Cloud Computing and Big Data Analytics 2022 Fall
Homework 5: Recommendation System with PySpark

BASIC LAB

TA：曾偉倫

Email: wltseng.ee06@nycu.edu.tw

NYCU

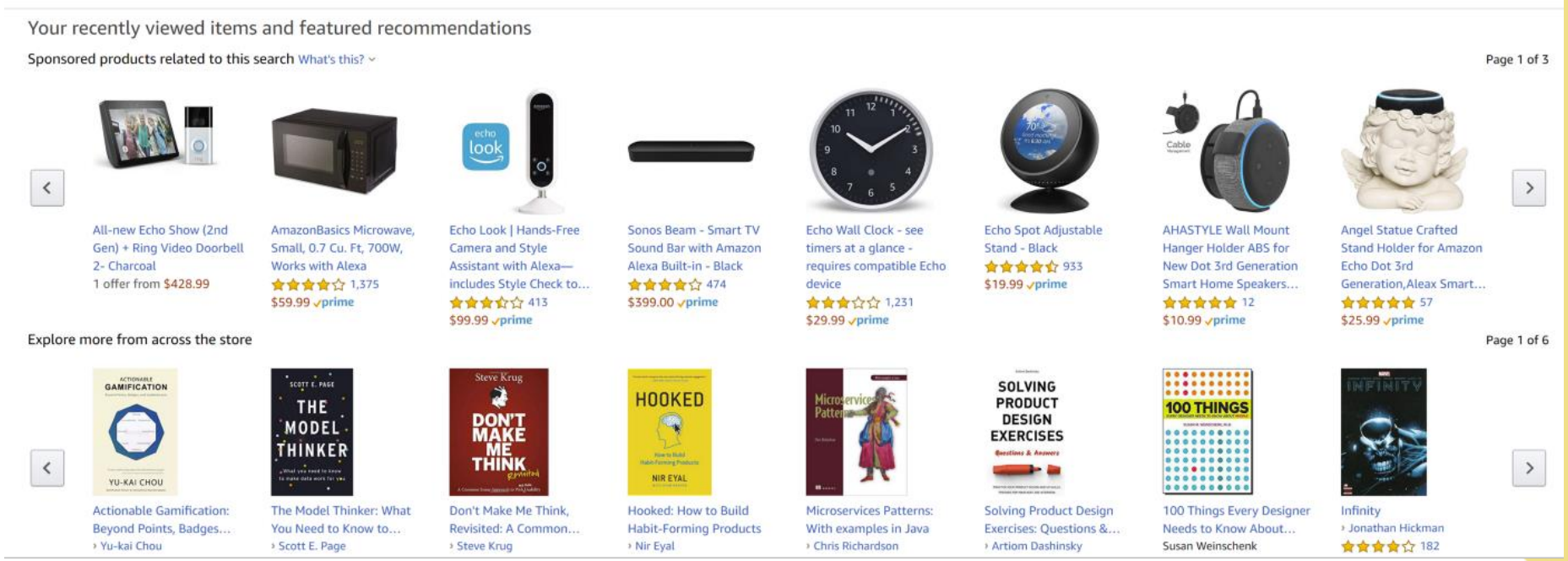
Outline

- Introduction
- Problem Description
- Dataset
- Grading Policy
- Submission
  - Kaggle Competition
  - E3
- Timeline

Introduction

BASIC LAB

- Recommendation system makes e-commerce platform better.



Problem Description

BASIC LAB

- Given : user-item pairs
- Objective: Predict ratings
- Limitation:
  - Build your recommendation system with PySpark framework.
  - PySpark ML-lib or Other Model-based methods

NYCU

Dataset

BASIC LAB

- File List
  - train_student.csv
    - item,user,rating
  - train_meta.csv
    - user,item,rating with additional with additional review content
  - test_public.csv (50% of testing set)
    - item,user,rating
  - test_private.csv (50% of testing set)
    - item,user
  - test_all.csv
    - item,user

Grading Policy

BASIC LAB

- Kaggle Competition Rank
  - Over baseline:
    - Top 20% : 100 pts
    - 20% ~ 30%: 97 pts
    - 30% ~ 40%: 95 pts
    - 40% ~ 50%: 93 pts
    - 50% ~ 60%: 91 pts
    - 60% ~ 70%: 89 pts
    - 70% ~ 80%: 87 pts
    - 80% ~ 90%: 85 pts
    - 90% ~ 100%: 83 pts(> baseline)
  - Valid submission, but $\leq$ baseline benchmark: 75 pts
  - Invalid submission, but submit code & readme to E3: 50 pts
    (Student must write down what you have done and tell TA what Algorithm / Method you plan to implement.)

