Проект 1

Проект на обработку данных. Дедлайн - 03.11.2020 в 23:59

- Выбирайте
    - Свои данные, своя обработка - 15 баллов
    - Общий проект (я даю датасет и задание, см. ниже) - 10 баллов
- В проекте должны использоваться
    - Датасет (<20 МБ)
    - Вычисления 3-х каких-либо характеристик (см. Как пример задание к общему проекту) и вывод в терминал
    - Отрисовка графика либо запуск тетрадки юпитера либо запуск программы на питоне

- Примечания к инд. проекту
    - Я не буду запускать process.sh для вашего проекта, только для общего. Так что о пакетах для питона, пути запуска и т д не надо беспокоиться.
    - Инд проект по сложности должен быть не хуже чем общий. Если хотите пойти по пути наименьшего сопротивления, можете реализовать минимально приемлемые (несложные) решения и поставить тег #п1хочухаляву в README.md. Тогда я не приду к вам в требованием переделать и поставлю 5 баллов.

Что сдавать

В репо linux-git1

- README.md с описанием проекта, данных, и скрипта process.sh - что он принимает на вход и что выводит
- Bash скрипт process.sh, который должен запускаться как ./process.sh и принимать файл с датасетом в качестве аргумента.

Дополнительно для индивидуального проекта:

- Файл p1_ind как индикатор индивидуального проекта.
- Файл tguserid, с вашим id телеграмма который получается командой /id в @ozonm_big_data_bot - через бот получите оценку.

Проверка

Запустите чекер: **checker.sh p1**

Общий проект проверяется автоматически. Вы должны добиться сообщения **PASSED 1** в самом конце. Если такого сообщения нет, то смотрите в вывод чекера на наличие ошибок.

Индивидуальный проект - я в момент дедлайна скачиваю ваши репо, проверяю. Отсылаю отметку через бот.

Общий проект

- Датасет /home/datamove/hotels.csv
- Вычислите
    - Средний рейтинг (overall_ratingsource)
        - Формат **RATING_AVG число**
    - Число отелей в каждой стране (название страны должно быть маленькими буквами. Например, его можно вычленять из первого поля (doc_id), либо еще как, погуглите :)
        - Формат каждой строчки: **HOTELNUMBER страна число**
    - Средний балл cleanliness по стране (колонка cleanliness) для отелей сети Holiday Inn vs. отелей Hilton
        - Формат: **CLEANLINESS страна холидейинн хилтон**

- Используя **gnuplot** рассчитайте коэффициенты линейной регрессии для зависимости чистоты (cleanliness) от общей оценки (overall_ratingsource). Нарисуйте этот линейный график и точки данных на нем, сохраните в png.
- Примечания
    - Если используете временные файлы - то пишите их в /tmp
    - Не добавляйте результаты работы gnuplot (fit.log, *png) в репо