

HW #06: Spark SQL

1. Описание задания	2
2. Критерии оценивания	2
3. Описание данных	2
4: (Task ID: spark.sssp) Single Source Shortest Path algorithm	3
5. Правила оформления задания	3

автор задания: BigData Team, коллективная работа.

1. Описание задания

В данном ДЗ нужно решить **1 задачу**. Решение надо выполнить с помощью Spark SQL (Dataframe).

2. Критерии оценивания

Балл за задачу складывается из:

- **60%** - правильное решение задачи
- **20%** - поддерживаемость и читаемость кода
 - в общем случае см. Clean Code и [Google Python Style Guide](#)
 - оценка качества будет проводиться автоматическим вызовом pylint:
 - `pylint *.py -d invalid-name,missing-docstring`
 - качество кода должно оцениваться выше 8.0 / 10.0
 - проверяем код **Python версии 3** с помощью `pylint==2.5.3`
- **20%** - эффективность решения (такие как потребляемые CPU-ресурсы, скорость выполнения (в предположении свободного кластера)).

Discounts (скидки и другие акции):

- **100%** за плагиат в решениях (всем участникам процесса)
- **100%** за посылку решения после hard deadline
- **30%** за посылку решения в после soft deadline и до hard deadline
- **5%** за каждую новую посылку (одна дополнительная посылка бесплатно)

Формула подсчета финальной оценки¹:

$\max(0, 0.95^{\max(0, \# \text{доп.посылок} - 1)} * (1 - \text{штраф. за. дедлайн. и. списывание})) * \text{последняя. оценка. из. grader}$

3. Описание данных

3.1 Социальный граф Twitter

twitter:

- Путь на кластере:
 - полный датасет: `/data/twitter/twitter.txt`
 - Семпл (для тестирования): `/data/twitter/twitter_sample_small.txt`
 - Семпл-2 (для тестирования): `/data/twitter/twitter_sample.txt`
- Формат: текст

¹ результат округляется до целого



- В каждой строке находятся следующие поля, разделенные знаком табуляции:
 - INT - ID пользователя
 - INT - ID follower'a
- Граф считаем направленным: follower → user.

Пример:

```
12    18
12    41
12    57
12    62
12   235
12   278
12   291
12   338
12   456
12   614
...
```

4: (Task ID: spark.sssp) Single Source Shortest Path algorithm

В этом домашнем задании вам предстоит реализовать алгоритм поиска кратчайшего пути в графе. Вам необходимо реализовать алгоритм поиска кратчайшего пути от одного пользователя Twitter к другому, используя поиск в ширину ([BFS](#)). Для успешной сдачи задания необходимо найти кратчайший путь от пользователя **12** к пользователю **34**.

Для тестирования решения предлагается пользоваться неполными датасетами. Длина кратчайшего пути между заданными вершинами в каждом датасете будет разная!

Условия:

- ваше решение должно вывести в STDOUT ровно одно число - длину кратчайшего пути между этими пользователями
- если для выполнения этого задания вам потребуется реализовать UDF, то ее необходимо реализовать именно как `pandas_udf` для ускорения работы алгоритма. Также посмотрите, нет ли необходимой вам функции в модуле `pyspark.sql.functions` (возможно, она там действительно есть)

Пример вывода:

5. Правила оформления задания

Оформление задания:

- Код задания (Short name): **HW6:Spark-SQL(SSSP)**.
- Выполненное ДЗ запакуйте в архив `X5BD2021Q1_<Surname>_<Name>_HW#.zip`, например, для Алексея Драля -- `X5BD2021Q1_Dral_Alexey_HW6.zip`. Если ваше решение лежит в папке `my_solution_folder`, то для создания архива `hw.zip` на Linux и Mac OS выполните команду²:
 - `zip -r hw.zip my_solution_folder/*`
- На Windows 7/8/10: необходимо выделить все содержимое директории `my_solution_folder/` нажать правую кнопку мыши на одном из выделенных объектов, выбрать в открывшемся меню "Отправить >", затем "Сжатая ZIP-папка". Теперь можно переименовать архив.
- Решения заданий должно содержаться в одной папке.
- PySpark-скрипт для запуска решения следует назвать `task_<Surname>_<Name>_sssp.py`:
 - решение будет запускаться с помощью команды:
 - `PYSPARK_DRIVER_PYTHON=python3.6 PYSPARK_PYTHON=python3.6 spark-submit "task_*_sssp.py"`
 - скрипт выводит на экран (STDOUT) указанное в задании число строк в нужном формате
- Вывод **STDOUT** задач нужно сохранить в соответствующих файлах в архиве **посылке домашнего задания (например, `task_*_sssp.out`)**.³
- Перед проверкой убедитесь, что дерево вашего архива выглядит так:
 - `| X5BD2021Q1_<Surname>_<Name>_HW6.zip`
 - `| ---- task_<Surname>_<Name>_sssp.py`
 - `| ---- task_<Surname>_<Name>_sssp.out`
 - При несовпадении дерева вашего архива с представленным деревом, ваше решение будет невозможно автоматически проверить, а значит, и оценить его.
- Для того, чтобы сдать задание необходимо:
 - Зарегистрироваться и залогиниться в сервисе [Everest](#)
 - Перейти на страницу приложения: [BDT-grader-X5-BD](#)
 - Выбрать вкладку Submit Job (если отображается иная).

² Флаг `-r` значит, что будет совершен рекурсивный обход по структуре директории

³ Для подготовки архива с решением и выводом результатов запуска можно воспользоваться командой `"tee"`



- Выбрать в качестве "Task" значение: **HW6:Spark-SQL(SSSP)**⁴
 - Загрузить в качестве "Task solution" файл с решением
 - В качестве Sender ID указать тот, который был выслан по почте
- Если Вы видите надпись "You are not allowed to run this application" во вкладке Submit Job в Everest, то на данный момент сдача закрыта (нет доступных для сдачи домашних заданий, по техническим причинам или другое). Попробуйте, пожалуйста, еще раз через некоторое время. Если Вы еще ни разу не сдавали, у коллег сдача работает, но Вы видите такое сообщение, сообщите нам об этом.
- Ситуации:
 - * система оценивания показывает оценку (Grade) < 0, а отчет (Grading report) не помогает решить проблему (пример помощи: в случае неправильно указанного Sender ID система вернет -2 и информацию о том, что его нужно поправить);
 - * показывает 0 и в отчете (Grading report) не указано, какие тесты не пройдены. Если Вы столкнулись с какой-то из них присылайте ссылку на выполненное задание (Job) на почту с темой письма "Short name. ФИО.". Например: **"HW6:Spark-SQL(SSSP). Иванов Иван Иванович."**
Пример ссылки: <https://everest.distcomp.org/jobs/67893456230000abc0123def>
- **Внимание:** Если до дедлайна остается меньше суток, и Вы знаете (сами проверили или коллеги сообщили), что сдача решений сломана, обязательно сдайте свое решение и напишите письмо, как написано выше, чтобы мы видели, какое решение Вы имели до дедлайна и смогли его оценить.
- Перед отправкой задания, оставьте, пожалуйста, отзыв о домашнем задании по ссылке: http://rebrand.ly/x5bd2021q1_feedback_hw06. Это позволит нам скорректировать учебную нагрузку по следующим заданиям (в зависимости от того, сколько часов уходит на решение ДЗ), а также ответить на интересные вопросы.

Любые вопросы / комментарии / предложения можно писать в телеграм-канал курса или на почту bigdata_x52021q1@bigdatateam.org .

Всем удачи!

⁴ Сервисный ID: spark.sssp