

Привет!

Тестовое задание для Wargaming Forge состоит из двух задач. Первая — алгоритмическая, а вторая — о машинном обучении. Ты можешь решить только одну из них, но лучше сделать обе. Это повысит твой шанс попасть в Wargaming Forge.

Задача 1. Алгоритмы и неравные матчи!

Данные для задачи находятся в папке `task_1_data`. Нужно написать программу, которая составит матчи между командами игроков. Программа должна учитывать рейтинги игроков, чтобы составленное распределение было сбалансированным.

Сбалансированное распределение — набор матчей, в которых разница между рейтингом играющих команд будет минимальна.

Рейтинг команды — сумма рейтингов игроков, состоящих в команде.

Входные данные

Рейтинги игроков	<code>task_1_data/{test_name}/players.txt</code>
Команды	<code>task_1_data/{test_name}/teams.txt</code>

В папке `task_1_data` ты найдешь папки с юнит-тестами (`test_A`, `test_b` и т. д.). Каждая из папок содержит 2 файла:

- Текстовый файл `players.txt` содержит записи об игроках и их рейтингах. Один игрок в строке, без заголовка, в формате "`<user_id> <rating>`".
- Текстовый файл `teams.txt` содержит строки-записи о командах и список игроков каждой команды. Без заголовка, в формате "`<team_id> <user_id> <user_id> ... <user_id>`".

Выходные данные

Сбалансированные пары команды	<code>Username_task_1_team_pairs/{test_name}_pairs.txt</code>
Обоснование алгоритма	<code>Username_task_1_Report.pdf</code>
Исходный код программы	<code>Username_task_1_src/*</code>

Вместо каждого `{test_name}`, подставь названия тестов (название папок, в которых лежат тесты). Каждый выходной файл `{test_name}_pairs.txt` должен иметь формат "`<team_id> <team_id>`". Если команд нечетное количество, то без пары должна остаться только одна команда.

Разработанный алгоритм нужно обосновать! Мы ожидаем, что твой код не просто быстрый, а алгоритмически эффективный. Обязательно напиши обоснование своего решения в любом текстовом редакторе и сохрани в `Username_task_1_Report.pdf`

Задача 2. Машинное обучение «Найди комедии!»

Данные для задачи находятся в папке task_2_data. Предскажи, является ли фильм комедией, используя его атрибуты и изображение постера. Лучшим будет считаться решение, дающее максимально точный прогноз.

Метрика качества прогноза

Качество предсказания измеряется площадью под ROC-кривой на тестовой выборке.

Входные данные:

Обучающая выборка	task_2_data/train.csv
Тестовая выборка	task_2_data/test.csv
Изображения постеров фильмов	task_2_data/posters/*.jpg

Данные взяты на сайте <http://www.imdb.com>

Описание фильмов представлено в виде CSV-файлов (первая строка – заголовок):

Название колонки	Описание	Тип
Id	Уникальный идентификатор фильма	int
Target	Целевая переменная (является комедией или нет). <i>Присутствует только в обучающей выборке!!!</i>	boolean
Duration	Продолжительность фильма в минутах	int
Language	Оригинальный язык фильма	string
Country	Страна производства фильма	string
Rating	Возрастное ограничение	string
Action	Принадлежит ли данному жанру	boolean
Adventure	Принадлежит ли данному жанру	boolean
Animation	Принадлежит ли данному жанру	boolean
Biography	Принадлежит ли данному жанру	boolean
Crime	Принадлежит ли данному жанру	boolean
Documentary	Принадлежит ли данному жанру	boolean
Drama	Принадлежит ли данному жанру	boolean
Family	Принадлежит ли данному жанру	boolean
Fantasy	Принадлежит ли данному жанру	boolean
History	Принадлежит ли данному жанру	boolean
Horror	Принадлежит ли данному жанру	boolean
Music	Принадлежит ли данному жанру	boolean
Musical	Принадлежит ли данному жанру	boolean
Mystery	Принадлежит ли данному жанру	boolean
News	Принадлежит ли данному жанру	boolean
Romance	Принадлежит ли данному жанру	boolean
Sci-Fi	Принадлежит ли данному жанру	boolean
Sport	Принадлежит ли данному жанру	boolean
Thriller	Принадлежит ли данному жанру	boolean
War	Принадлежит ли данному жанру	boolean
Western	Принадлежит ли данному жанру	boolean
Poster	Путь к изображению постера в папке task_2_data/posters	string

Выходные файлы по задаче 2:

Предсказание на тестовой выборке	Username_task_2_prediction.csv
Описание решения	Username_task_2_Report.pdf или Username_task_2_Report.ipynb
Исходный код решения	Username_task_2_src/*

Файл Username_task_2_prediction.csv должен содержать:

Название колонок	Описание	Тип
Id	Уникальный идентификатор фильма из тестовой выборки	int
Probability	Предсказанная вероятность того, что фильм является комедией (вещественное число от 0 до 1).	float

Первая строка CSV-файла с предсказаниями должна содержать заголовок таблицы. Всего в файле должно быть ровно 910 строк (заголовок и далее 909 строк с предсказаниями). Пример правильно оформленного предсказания смотри в файле test_2_data/sample_submission.csv.

В этом задании нам интересно увидеть ход твоих рассуждений. Напиши, пожалуйста, в свободной форме описание твоего решения (Username_task_2_Report). О чем писать:

- Как ты пришел к финальному решению? Какие удачные и неудачные эксперименты ты проводил? Какие выводы из них сделал?
- Какие идеи ты не успел попробовать?

Срок и формат подачи решений

Для участия в конкурсном отборе необходимо заполнить заявку в Wargaming Forge не позднее 23:59:59 1 октября 2017 года. Чтобы отправить заявку, заполни анкету участника. Там прикрепи ZIP-архив с выполненным тестовым заданием (или ссылку на файлообменник, куда залит архив).

Назови ZIP-архив своими Ф.И.О. Он должен содержать:

Категория	Описание	Название файла
Задача 1	Сбалансированные пары команд	Username_task_1_team_pairs/{test_name}_pairs.txt (подробнее далее)
Задача 1	Обоснование алгоритма	Username_task_1_Report.pdf
Задача 1	Исходный код программы	Username_task_1_src/*
Задача 2	Предсказание на тестовой выборке	Username_task_2_prediction.csv
Задача 2	Описание решения	Username_task_2_Report.pdf или Username_task_2_Report.ipynb
Задача 2	Исходный код решения	Username_task_2_src/*

Вместо Username подставь, пожалуйста, свои Ф.И.О.

Если ты решаешь одну задачу, то прикрепляй только файлы, относящиеся к ней. Пожалуйста, не включай в архив предоставляемые нами наборы данных.

Ответы на возможные вопросы:

1. *Можно для построения предсказаний использовать не все атрибуты фильмов?* — Да, можно.
2. *На каком языке программирования можно решать задачи?* — На любом языке, который считаешь подходящим. Главное, не забудь приложить файлы с кодом к решению. Например, один из авторов при подготовке заданий использовал интерактивный Python через оболочку Jupyter Notebook (дистрибутив Anaconda Python).
3. *Можно совместить описание решения второй задачи и код самого решения в одном Jupyter Notebook-е (ipynb-файл)?* — Да, можно.
4. *Обязательно ли использовать постеры во втором задании?* — Нет, не обязательно

Остались вопросы? Пиши на wg_forge@wargaming.net. Мы постараемся ответить максимально быстро.

До встречи в Wargaming Forge!