INF553 Foundations and Applications of Data Mining

Spring 2020

Competition Project

Deadline: May. 6th 11:59 PM PST

1. Overview of the Competition Project

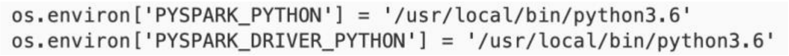
In this competition project, you need to build a recommendation system (e.g., a hybrid recommendation systems) to provide accurate predictions.

2. Competition Requirements

2.1 Programming Language and Library Requirements

a. **You must use Python to implement the competition project**. You can use the Python libraries that are available on the Vocareum.

b. You can re-use **your own code** in assignment 3. If you want to use Spark, please specify the following environment in your code:



```
os.environ['PYSPARK_PYTHON'] = '/usr/local/bin/python3.6'
os.environ['PYSPARK_DRIVER_PYTHON'] = '/usr/local/bin/python3.6'
```

2.2 Programming Environment

Python 3.6 and Spark 2.3.0

We will use Vocareum to automatically run and grade your submission. You must test your scripts on **the local machine** and **the Vocareum terminal** before submission.

2.3 Write your own code

Do not share code with other students!!

For this assignment to be an effective learning experience, you must write your own code! We emphasize this point because you will be able to find Python implementations of some of the required functions on the web. Please do not look for or at any such code!

TAs will combine all the code we can find from the web (e.g., Github) as well as other students' code from this and other (previous) sections for plagiarism detection. We will report all detected plagiarism to the university.

3. Yelp Data

In this assignment, we generated the review data from the original Yelp datasets with some filters, such as the condition: “state” == “CA”. We randomly took 80% of the data for training, 10% of the data for testing, and 10% of the data as the blind dataset. **We do not share the blind dataset.**

You can access the files (a-e) under the fixed directory on the Vocareum: *resource/asnlib/publicdata/*

a. train_review.json

b. user.json – user metadata

c. business.json – business metadata, including locations, attributes, and categories

d. user_avg.json – containing the average stars for the users in the train dataset

e. business_avg.json – containing the average stars for the businesses in the train dataset

Besides, the Google Drive provides the above files (a-e) and the following testing files (f and g)
https://drive.google.com/open?id=1ss6Tq-hxeRfyst8u-n8Tx8Ykn1jD8GB8 (USC email only)

f. test_review.json – containing only the target user and business pairs for the prediction task

g. test_review_ratings.json – containing the ground truth rating for the testing pairs

4. Task (5 points)

You need to submit the following files on Vocareum: (all lowercase)

a. [REQUIRED] Two Python scripts: **train.py**, **predict.py**

b. [REQUIRED] Model files/folders (you can name them yourself)

c. [REQUIRED] One PDF file: **model.pdf** (describing your model in 200 words)

d. You can include other Python scripts to support your programs (e.g., callable functions).

4.1 Task description

In the competition project, you will build a recommendation system with the provided datasets on the Vocareum and use the model(s) to predict the ratings for a given pair of user and business.

4.2 Execution commands

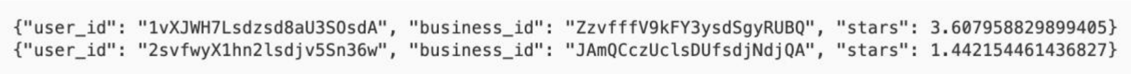
Training commands: $ python3 train.py

Predicting commands: $ python3 predict.py <test_file> <output_file>

Param	<test_file>: containing the target pairs for prediction, e.g., test_review.json <output_file>: the prediction results

4.3 Output format:

You must write a target pair and its prediction in the JSON format using **exactly the same tags as the example in Figure 1**. Each line represents for a predicted pair of (“user_id”, “business_id”).



```
{"user_id": "1vXJWH7Lsdzsd8aU3SOsdA", "business_id": "ZzvfffV9kFY3ysdSgyRUBQ", "stars": 3.607958829899405}
{"user_id": "2svfwyX1hn2lsdjv5Sn36w", "business_id": "JAmQCczUclsDUfsdjNdjQA", "stars": 1.442154461436827}
```

Figure 1: An example output in JSON format

4.4 Grading

You **MUST** submit **your model file(s) (0.5pt)** and **a PDF file (0.5pt)** to describe how you design/build your model(s) in 200 words. We will compare your prediction results against the ground truth. You **MUST** output the predictions for **ALL** the target pairs **(0.5pt for the test dataset and 0.5pt for the blind dataset).** We use RMSE (Root Mean Squared Error) to evaluate the performance:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}(Pred_i - Rate_i)^2}$$

Where $Pred_i$ is the prediction for business i and $Rate_i$ is the true rating for business $i. n$ is the total number of the user and business.

The execution time of the training process on Vocareum should be **less than 1,200 seconds**. The execution time of the predicting process on Vocareum should be **less than 300 seconds**. The table below shows the **CURRENT RMSE** for the prediction task.

	Test set	Blind set
RMSE	1.23	1.23

To get the full points for the competition project, **your RMSE should be lower than TAs’ (1.5pt for the test dataset and 1.5pt for the blind dataset)**. TAs will continuously improve their systems and update the accuracy on the Piazza. They will fix their results **one week before the competition due (i.e., Apr. 29th 11:59 PM PST)**.

You will also compete your model performance with other students in the competition project. You can check the **Leaderboard** on the Vocareum to see the accuracy results from other students (anonymous). On the Vocareum, you will be simply ranked by the sum of the RMSE of test dataset and blind dataset. The final submission with the highest accuracy (**on both dataset**) will receive extra 3 points on the final grade. The second place will receive extra 2 points. The third one will receive extra 1 point. **Note that if your model performs the best only on one of the datasets, you will not be considered as the winners.**

5. About Vocareum

a. Your code can directly access the datasets under the directory: ../resource/asnlib/publicdata/

b. You should upload the required files under your workspace: work/

c. You must test your scripts on both the local machine and the Vocareum terminal before submission.

d. During submission period, the Vocareum will run predict scripts and evaluate the prediction results for both test and blind sets.

e. During grading period, the Vocareum will run both train and predict scripts. If the training or predicting process fail to run, you can get 50% of the score only if the submission report shows that your submitted models or results are correct.

f. You will receive a submission report after Vocareum finishes executing your scripts. The submission report should show **the accuracy information** for each task.

g. The total execution time of submission period should be less than 600 seconds. The execution time of grading period need to be less than 1800 seconds.

h. Please start your assignment early! You can resubmit any script on Vocareum. We will only grade on your last submission.

6. Grading Criteria

(% penalty = % penalty of possible points you get)

a. You cannot use late day for the competition project. Late submission is not allowed.

b. There is no regrading. Once the grade is posted on the Blackboard, we will only regrade your assignments if there is a grading error. No exceptions.