Submission

- Kaggle
  - Register with invitation link: <URL>
  - Team name: <student_ID> .
  - .csv file, containing two columns
    - Column name: U_I,rating (I is capital "i")
    - U_I: user-item pairs, unique identifier (string)
    - rating: predicted rating (double)
    - Make sure your total line number in submission file is 1 (column name) +83799 (U_I,rating pair).
    - 10 submissions per day

```
U_I = str(user) +"_" + str(item)
user = 12345
item = CXGZE

U_I = 12345_CXGZE
```

Example submission:

```
U_I,rating
12345CXGZE,5
12345CXGZK,1
…
```

BASIC LAB

Submission

BASIC LAB

- E3
  - <studet_ID>.zip : Source code (in one zip file)
  - <studet_ID>.pdf : Readme File
    - 1 page A4
    - Tell TA how to execute your code
    - Briefly explain your method
    - Reference

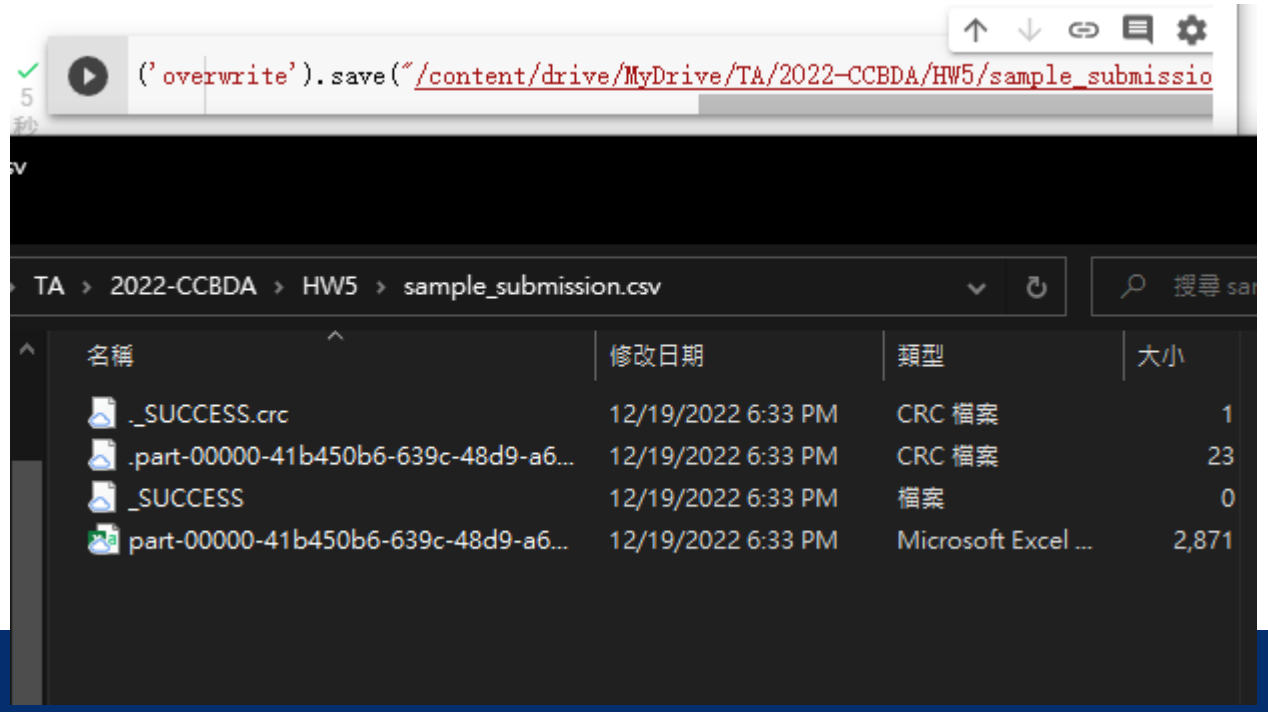
Timeline

- 12/13: HW5 release
- 12/21-12/26: Kaggle Competition Submission (10 times per day)
- 12/30 : E3 Submission Deadline (before 23:59:59)

- Any problems? Contact TA:
  - TA：曾偉倫
  - Email: wltseng.ee06@nycu.edu.tw
  - TA Hour needs to make an appointment

Appendix

- PySpark will output csv file in the following screenshot.
  - Please select the csv file in your output path.



Appendix

BASIC LAB

- Useful example project:
  - Building a Recommendation System with Spark ML and Elasticsearch | by Lijo Abraham | Towards Data Science
  - PySpark Recommender System with ALS | Towards Data Science