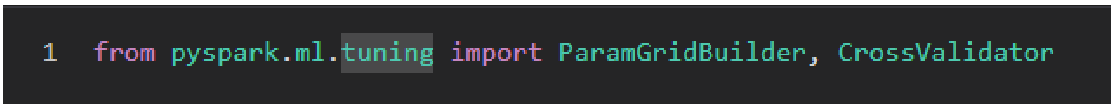
CCBDA　HW#5

- How to run：使用 jupyter-notebook 或 google colab 逐格執行即可

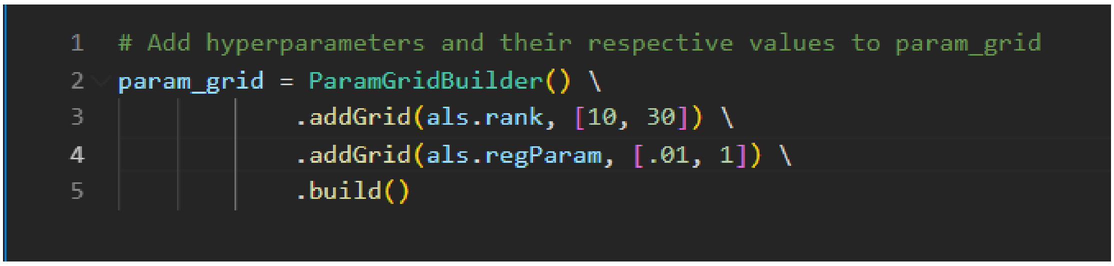
- Method:

使用 model 為 pyspark.ml.recommendation 中的 ALS algorithm.

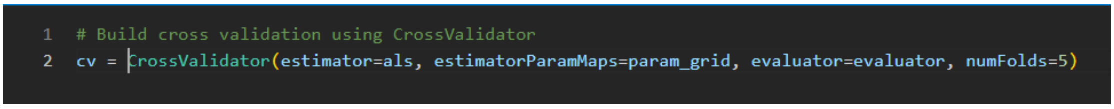


```python
from pyspark.ml.tuning import ParamGridBuilder, CrossValidator
```

在一開始的時候我嘗試使用 CrossValidator，想找出 best model，使用的參數為

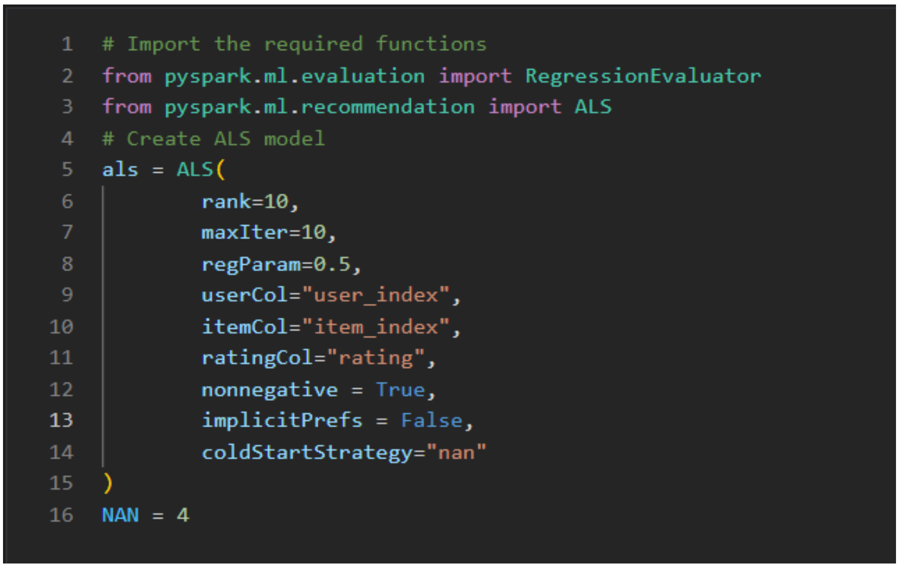


```python
# Add hyperparameters and their respective values to param_grid
param_grid = ParamGridBuilder() \
            .addGrid(als.rank, [10, 30]) \
            .addGrid(als.regParam, [.01, 1]) \
            .build()
```



