version > 3.6

PostgreSQL version > 9.6

Web Server: NGINX or Apache

PyPI Packages Used

- psycopg2
- numpy
- scipy
- pandas
- sklearn
- flask

Preparing Data

In `data` directory, we have already prepared the dataset to you.

If you wish to prepare the data by your self, you can download the 1M MovieLens data from here. Unzip it and put all the `.dat` files into `data` directory under the project root directory (if it doesn't exist, create one). They must be converted into UTF-8 encoding. You can use `iconv` tool to convert encoding of a file:

```
iconv -f iso-8859-1 -t utf-8 ratings.dat > ratings.dat.conv
```

Don't forget to change the file names back to `.dat` after conversion.

Then you should run the `split_dat.py` (get it from `dbmisc` directory) to split the original rating data into `test.dat` and `train.dat`.

Populate Database

Install and configure PostgreSQL. Creating a new user and database for the project is recommended. Then from the project root directory, run `psql` to create database schemas:

```
psql -h localhost -U [db_name] < dbmisc/tables.sql
```

First copy populate_db.py from `dbmisc` directory to project root directory. Then change the database configuration in it, run it and wait for it complete running. The script will automatically insert all the data into database.

Running the Scripts

Before running the scripts, you must first specify correct database configuration in all the scripts mentioned below.

The `knn.py` script will calculate all the recommendation result using kNN algorithm and store them into database. It will also evaluate the prediction result by utilizing RMSE method.

The `svdRecommender.py` script will calculate all the recommendation result using SVD algorithms and store them into database. It will also evaluate the prediction result by utilizing RMSE method.

The scripts will run about 9 threads concurrently, and finish training in about 8 hours.

Check out the Recommendation Result

You can check out all the recommendation result by running raw query in database, or using our simple web applicaiton. You will also need to specify correct database configuration in `api.py`.

To run the web application, run the command below:

```
set FLASK_APP="api.py"
flask run
```

To use the front-end web application in browser, you will need to install a web server to host the web pages, and visit them from your browser.