Task 1

Command to run:

```
userName$ ./bin/spark-submit --class recommenderALS --master local[4]
Yu_Dong_task1+2.jar "ratings.csv" "testing_small.csv"
```

Output file:

This command will create a folder named “predictionsALS”, and the “part-00000” text file inside it is the predictions for all the (user, movie) combinations in the testing set.

Accuracy:

```
>=0 and <1: 637
>=1 and <2: 1706
>=2 and <3: 4629
>=3 and <4: 8250
>=4: 5034
RMSE = 1.1718064472176049
The total execution time taken is 18.72545199 sec.
```

Note: Due to the randomness of this algorithm, the predictions and RMSE will vary between 1.17 to 1.19.

Missing values handling:
For those movies that have never been rated by other users, except in the record in the testing dataset, or those movies that are not rated by the ALS algorithm, I impute the prediction by the average rating of that user.

Outlier ratings handling:
For those predicted value exceeds 5, I rounded down all of them to 5;
For those predicted value under 0, I rounded up all of them to 0.

Task 2

Description of the Algorithm:

I used User-based CF algorithm in this task. The basic process is:

1. Exclude those records in test set from the training set
2. Find co-rated items for each user pair
3. Calculate Pearson correlation for each user pair
4. For each user, find the top 2 similar users based on Pearson correlation
5. Make predictions based on the ratings of the 2 similar users
6. Impute the rating of un-predicted (user, movie) pairs by the average rating of that user

And I used the same missing value imputation and outlier handling principles as in the task1.

Command to run:

```
userName$ ./bin/spark-submit --class recommenderUserBased --master
local[4] Yu_Dong_task1+2.jar "ratings.csv" "testing_small.csv"
```

Output file:

This command will create a folder named “predictionsUserBased”, and the “part-00000” text file inside it is the predictions for all the (user, movie) combinations in the testing set.

Accuracy:

```
>=0 and <1: 19
>=1 and <2: 66
>=2 and <3: 2371
>=3 and <4: 13831
>=4: 3969
RMSE = 0.984278367793504
The total execution time taken is 61.495399044 sec.
```