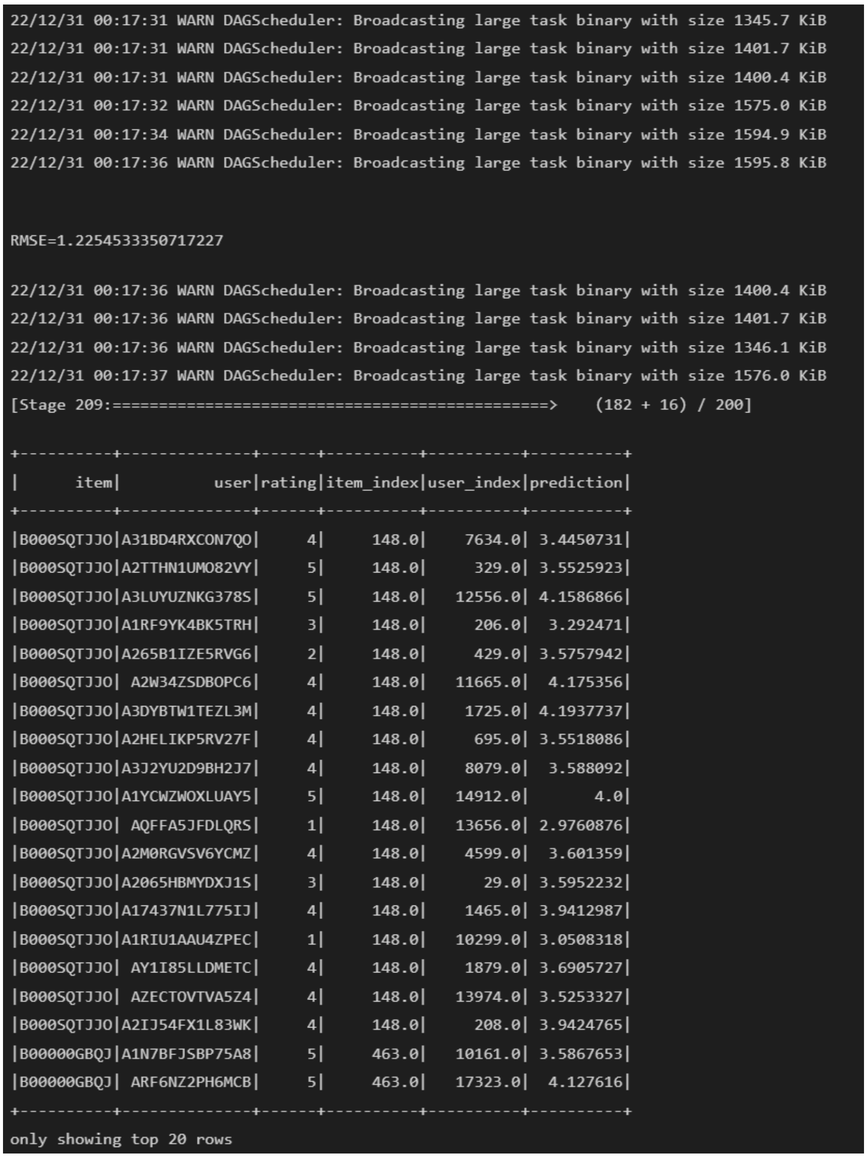
```python
# Build cross validation using CrossValidator
cv = CrossValidator(estimator=als, estimatorParamMaps=param_grid, evaluator=evaluator, numFolds=5)
```

但執行結果的 RMSE 結果都偏高，最後自己手動 tune 出以下的參數才得到最好的 RMSE ： 1.2254



```python
# Import the required functions
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.recommendation import ALS
# Create ALS model
als = ALS(
        rank=10,
        maxIter=10,
        regParam=0.5,
        userCol="user_index",
        itemCol="item_index",
        ratingCol="rating",
        nonnegative = True,
        implicitPrefs = False,
        coldStartStrategy="nan"
)
NAN = 4
```

