Введение

Предлагается решить 2 задачи. 1-я задача для всех одинакова. Номер 2-й задачи определяется с помощью этого скрипта. На вход скрипту подается ваш **username в gitlab.atp-fivt.org**. Например:

```
~$ ./get_mr_variant.py velkerr
120
```

- ❏ Репозиторий для сдачи: http://gitlab.atp-fivt.org/hobod2021/USERNAME-hobmapreduce
- ❏ Ветки: hobmapreducetask1, hobmapreducetask2

Сроки

- ➢ Мягкий deadline: 09.03, 23:59.
- ➢ Жесткий deadline: 14.03, 23:59.

Как сдать задание

1. В каждой ветке создать директорию, имя директории = имени ветки. Например, нужно создать директории с названиями hobmapreducetask1 и hobmapreducetask2 в ветках hobmapreducetask1 и hobmapreducetask2 соответственно.
2. В созданных директориях положите файл run.sh. Это точка входа в вашу программу и именно её будет запускать система проверки. В run.sh может быть как всё решение задачи так и вызов других файлов.
3. Сделать merge request из ветки в master. Высылать ссылку на MR не нужно, достаточно убедиться что ваша ссылка появилась во 2й вкладке таблицы с оценками (обновление таблицы происходит раз в час).

Задание 1

Исходные данные

Википедия:

- Путь на кластере: `/data/wiki/en_articles`, семпл: `/data/wiki/en_articles_part`
- Формат: текст, в каждой строке: идентификатор статьи <tab> текст статьи

Задача 1 (111)

Посчитайте число вхождений имён собственных длиной от 6 до 9 символов. Имя собственное - это слово, начинающееся с заглавной буквы и последующих маленьких букв (и только маленьких) и которое ни разу не встретилось в тексте с маленькой буквы. При разборе статей очищайте слова от от знаков пунктуации. Результат приведите к нижнему регистру и отсортируйте по убыванию числа вхождений, в случае равенства - лексикографически.

- Входные данные: википедия.

- Формат вывода в HDFS: имя количество
- Вывод на печать: топ10 имен.

Пример вывода:

```
english 8358
states  7264
british 6829
...
```

Задание 2

Исходные данные

Путь к данным в HDFS: `/data/minecraft-server-logs`

Данные представляют собой логи сервера игры Minecraft, собранные с 29.11.17, 16:55 по 31.12.17, 23:53. Строчка логов имеет вид:

```
[YYYY-MM-dd.HH:mm:ss] [Thread name/SEVERITY]: Message
```

Например:

```
[2017-11-29.17:22:26] [Server thread/INFO]: [0;37;22m[[0;36;1mAdministration[0;36;22
```

Message содержит информацию о событии, произошедшем на игровом сервере. В частности, сообщения могут быть таких типов:

➢ Обычные события:

```
[2017-11-29.17:22:26] [Server thread/INFO]: Luck20 lost connection: Server closed
```

➢ Предупреждения:

```
[2017-12-0216:33:12] [Server thread/WARN]: Plugin `Administration_Panel v1.1.0' uses the
space-character (0x20) in its name `Administration Panel' - this is discouraged
```

➢ Событие, связанное с плагином или модом. Например:

```
[2017-11-29.17:22:26] [Server thread/INFO]: [AntiAd] Disabling AntiAd v2.3.4
```

Видим, что сообщение инициировано плагином [AntiAd].

