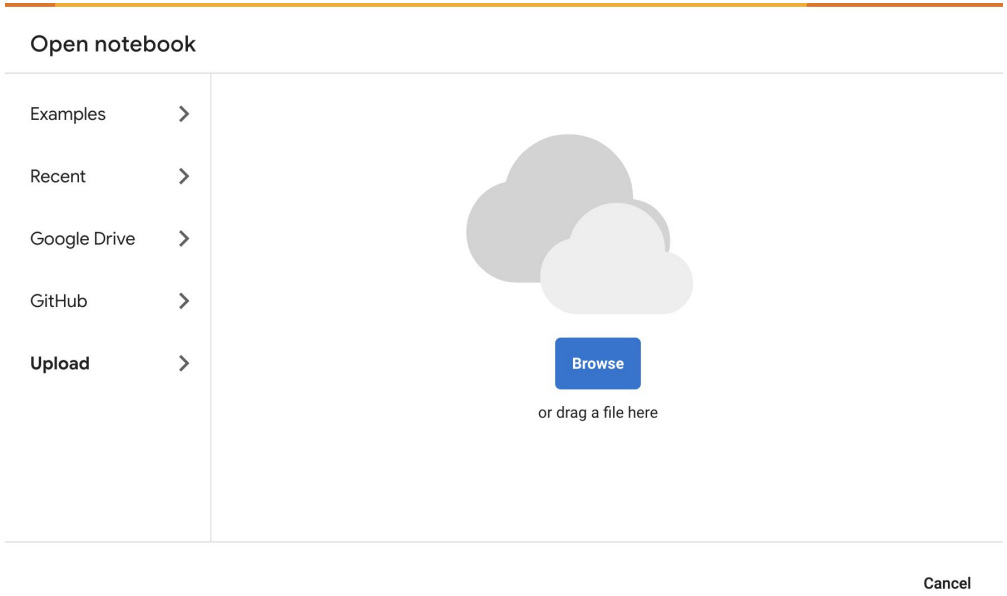
Project: Movie Recommendation with MLlib - Collaborative Filtering

Emily Weng 20016

Step 1: Download the Pyspark code (ipynb)

- Download at this link:
  - https://github.com/snehalnair/als-recommender-pyspark/blob/master/Recommendation_Engine_MovieLens.ipynb
- Upload it into Google Colab

Step 2: Upload the ipynb file to your Colab



Step 3: Experiment Pyspark code (ipynb) by modifying the ipynb file

- Modifications were made in this cell since it was running for too long

```
#Fit cross validator to the 'train' dataset
model = cv.fit(train)

#Extract best model from the cv model above
best_model = model.bestModel
```

Continue

- These were the modifications made:

