COMMAND:

Task1:

add signature licence(if need):

zip -d /YOUR_PATH_TO_JAR_FILE/weinung_chao_task1.jar META-INF/*.RSA META-INF/*.DSA META-INF/*.SF

to run this program:

```
/YOUR_PATH_TO_SPARK /spark-1.6.1-bin-hadoop2.4/bin/spark-submit --packages
com.databricks:spark-csv_2.10:1.4.0 --class Weinung_Chao_task1
weinung_chao_task1.jar ratings.csv testing_small.csv
```

Task2:

add signature licence(if need):

zip -d /YOUR_PATH_TO_JAR_FILE/weinung_chao_task1.jar META-INF/*.RSA META-INF/*.DSA META-INF/*.SF

to run this program:

```
/YOUR_PATH_TO_SPARK /spark-1.6.1-bin-hadoop2.4/bin/spark-submit --packages
com.databricks:spark-csv_2.10:1.4.0 --class Weinung_Chao_task2
weinung_chao_task2.jar ratings.csv testing_small.csv
```

DESCRIPTION:

1. I use user-based CF
2. I handle the missing data by using imputation(mean from other data)
3. If data over 5 or less than 0, they will be count as 5 or 0
4. The accuracy for task1 and task2 is

```
>=0 and <1: 627
>=1 and <2: 1687
>=2 and <3: 4401
>=3 and <4: 8051
>=4: 5490
RMSE: 1.1115867219665387
The total execution time taken is44.816647664sec.

>=0 and <1: 52
>=1 and <2: 403
>=2 and <3: 2742
>=3 and <4: 10444
>=4: 6615
RMSE: 0.9276192215500006
The total execution time taken is298.349467179sec.
```