➢ Ошибка. Отличается тем, что может содержать несколько строк, например:

```
[2017-11-29.16:55:43]       [Main       thread/ERROR]:       Could       not       load
'plugins/EssentialsSpawn-2.x-SNAPSHOT.jar' in folder 'plugins'
org.bukkit.plugin.InvalidDescriptionException: Invalid plugin.yml
      at
org.bukkit.plugin.java.JavaPluginLoader.getPluginDescription(JavaPluginLoader.java:162)
~[spigot-1.8.8.jar:git-Spigot-db6de12-18fbb24]
      at  org.bukkit.plugin.SimplePluginManager.loadPlugins(SimplePluginManager.java:133)
[spigot-1.8.8.jar:git-Spigot-db6de12-18fbb24]
      at     org.bukkit.craftbukkit.v1_8_R3.CraftServer.loadPlugins(CraftServer.java:292)
[spigot-1.8.8.jar:git-Spigot-db6de12-18fbb24]
      at      net.minecraft.server.v1_8_R3.DedicatedServer.init(DedicatedServer.java:198)
[spigot-1.8.8.jar:git-Spigot-db6de12-18fbb24]
      at       net.minecraft.server.v1_8_R3.MinecraftServer.run(MinecraftServer.java:525)
[spigot-1.8.8.jar:git-Spigot-db6de12-18fbb24]
      at java.lang.Thread.run(Thread.java:745) [?:1.7.0_85]
```

Дополнительные комментарии

1. Во всех задачах, где не написано обратное, вывести TOP-10 записей.

2. Не стоит удивляться, если результаты в задачах будут получаться очень несбалансированными. Очень часто, так обстоят дела и в реальной жизни.

Задачи

Задача 1 [120]. “Мониторинг”. Посчитать кол-во warning’ов и ошибок по дням. Результат отсортировать[1] по кол-ву ошибок. При равном кол-ве ошибок отсортировать по кол-ву предупреждений.

Формат вывода:

```
YYYY-MM-dd <tab> errors <tab> warnings
```

Пример вывода:

```
2017-11-29    20   11
2017-12-02    16   61
```

Задача 2 [121]. “Исследование ошибок”. Сервер настроен достаточно сыро, поэтому в логах часто выкидываются ошибки. Админу сервера нужно найти самый забагованный плагин чтоб понять, стоит ли его вообще оставлять на сервере.

Проанализировать стектрейсы ошибок и вывести кол-во ошибок для каждого Java-класса, который их инициирует. Например, такая ошибка

```
[2017-11-29.16:55:43] [Server thread/ERROR]: Could not load 'plugins/figadmin.jar' in
folder 'plugins'
org.bukkit.plugin.InvalidDescriptionException: Invalid plugin.yml
      at
org.bukkit.plugin.java.JavaPluginLoader.getPluginDescription(JavaPluginLoader.java:162)
~[spigot-1.8.8.jar:git-Spigot-db6de12-18fbb24]
      at
org.bukkit.plugin.SimplePluginManager.loadPlugins(SimplePluginManager.java:133)
[spigot-1.8.8.jar:git-Spigot-db6de12-18fbb24]
      at org.bukkit.craftbukkit.v1_8_R3.CraftServer.loadPlugins(CraftServer.java:292)
[spigot-1.8.8.jar:git-Spigot-db6de12-18fbb24]
      at net.minecraft.server.v1_8_R3.DedicatedServer.init(DedicatedServer.java:198)
[spigot-1.8.8.jar:git-Spigot-db6de12-18fbb24]
      at net.minecraft.server.v1_8_R3.MinecraftServer.run(MinecraftServer.java:525)
[spigot-1.8.8.jar:git-Spigot-db6de12-18fbb24]
      at java.lang.Thread.run(Thread.java:745) [?:1.7.0_85]
Caused by: java.util.zip.ZipException: error in opening zip file
      at java.util.zip.ZipFile.open(Native Method) ~[?:1.7.0_85]
      at java.util.zip.ZipFile.<init>(ZipFile.java:215) ~[?:1.7.0_85]
      at java.util.zip.ZipFile.<init>(ZipFile.java:145) ~[?:1.7.0_85]
      at java.util.jar.JarFile.<init>(JarFile.java:154) ~[?:1.7.0_85]
      at java.util.jar.JarFile.<init>(JarFile.java:118) ~[?:1.7.0_85]
      at
org.bukkit.plugin.java.JavaPluginLoader.getPluginDescription(JavaPluginLoader.java:150)
```

инициирована классом org.bukkit.plugin.java.JavaPluginLoader.getPluginDescription.

