BIGDATA TEAM

HW #07: Spark SQL

Deadline: 05.08.2019, 08:00

---

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 (Clean Code, см. например Google Python Style Guide)
- **20%** - эффективность решения

Штрафы:

- **10%** за несоответствие правилам оформления задания
- **30%** за просрочку дедлайн

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
- 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: Single Source Shortest Path (SSSP) algorithm[1].

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**.

[1] Внутренний ID задачи (для проверяющих) - 803

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:

```
1234
```

5. Сроки сдачи и правила оформления задания.

Deadline: 05.08.2019, 08:00

Оформление задания:

- Код задания (Short name): **HW7:Spark-SQL**.
- Решения задач должны содержаться в одной папке.
- Выполненное ДЗ запакуйте в архив **MF**2019Q2_<фамилия>_HW#.zip , например -- **MF**2019Q2_Ivanov_**HW7**.zip. Например, ваше решение лежит в папке my_solution_folder, тогда чтобы на Linux и Mac OS создать архив под названием hw.zip и пожать его с помощью zip выполните команду[2]:
  - zip -r hw.zip my_solution_folder/

  На Windows 7/8/10: необходимо нажать правую кнопку мыши на директорию my_solution_folder/, выбрать в открывшемся меню "Отправить >", затем "Сжатая ZIP-папка". Теперь можно переименовать архив.
- Присылайте выполненное задание на почту bigdata_mf2019q2@bigdatateam.org с темой письма “Short name. ФИО.”. Например: “**HW7:Spark-SQL**. Иванов Иван Иванович.”
- Перед отправкой письма, оставьте, пожалуйста, отзыв о домашнем задании по ссылке: http://rebrand.ly/mf2019q2_feedback_hw07. Это позволит нам скорректировать учебную нагрузку по следующим заданиям (в зависимости от того, сколько часов уходит на решение ДЗ), а также ответить на интересующие вопросы.

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

Любые вопросы / комментарии / предложения можно писать:

- в телеграм-канале: http://rebrand.ly/mf2019q2_telegram_join
- На почту: bigdata_mf2019q2@bigdatateam.org

Всем удачи!

6. Дорешка.

Решения после получения фидбека на решение ДЗ можно улучшить. Разрешается одна досылка в течение 1й недели после окончания дедлайна за ДЗ. Соответственно, фидбек за дорешенные ДЗ вы получите в течение 24 часов после окончания deadline следующего ДЗ.

Дорешивать неработающие задания - нельзя. Это позволит исправить дисбаланс между

присланными **работающими** заданиями **после** deadline

VS

присланными **НЕработающими** заданиями **до** deadline