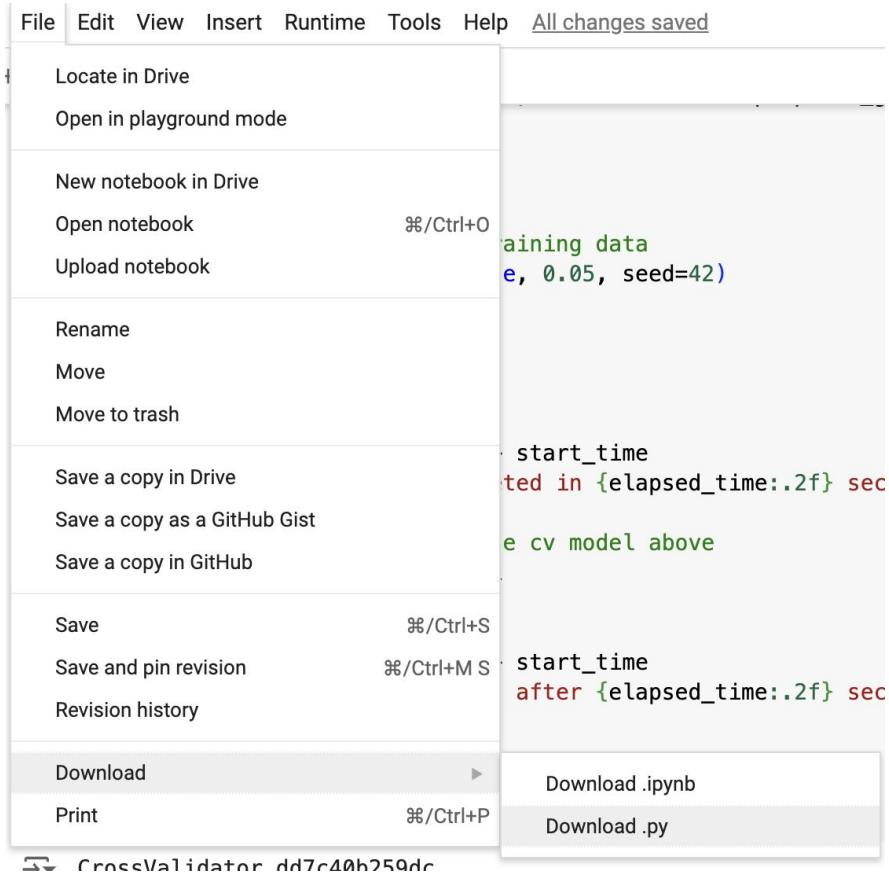
```
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.recommendation import ALS
from pyspark.ml.tuning import ParamGridBuilder, CrossValidator
import time
```

```python
# Use a smaller subset of the training data
train_subset = train.sample(False, 0.05, seed=42)

start_time = time.time()

try:
    model = cv.fit(train_subset)
    elapsed_time = time.time() - start_time
    print(f"Model fitting completed in {elapsed_time:.2f} seconds")

    # Extract best model from the cv model above
    best_model = model.bestModel

except Exception as e:
    elapsed_time = time.time() - start_time
    print(f"Model fitting failed after {elapsed_time:.2f} seconds")
    print(f"Error: {e}")
```

Continue:

- The modifications can reduces the subset size and the number of folds lessen potential memory and processing issues.
- Also ensures that the train dataset does not contain any anomalies or null values.

Step 4: Save the modified ipynb file as py format



Step 6: Save the modified ipynb file as HTML format which can be used on Step 9 of this project

```
%%shell
jupyter nbconvert --to html PATH_TO_FILE
```

```
%%shell
jupyter nbconvert --to html /content/Modified_Recommendation_Engine_MovieLens.ipynb
```

```
[NbConvertApp] Converting notebook /content/Modified_Recommendation_Engine_MovieLens.ipynb
[NbConvertApp] Writing 630330 bytes to /content/Modified_Recommendation_Engine_MovieLens.h
```

Step 7: Run the py file saved at Step 3.4 on GCP

```
driverControlFilesUri: gs://dataproc-staging-asia-eas
5bea10515fe0fe351/
driverOutputResourceUri: gs://dataproc-staging-asia-e
d75bea10515fe0fe351/driveroutput
jobUuid: 177d481d-e096-32f6-b03e-13f6bf837ac1
placement:
  clusterName: cluster-2335
  clusterUuid: a0c4b3d5-2e2d-498d-a9af-53eefd4e0980
pysparkJob:
  mainPythonFileUri: gs://py1/movie_rec.py
reference:
  jobId: b3ae16151e514d75bea10515fe0fe351
  projectId: cs570-project3-426016
status:
  state: DONE
  stateStartTime: '2024-07-26T06:36:10.924987Z'
statusHistory:
- state: PENDING
  stateStartTime: '2024-07-26T06:33:44.147203Z'
- state: SETUP_DONE
  stateStartTime: '2024-07-26T06:33:44.191777Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2024-07-26T06:33:44.423726Z'
yarnApplications:
- name: MovieLensRecommendation
  progress: 1.0
  state: FINISHED
  trackingUrl: http://cluster-2335-m.asia-east1-a.c.c
```

Step 8: Remember to delete and close everything on GCP

Github link:

https://github.com/emilywengster/sfbu/tree/68e5bd42fe250f0cb46cf4185be6b921003e0a7e/Cloud%20Computing/Machine%20Learning/%20Movie%20Recommendation%20System