CM50266 Applied Data Science

Lab 2: Movie Recommendations

Deadline

Set	31st Oct 2018
Due	16th Nov 2018, 5pm.
Marks	/10 (10% of overall unit mark.)

Data

For this lab you will use the MovieLens 100K Dataset consisting of 100,000 movie ratings. This can be found at:

http://grouplens.org/datasets/movielens/100k/

This is a cut-down version of the full dataset that contains 20 million ratings. Your objective is to use it to provide users with recommendations of films they may enjoy. The data is CSV format with a variety of separators used. The three most important files are:

u.user	Each line provides the details of one user.
u.item	Each line provides details of one movie.
u.item	Each line represents one rating of one movie by one user.
u.genre	The movie genres.

Movie IDs are consistent across all files. They are consistent with the full database and may therefore contain gaps.

Task1 (3 marks)

Write a Python program to parse the data files and identify the highest rated film in each genre. If more than one film in a genre scores the same rating, provide the complete list.

Task 2 (4 marks)

Extend your code using a user based collaborative filtering approach to make an individual recommendation to a user based on the set of movie ratings they have provided. It is suggested that you use k-nearest neighbour in order to identify the closest matches, however you are free to use other methods so long as you explain your chosen method. Experiment with changing the number of users that are considered similar to the target user. What if any impact does this have on the recommendations made? You should reserve some users as test subjects.

Task 3 (3 marks)

Extend your code from Task 2 to include the user details from u.user as part of the matching process. What if any difference does this make to the results? You will need to think carefully about how you convert the fields in u.user into suitable parameters.

Submission

Submission:	Moodle Assignment
Submit:	Report (PDF/Word) + ZIP containing code/images/data.

You should include a brief report that:

1. The methods used to achieve the two tasks.
2. Describes the code you've written and any issues you had developing it.
3. The results of your experimentation and any conclusions you draw from it.

The report should not be any longer than two pages. In addition to your report you should submit a ZIP file containing your code with instruction on how to run it.