ML challenge Yandex Cup 2021 (RecSys)

14 окт 2021, 17:17:16
старт: 20 сен 2021, 10:00:00
финиш: 19 окт 2021, 16:00:00
до финиша: 4д. 22ч.
начало: 20 сен 2021, 10:00:00
конец: 19 окт 2021, 16:00:00
длительность: 29д. 6ч.

A. RecSys

Иннокентий — москвич в пятом поколении.

Блуждание между тремя тополями на Плющихе, ежемесячное посещение Мавзолея, кормление тапиров в Московском зоопарке — всё это в крови Иннокентия.

Помимо этого, Иннокентий очень любит поесть. Как типичный среднестатистический москвич, Иннокентий посещает исключительно рестораны со средним чеком от 5000 рублей, любит устрицы и хорошие стейки зернового откорма, избегает кафе с шаурмой — боится, что его нечаянно накормят шавермой. Будучи добропорядочным горожанином, Иннокентий исправно оставляет на *Яндекс.Картах* отзывы на посещенные рестораны, живописно описывая свой гастрономический опыт.

Однажды Иннокентий понимает, что за пределами МКАД лежит целый мир, дикий и неизведанный. Он решается на самое опасное и рискованное путешествие в своей жизни — посещение Санкт-Петербурга. Иннокентий выбрал поезд на *Яндекс.Расписаниях*, забронировал отель на *Яндекс.Путешествиях* и отправился в путь.

Выйдя утром из вагона поезда «Красная Стрела», Иннокентий опасливо огляделся и решил хорошо подкрепиться. Дело за малым — подобрать хороший ресторан, удовлетворяющий его утонченным вкусам. Рекомендации друзей оказались бесполезны, ведь все они москвичи и плохо разбираются в петербургских гастрономических трендах. Давайте поможем Иннокентию найти вкусную еду.

В этой задаче вам предстоит построить рекомендательную систему, которая предложит пользователям *Яндекс.Карт* соответствующие их вкусу кафе, бары и рестораны в неродном городе: москвичам — в Санкт-Петербурге, а петербуржцам — в Москве.

В качестве данных используйте анонимизированную информацию о реальных отзывах и оценках, оставляемых пользователями Яндекс.Карт на заведения общепита Москвы и Санкт-Петербурга, и различную информацию о самих заведениях.

В частности, каждый отзыв содержит множество аспектов (упомянутые в отзыве блюда, особенности и т. п.), извлеченных из отзыва с помощью NLP-алгоритма. Для заданного множества москвичей и петербуржцев нужно предсказать, какие заведения в неродном городе они посетят, оставив при этом положительный отзыв с оценкой 4 или 5.

Baseline к задаче с метрикой, train-test split и очевидными решениями можно найти по ссылке.

Архив с данными можно найти здесь.

Формат ввода

Обучающее множество собрано за X = 1217 дней, тестовое множество – за последующие Y = 107 дней.

reviews.csv

В этом файле дана информация об отзывах и оценках, оставленных некоторым множеством жителей Москвы и Санкт-Петерубрга в течение обучающего периода:

```
user_id,org_id,ts,rating,aspect_ids
18a7276b,14e1b7bb,120,4,2 4 23
...
```

- user_id: идентификатор пользователя
- org_id: идентификатор организации
- ts: время отзыва (в днях от начала обучающего периода)
- rating: поставленная оценка
- aspect_ids: набор упомянутых в тексте отзыва аспектов.

organisations.csv

Информация об организациях:

```
org_id,city,average_bill,rubric_id,avg_rating,feature_ids
14e1b7bb,msk,2000,6,4.3,3 5 14 28
0ed69bff,spb,1500,2,4.8,2 5 6
...
```

- org_id: идентификатор организации
- city_id: город организации
- average_bill: средний чек в рублях (округленный с точностью до 500 рублей)
- avg_rating: средний рейтинг (в том числе с учетом не перечисленных в файле reviews.csv отзывов и оценок)
- rubric_id: рубрика опрганизации
- feature_ids: набор известных особенностей данной организации.

users.csv

Информация о городе проживания пользователя:

```
user_id,city
18a7276b,msk
270cc9fee,spb
4bf7ffc,msk
...
```

aspects.csv

Описание извлекаемых из отзывов аспектов. Множество аспектов извлекается из отзыва с помощью NLP-алгоритма и может быть неточным.

```
aspect_id,aspect_name
1,Бургеры
2,Кофе
3,Интерьер
4,Веранда
5,Устрицы
...
```

features.csv

Описание особенностей организаций. Как правило, множество особенностей организации заполняется ее владельцем и может быть неточным.

```
feature_id,feature_name
1,Wi-Fi
2,Доставка
3,Кофе с собой
...
```

rubrics.csv

Описание рубрик организаций:

```
rubric_id,rubric_name
1,Ресторан
2,Кафе
3,Бар
...
```

test_users.csv

Множество пользователей, для которых необходимо сделать предсказание:

```
user_id
270cc9fee
4bf7ffc
...
```

Формат вывода

answers.csv

Для каждого пользователя из файла **test_users.csv** необходимо приложить список из не более чем 20 организаций, относящихся к городу, отличному от города проживания пользователя.

```
user_id,target
270cc9fee,14e1b7bb 169a320c 75e004ad
4bf7ffc,0ed69bff f13d2de1 c5b05e22
...
```

Примечания

В качестве метрики используется **MNAP@20**.

$$\mathrm{MNAP@20} = \frac{1}{|U|}\sum_{u \in U}\frac{1}{\min(n_u, 20)}\sum_{i=1}^{20} r_u(i) p_u@i,$$

$$p_u@k = \frac{1}{k}\sum_{i=1}^{k} r_u(i),$$

где:

- $r_u(i)$ — оставил ли в течение тестового периода пользователь u оценку 4 или 5 организации, рекомендованной ему на месте i (1 либо 0),
- n_u — количество организаций из противоположного города, которым пользователь u поставил оценку 4 или 5 за тестовый период,
- U — множество тестовых пользователей.

Тестовый список пользователей неизвестным для участников образом разделен на две половины — public и private. Оценка посылок в публичном лидерборде считается по первой половине, а в приватном — по второй. Финальной метрикой соревнования будет являться точность на приватной части тестового набора, которая будет показана только после завершения контеста.

Для удобства значение метрики при выставлении баллов умножается на 100.

Посылки

Изначально ограничение числа посылок равно 5. Далее каждый день лимит посылок будет увеличиваться на 5 штук. То есть суммарно за весь контест можно будет отправить 150 посылок. Также стоит ограничение на то, что посылки можно делать не чаще, чем раз в 10 минут.

Выбрать Файл не выбран

Отправить