Welcome to the competition reserved to the students of the Recommender Systems course in Politecnico di Milano.

Please read carefully till the end of the page!

Goal

The application domain is TV shows recommendation. The datasets we provide contains both interactions of users with TV shows, as well as features related to the TV shows. The main goal of the competition is to discover which previously unseen items (TV shows) a user will interact with.
Each TV show (for instance, "The Big Bang Theory") can be composed by several episodes (for instance, episode 5, season 3) but the data does not contain the specific episode, only the TV show. If a user has seen 5 episodes of a TV show, there will be 5 interactions with that TV show. The goal of the recommender system is not to recommend a specific episode, but to recommend a TV show the user has not yet interacted with.

Description

The datasets includes around 1.8M interactions, 41k users, 27k items (TV shows) and two features: the TV shows length (number of episodes or movies) and 4 categories. For some user interactions the data also includes the impressions, representing which items were available on the screen when the user clicked on that TV shows.
The training-test split is done via random holdout, 85% training, 15% test.
The goal is to recommend a list of 10 potentially relevant items for each user. MAP@10 is used for evaluation. You can use any kind of recommender algorithm you wish e.g., collaborative-filtering, content-based, hybrid, etc. written in Python. Note that the impressions can be used to improve the recommendation quality (for example as additional features, a context, to estimate/reduce the recommender bias or as a negative interaction for the user) but are not used in any of the baselines.

Score

Each team will receive a final score according to the quality of recommendations computed on the private leaderboard.

The score is computed with the following formula:

$$score=baseline\_points+standing\_points+team\_bonus$$score=baseline_points+standing_points+team_bonus

Maximum score is 27 points (28 points for one-person team).

Baseline Points

You are provided with 5 baselines (the Random baseline is not counted). Each baseline is computed with a different algorithm. If you are able to do better than *n* baselines, you will receive a bonus score that adds to your final score.

Better than

- B1: baseline_points = 1
- B2: baseline_points = 2
- B3: baseline_points = 4

- B4: baseline_points = 6
- B5: baseline_points = 8

Maximum baseline score is 8 points.

Standing Points

At each deadline *i*, standing points will be assigned by using the following formula:

$$s_i = 19 - 19 \cdot \log\_2\left(r_i - 1 N_{\text{teams}} - 1 + 1\right) \text{si=19−19·log\_2(ri−1Nteams−1+1)}$$

where

$$N_{\text{teams}} = \text{number of teams} \text{Nteams=number of teams}$$

and

$$r_i = \text{ranking of the team in the leaderboard at deadline } i = 1 \ldots N_{\text{teams}} \text{ri=ranking of the team in the leaderboard at deadline i=1...Nteams}$$

The final standing points are computed with a weighted average over the deadlines with the following formula:

$$\text{standing\_points} = s_1 + 9 \cdot s_2 10 \text{standing\_points=s1+9·s210}$$

Maximum standing score is 19 points.

Team bonus

Single-person teams receive one point of bonus

$$t = \begin{cases} 0 & \text{for two-person teams} \\ 1 & \text{for one-person teams} \end{cases} \text{t={0for two-person teams1for one-person teams}}$$

Notes

Attention: Results on the public leaderboard are computed on a different subset of the test set, so they may differ from the private one.

Enrolling to the Competition

When registering on Kaggle, you must use your codice persona, 8 digits, as both your Display Name and Team Name.

If you use a different display or team name, your mark will not be registered.
If a team is composed by two members, the team name and display name must be *xxxxxxxx_yyyyyyyy*, where *xxxxxxxx* and *yyyyyyyy* are the codice persona of the two members.
Example: if two students with codice persona *10099001* and *20044002* merge into a team, the team name and display name must be *10099001_20044002*.

Team Merging

Team merging won't be allowed after the first deadline.

Attention: At the end of the competition, we will evaluate the activity and contributions of each team member. If we decide that a member has provided only a minimal contribution, we reserve the right to reduce or cancel his/her mark and, eventually, to add a bonus to the mark of the other member.

Team Splitting

Team splitting is not allowed, unless you cancel your Kaggle account and create a new account with the same email address. In this case, you will lose all of your previous submissions and the related points.

Deadlines

Deadlines will be on the following dates (at 23.59 CET):

- Deadline 1 (intermediate deadline): 15 December (best submission in private)
- Deadline 2 (final deadline): 15 January (the best two submissions in public are used for private)

Evaluation

The evaluation metric for this competition is MAP@10.

MAP@10

The average precision at 10, for a user is defined as:

$$\mathrm{AP@10}=\sum_{k=1}^{10} P(k)\cdot \mathrm{rel}(k) \min(m,10) \quad \text{AP@10=}\sum\text{k=110P(k)}\cdot\text{rel(k)min(m,10)}$$

where *P(k)* is the precision at cut-off *k*, rel(*i*) is 1 if item in position *k* is relevant, 0 otherwise, and *m* is the number of relevant items in the test set.

The mean average precision for N users at position 10 is the average of the average precision of each user, i.e.,

$$\mathrm{MAP@10}=\sum_{u=1}^{N} \mathrm{AP@10}_u N \quad \text{MAP@10=}\sum\text{u=1NAP@10uN}$$

Submission Format

The submission format is defined in the file sample_submission.csv.

IMPORTANT: All files are comma-separated (columns are separated with ',').
Also the submission file must be comma-separated.

- **interactions_and_impressions.csv** : Contains the training set, describing implicit preferences expressed by the users.
    - **user_id** : identifier of the user

- **item_id** : identifier of the item (TV series)
- **impression_list** : string containing the items that were present on the screen when the user interacted with the item in column item_id. Not all interactions have a corresponding impressions list.
- **data** : "0" if the user viewed the item, "1" if the user opened the item details page.

Note that there are multiple interactions between the same user and item when a user watches multiple episodes of a TV series (if a user has watched 5 episodes there will be 5 interactions with that item_id).

- **data_ICM_length.csv** :
  Contains the *number of episodes* of the items. TV series may have multiple episodes.
  - **item_id** : identifier of the item
  - **feature_id** : identifier of the feature, only one value (0) exists since this ICM only contains the feature "length"
  - **data** : number of episodes. Some values may be 0 due to incomplete data.
- **data_ICM_type.csv**:
  Contains the *type* of the items. An item can only have one type.
  All types are anonymized and described only by a numerical identifier.
  - **item_id** : identifier of the item
  - **feature_id** : identifier of the type
  - **data** : "1" if the item is described by the type
- **data_target_users_test.csv**:
  Contains the ids of the users that should appear in your submission file.
  The submission file should contain all and only these users.
- **sample_submission.csv**:
  A sample submission file in the correct format: [user_id],[ordered list of recommended items].
  Be careful with the spaces and be sure to recommend the correct number of items to every user.
  IMPORTANT: first line is mandatory and must be properly formatted.

```
user_id,item_list
1,0 1 2 3 4 5 6 7 8 9
2,0 1 2 3 4 5 6 7 8 9
[ . . . ]
```