- ➢ Есть ошибки, в которых нет Stacktrace’а. Их в данной задаче просто игнорируем.
- ➢ Номер строки и название файла с кодом (`JavaPluginLoader.java:150`) следует отсеять при обработке.
- ➢ Следует понимать, что stacktrace читается снизу вверх, поэтому самым информативным для отладки является самое нижнее значение.

[1] Во всех задачах сортировка только по убыванию.

Результат отсортировать по количеству.

Формат вывода:

```
SomeException <tab> N
```

Пример вывода:

```
org.bukkit.plugin.java.JavaPluginLoader.getPluginDescription    20
java.lang.Thread.run    11
```

Задача 3 [122]. “Смертность”. Посчитать среднее кол-во смертей (с точностью до сотых) за одну сессию по каждому пользователю. Также для каждого пользователя вывести общее кол-во его сессий.

- ➢ Сессия начинается сообщением: **`UUID of player some_user is XXXX`**
- ➢ А заканчивается **`some_user lost connection: Disconnected`**
- ➢ Смерть пользователя: **`some_user died`**

Отсортировать по среднему кол-ву смертей. При равном среднем количестве смертей, отсортировать лексикографически по никам пользователей.

Формат вывода:

```
some_user <tab> смертность <tab> число_сессий
```

Пример вывода:

```
avivzusim   0.24    15
lucky20    0    10
```

Задача 4 [123]. “Самый активный пользователь”. Посчитать среднее количество команд, введённых пользователем на сервере за 1 сессию. Также для каждого пользователя вывести общее кол-во его сессий.

- ➢ Сессия начинается сообщением: **`UUID of player some_user is XXXX`**
- ➢ А заканчивается
  - **`some_user lost connection: Disconnected`**
  - **`com.mojang.authlib.GameProfile … lost connection:...`**
- ➢ Команда, введённая игроком, обозначается так: **`some_user issued server command: /menu`**

Отсортировать по кол-ву команд. При равном среднем количестве команд, отсортировать лексикографически по никам пользователей.

Дополнительные комментарии[2]

1. Есть технические пользователи, которые генерируют команды, но при этом не запускают сессий. Такие команды не учитываются.
2. Если сессия не закрылась, то считаем что она закрывается в момент окончания лога.
3. Если сессия не открывалась (открылась до начала логов), то её не учитываем.

Формат вывода:

```
some_user <tab> ср. кол-во команд <tab> число_сессий
```

Пример вывода:

[2] Комментарии от А. Поповкина.

```
lucky20    25.0    10
avivzusim   5.0    15
```

Задача 5 [124]. “Популярные команды по дням”. Посчитать суммарное количество введённых игроками команд по каждому дню.

- ➢ В логах команда, введённая на сервере, обозначается так: **`some_user issued server command: /menu`**

Команда должна состоять из 1 слова. Аргументы команды не учитываем. Т.е. если при извлечении данных появились команды типа **/tp user 25**, оставляем только **/tp**.

Результат отсортировать по дням (по возрастанию даты). Строки с одинаковыми днями отсортировать по количеству.

Формат вывода:

```
день <tab> команда <tab> кол-во
```

Пример вывода:

```
2017-11-29    /menu    11
2017-12-02    /hack    9
```

Задача 6 [125]. “Многопоточность”. Сервер Minecraft - параллельная программа и работает в несколько потоков. Нужно определить сбалансированность нагрузки на потоки. Для этого подсчитайте среднее количество сообщений, генерируемых различными потоками в день.

Если в какой-то день определённый поток не сгенерировал ни одного сообщения, этот день из рассмотрения не выбрасываем. Считаем, что в этот день было 0 сообщений и среднее всё равно берём по всем дням.

Результат отсортировать по количеству сообщений.

Формат вывода:

```
Название потока <tab> среднее кол-во сообщений в день
```

Среднее значение нужно выводить в формате с плавающей точкой.

Пример вывода:

```
Server    566.6
Main   25.7
```

Технические особенности

1. Презентация “Как сдать задание”.
2. Система тестирования читает ответ из STDOUT поэтому при выводе из STDOUT требуется вычистить всё, что не является ответом задачи. STDERR наоборот нужно оставить нетронутым.