執行結果



```
22/12/31 00:17:31 WARN DAGScheduler: Broadcasting large task binary with size 1345.7 KiB
22/12/31 00:17:31 WARN DAGScheduler: Broadcasting large task binary with size 1401.7 KiB
22/12/31 00:17:31 WARN DAGScheduler: Broadcasting large task binary with size 1400.4 KiB
22/12/31 00:17:32 WARN DAGScheduler: Broadcasting large task binary with size 1575.0 KiB
22/12/31 00:17:34 WARN DAGScheduler: Broadcasting large task binary with size 1594.9 KiB
22/12/31 00:17:36 WARN DAGScheduler: Broadcasting large task binary with size 1595.8 KiB


RMSE=1.2254533350717227

22/12/31 00:17:36 WARN DAGScheduler: Broadcasting large task binary with size 1400.4 KiB
22/12/31 00:17:36 WARN DAGScheduler: Broadcasting large task binary with size 1401.7 KiB
22/12/31 00:17:36 WARN DAGScheduler: Broadcasting large task binary with size 1346.1 KiB
22/12/31 00:17:37 WARN DAGScheduler: Broadcasting large task binary with size 1576.0 KiB
[Stage 209:=============================================>    (182 + 16) / 200]

+----------+--------------+------+----------+----------+----------+
|      item|          user|rating|item_index|user_index|prediction|
+----------+--------------+------+----------+----------+----------+
|B000SQTJJO|A31BD4RXCON7QO|     4|     148.0|    7634.0| 3.4450731|
|B000SQTJJO|A2TTHN1UMO82VY|     5|     148.0|     329.0| 3.5525923|
|B000SQTJJO|A3LUYUZNKG378S|     5|     148.0|   12556.0| 4.1586866|
|B000SQTJJO|A1RF9YK4BK5TRH|     3|     148.0|     206.0|  3.292471|
|B000SQTJJO|A265B1IZE5RVG6|     2|     148.0|     429.0| 3.5757942|
|B000SQTJJO| A2W34ZSDBOPC6|     4|     148.0|   11665.0|  4.175356|
|B000SQTJJO|A3DYBTW1TEZL3M|     4|     148.0|    1725.0| 4.1937737|
|B000SQTJJO|A2HELIKP5RV27F|     4|     148.0|     695.0| 3.5518086|
|B000SQTJJO|A3J2YU2D9BH2J7|     4|     148.0|    8079.0|  3.588092|
|B000SQTJJO|A1YCWZWOXLUAY5|     5|     148.0|   14912.0|       4.0|
|B000SQTJJO| AQFFA5JFDLQRS|     1|     148.0|   13656.0| 2.9760876|
|B000SQTJJO|A2M0RGVSV6YCMZ|     4|     148.0|    4599.0|  3.601359|
|B000SQTJJO|A2065HBMYDXJ1S|     3|     148.0|      29.0| 3.5952232|
|B000SQTJJO|A17437N1L775IJ|     4|     148.0|    1465.0| 3.9412987|
|B000SQTJJO|A1RIU1AAU4ZPEC|     1|     148.0|   10299.0| 3.0508318|
|B000SQTJJO| AY1I85LLDMETC|     4|     148.0|    1879.0| 3.6905727|
|B000SQTJJO| AZECTOVTVA5Z4|     4|     148.0|   13974.0| 3.5253327|
|B000SQTJJO|A2IJ54FX1L83WK|     4|     148.0|     208.0| 3.9424765|
|B00000GBQJ|A1N7BFJSBP75A8|     5|     463.0|   10161.0| 3.5867653|
|B00000GBQJ| ARF6NZ2PH6MCB|     5|     463.0|   17323.0|  4.127616|
+----------+--------------+------+----------+----------+----------+
only showing top 20 rows
```

- Reference:

Collaborative Filtering - Spark 2.2.0 Documentation (apache.org)

Building a Recommendation System with Spark ML and Elasticsearch | by Lijo Abraham | Towards Data Science

https://miro.medium.com/max/828/1*D34HqTvyzuvrCerHWCZSHQ.webp