BIGDATA TEAM

HW #06: Spark RDD

Deadline: 29.07.2019, 08:00

---

1. Описание задания.

В данном ДЗ нужно решить **2 задачи**. Задачи общие для всех. Решение надо выполнить с помощью Apache Spark, можно использовать только RDD API.

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;

Веса задач:

1. 40%
2. 60%

3. Описание данных

3.1 Дамп Википедии

`en_articles_part`:

- Путь на кластере: полный[1] 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.
```

3.2 Стоп-слова

`stop_words_en`:

- Путь на кластере: **`/data/wiki/stop_words_en-xpo6.txt`**
- Формат: одно стоп-слово на строчку

Пример:

```
...
wherein
whereupon
wherever
...
```

4.1 Задача #1[2]: народные биграммы.

Найдите все пары двух последовательных слов (биграмм), где первое слово:

```
narodnaya
```

[1] Да, здесь нет ошибки, работаем на части данных, чтобы побыстрее познакомиться со Spark RDD
[2] Внутренний ID задачи (для проверяющих) - 517

Для каждой пары подсчитайте количество вхождений в тексте статей Википедии. Выведите все пары с их частотой вхождений в лексикографическом порядке. Формат вывода:

```
word_pair <tab> count
```

Условия:

- очистить тексты от знаков пунктуации (см. `re.sub`)
- привести все слова к нижнему регистру;
- слова в паре объединить символом нижнего подчеркивания “_”;
- отсортировать слова в выводе по алфавиту;

Пример вывода:

```
...
crazy_zoo 42
red_apple 100500
...
```

4.2 Задача #2[3]: коллокации.

Коллокация - это комбинации слов, которые часто встречаются вместе. Например, «high school» или «Roman Empire». Чтобы определить является ли пара слов коллокацией, можно воспользоваться метрикой NPMI - нормализованная точечная взаимная информация.

Чтобы рассчитать NPMI введем несколько определений:

1. `P(a)` - вероятность увидеть слово “a” в датасете.
   `P(a) = num_of_occurrences_of_word_"a" / total_number_of_words`
   `total_number_of_words` - общее количество слов в тексте
2. `P(ab)` - вероятность увидеть пару слов “a” и “b”, идущих подряд.
   `P(ab) = num_of_occurrences_of_pair_"ab" / total_number_of_word_pairs`
   `total_number_of_word_pairs` - общее количество пар
3. `PMI(a,b) = ln( P(ab) / [P(a) x P(b)] )`
4. `NPMI(a,b) = PMI(a,b) / -ln(P(ab))` - величина PMI нормализованная в диапазон [-1, 1];

Примеры и комментарии:

[3] Внутренний ID задачи (для проверяющих) - 518

- значение NPMI равное “-1” будет означать, что пара слов никогда не встречается в датасете. Например такие пары как **“green idea”** или **“sleeps furiously”** никогда не встречаются вместе, поэтому `P(ab) = 0`, следовательно `PMI(a,b) = -inf`, NPMI = -1;
- значение NPMI равное “0” будет означать, что слова в паре встречается абсолютно независимо друг от друга. Рассмотрим пример **"the doors"**: “the” может встретится рядом с любым словом. Таким образом, `P(ab) = P(a) x P(b)` и `PMI(a,b) = ln(1) = 0`, NPMI = `0`.
- значение NPMI равное “1” будет означать, что это идеальная коллокация. Рассмотрим пример **"the doors"**: “the” может встретится рядом с любым словом. Таким образом, `P(ab) = P(a) x P(b)` и `PMI(a,b) = ln 1 = 0`, NPMI = `0`. Предположим, что **"Roman Empire"** - это уникальная комбинация, и за каждым появлением “Roman” следует “Empire”, и, наоборот, каждому появлению “Empire” предшествует “Roman”. В этом случае `P(ab) = P (a) = P(b)`, поэтому `PMI(a,b) = -ln(P(a)) = -ln(P(b))`, следовательно NPMI = 1.

Условия:

- найти самые популярные коллокации в Википедии;
- очистить тексты от знаков пунктуации (см. `re.sub`)
- привести все слова к нижнему регистру;
- удалить стоп-слова;
- слова в паре объединить символом нижнего подчеркивания “_”;
- отфильтровать биграммы, которые встретились **не реже** 500 раз (т.е. проводим все необходимые join'ы и считаем NPMI только для них, НО оценку вероятности встретить биграмму считаем на полном датасете);
- отсортировать слова в выводе по значению NPMI;
- вывести **TOP-39** самых популярных коллокаций и их значения NPMI (округляем до 3го знака после запятой, см. round).

Пример вывода.

```
...
south_africa      0.619
roman_empire      0.603
...
```

5. Сроки сдачи и правила оформления задания.

Deadline: 29.07.2019, 08:00

Оформление задания:

- Код задания (Short name): **HW6:Spark-RDD**.
- Решения задач должны содержаться в одной папке.
- Выполненное ДЗ запакуйте в архив **MF**2019Q2_<фамилия>_HW#.zip , например --MF2019Q2_Ivanov_HW5.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[4]:
    - 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: “**HW6:Spark-RDD**. Иванов Иван Иванович.”
- Перед отправкой письма, оставьте, пожалуйста, отзыв о домашнем задании по сссылке: http://rebrand.ly/mf2019q2_feedback_hw05. 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!

7. 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

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