BIGDATA TEAM

HW #08: Spark Optimization

Deadline: 18.08.2019, 08:00

1. Описание задания.

Решение надо выполнить с помощью Apache Spark и оптимизировать исполнение с помощью методов, изученных на занятии по оптимизации Spark.

2. Критерии оценивания.

Балл за задачу складывается из:

- **40%** - правильное решение задачи
- **40%** - эффективность решения
- 20% - поддерживаемость и читаемость кода (Clean Code, см. например Google Python Style Guide)

Штрафы:

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

3. Описание данных.

3.1 Дамп Википедии

`en_articles_part`:

- Путь на кластере: полный датасет - **`/data/wiki/en_articles_part`**
- Формат: текст
- В каждой строке находятся следующие поля, разделенные знаком табуляции:
    1. INT - id статьи,
    2. STRING - текст статьи,

Пример:

```
12    Anarchism         Anarchism is often defined as a political
philosophy which holds the state to be undesirable, unnecessary, or
harmful.
```

4. Задача[1]: TF-IDF.

Посчитайте TF-IDF для слов статей Википедии. Про TF-IDF подробнее можно почитать здесь: https://en.wikipedia.org/wiki/Tf-idf.

Чтобы рассчитать TF-IDF введем несколько определений:

1. `TF(w, d)` - вероятность увидеть слово “w” в документе (статье) “d”.
   `TF(w, d) = word_count_"w"_in_"d" / word_count_in_"d"`
2. `DF(w)` - число документов со словом “w” в заданном датасете
   `DF(w) = num_of_docs_with_word_"w"`
3. `IDF(w)` - величина уникальности слова “w” в заданном датасете (чем больше величина, тем более уникально слов)
   `IDF(w) = log(total_number_of_docs / (1 + DF(w))`
   `total_number_of_docs` - общее количество документов в датасете
   `log` - используется логарифм по основанию 10

Когда мы перемножаем величины TF и IDF, то получаем значимость слова для указанной статьи. Редкие слова (маленький DF, большой IDF), которые встречаются часто в одном документе, будут характеризовать указанную статью (например, слово “нуклеотид” в

[1] Для проверяющих - за основу взята задача 705

статье про РНК). Благодаря IDF можно также находить стоп-слова, но этот вопрос уже для другой задачи.

Выведите интересующие пары (см. условия ниже) с их значением TF-IDF по убыванию значения TF-IDF. Формат вывода:

tf_idf <tab> document <tab> word

Цель задания: минимизировать число используемых stage в период вычислений.

Условия:

- очистить тексты от знаков пунктуации (см. re.sub / re.split);
- привести все слова к нижнему регистру;
- для **каждой пары** (слово, документ) подсчитайте значение TF-IDF;[2]
- для статьи 12 вывести **TOP-50** самых важных тематических слов и их значения TF-IDF (округляем до 3го знака после запятой, рекомендуется использовать "{:10.3f}".format(tf_idf_value)). Слова в рамках указанной статьи отсортировать по значению TF-IDF по убыванию;

Пример вывода:

```
...
  0.003        12       revolutionary
...
```

5. Сроки сдачи и правила оформления задания.

Deadline: 18.08.2019, 08:00

Оформление задания:

- Код задания (Short name): **HW8:Spark-advanced**.
- Решения задач должны содержаться в одной папке.
- Выполненное ДЗ сохраните в файл **MF**2019Q2_<фамилия>_HW#.ipynb , например -- **MF**2019Q2_Ivanov_**HW8**.ipynb.
- В обязательном порядке оставьте вывод исполнения ipynb, чтобы можно было визуально проверить результат безе перезапуска решения.
- К письму приложите скриншот и ссылку на UI оптимизированного DAG.

[2] Фильтрацию по интересующим статьям проводить только в самом конце.

- Присылайте выполненное задание на почту bigdata_mf2019q2@bigdatateam.org с темой письма “Short name. ФИО”. Например: “**HW8:Spark-advanced**. Иванов Иван Иванович”.
- Перед отправкой письма, оставьте, пожалуйста, отзыв о домашнем задании по ссылке: http://rebrand.ly/mf2019q2_feedback_hw08. 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.

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** deadli