

# ДЗ 4

Кондраев Дмитрий

## 1 Какая связь между DataFrame и Dataset? (1 балл)

```
DataFrame = DataSet[Row]
```

## 2 Соберите WordCount приложение, запустите на датасете ppkm\_sentiment (1 балл)

Исходники см. в репозитории <https://github.com/mariohuq/2022-polis-ml>. Команды выполняются из корневой директории репозитория:

```
$ git clone https://github.com/mariohuq/2022-polis-ml.git kondraev-ml
$ cd kondraev-ml
```

1. Старт контейнера, созданного в прошлом ДЗ:

```
$ systemctl start docker.service
$ docker start -i zeppelin
```

2. Сборка и копирование WordCount

```
$ git checkout 8fffe2ce
$ cd 04-spark-sql/WordCount
$ sbt package
$ docker cp ./**/wordcount_2.11-0.1.0-SNAPSHOT.jar zeppelin:/home/hduser/
```

3. Запуск Spark

```
hduser@localhost:~$ hdfs dfs -rm -r /user/hduser/scala-ppkm-out
hduser@localhost:~$ spark-submit \
  --name WordCount \
  --class ok.ml.WordCount \
  --master yarn \
  --deploy-mode cluster \
  wordcount_2.11-0.1.0-SNAPSHOT.jar \
  /user/hduser/ppkm/ppkm_dataset.csv \
  /user/hduser/scala-ppkm-out
```

<output trimmed for brevity>

```
22/10/23 16:08:38 INFO yarn.Client: Application report for application_1666541178843_0001 (sta
```

```
22/10/23 16:08:38 INFO yarn.Client:
```

```
  client token: N/A
```

```
  diagnostics: N/A
```

```
  ApplicationMaster host: localhost
```

```
  ApplicationMaster RPC port: 37959
```

```
  queue: default
```

```
  start time: 1666541220035
```

```
  final status: SUCCEEDED
```

```
  tracking URL: http://localhost:8088/proxy/application_1666541178843_0001/
```

```
  user: hduser
```

```
22/10/23 16:08:38 INFO util.ShutdownHookManager: Shutdown hook called
```

```
22/10/23 16:08:38 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-
c4ff24bad-4f15-4a23-b2d2-51a58d6d05b0
```

```
22/10/23 16:08:38 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-9ba30636-c4ba-4136-a256-649d01a0876f
```

При выводе образовалось 200 файлов:

```
hduser@localhost:~$ hdfs dfs -ls /user/hduser/scala-ppkm-out/part-* | wc -l
200
```

```
hduser@localhost:~$ hdfs dfs -cat /user/hduser/scala-ppkm-out/part-*
```

<output trimmed for brevity>

```
Perpanjangan,8
"Mitra,",2
#beritajabar,1
Belum,1
mentok,1
kebudayaan,1
DP,3
empat,1
Percuma,1
maksud,1
```

### 3 Измените WordCount так, чтобы он удалял знаки препинания и приводил все слова к единому регистру (2 балла)

#### 1. Сборка и копирование WordCount

```
$ git checkout 754f8a89
$ cd 04-spark-sql/WordCount
$ sbt package
$ docker cp ./**/wordcount_2.11-0.2.0-SNAPSHOT.jar zeppelin:/home/hduser/
```

#### 2. Запуск Spark

```
hduser@localhost:~$ hdfs dfs -rm -r /user/hduser/scala-ppkm-out
hduser@localhost:~$ spark-submit \
```

```
--name WordCount \
--class ok.ml.WordCount \
--master yarn \
--deploy-mode cluster \
wordcount_2.11-0.2.0-SNAPSHOT.jar \
/user/hduser/ppkm/ppkm_dataset.csv \
/user/hduser/scala-ppkm-out
```

```
hduser@localhost:~$ hdfs dfs -cat /user/hduser/scala-ppkm-out/part-*
```

<output trimmed for brevity>

```
dilaksanakan,2
sebagai,2
februari,2
mentok,1
kebudayaan,1
empat,1
wartajateng,1
langgarpsbb,4
maksud,1
sosialisasi,2
```

#### 4 Измените выход WordCount так, чтобы сортировка была по количеству повторений, а список слов был во втором столбце, а не в первом (1 балл)

##### 1. Сборка и копирование WordCount

```
$ git checkout 46c8dae3
$ cd 04-spark-sql/WordCount
$ sbt package
$ docker cp ./**/wordcount_2.11-0.3.0-SNAPSHOT.jar zeppelin:/home/hduser/
```

##### 2. Запуск Spark

```
hduser@localhost:~$ hdfs dfs -rm -r /user/hduser/scala-ppkm-out
hduser@localhost:~$ spark-submit \
  --name WordCount \
  --class ok.ml.WordCount \
  --master yarn \
  --deploy-mode cluster \
  wordcount_2.11-0.3.0-SNAPSHOT.jar \
  /user/hduser/ppkm/ppkm_dataset.csv \
  /user/hduser/scala-ppkm-out

hduser@localhost:~$ hdfs dfs -cat /user/hduser/scala-ppkm-out/part-* | wc -l
1671
hduser@localhost:~$ hdfs dfs -cat /user/hduser/scala-ppkm-out/part-*

137,ppkm
111,mikro
105,positif
100,negatif
100,netral
93,covid
88,dan
84,co
83,t
82,https
<output trimmed for brevity>
```

#### 5 Добавьте в WordCount возможность через конфигурацию задать список стоп-слов, которые будут отфильтрованы во время работы приложения (2 балла)

##### 1. Сборка и копирование WordCount

```
$ git checkout 86ef2525
$ cd 04-spark-sql/WordCount
$ sbt package
$ docker cp ./**/wordcount_2.11-0.4.0-SNAPSHOT.jar zeppelin:/home/hduser/
```

##### 2. Запуск Spark

```
hduser@localhost:~$ hdfs dfs -rm -r /user/hduser/scala-ppkm-out
hduser@localhost:~$ spark-submit \
  --name WordCount \
  --class ok.ml.WordCount \
  --master yarn \
  --deploy-mode cluster \
  wordcount_2.11-0.4.0-SNAPSHOT.jar \
  /user/hduser/ppkm/ppkm_dataset.csv \
  /user/hduser/scala-ppkm-out \
  /user/hduser/ppkm/stopwordv1.txt
```

```
hduser@localhost:~$ hdfs dfs -cat /user/hduser/scala-ppkm-out/part-* | wc -l
1499
hduser@localhost:~$ hdfs dfs -cat /user/hduser/scala-ppkm-out/part-*
137,ppkm
111,mikro
105,positif
100,negatif
100,netral
93,covid
88,dan
84,co
<output trimmed for brevity>
```

## 6 Почему в примере в выходном файле получилось 200 партий? (3 балла)

`spark.sql.shuffle.partitions` по умолчанию 200, и это значение используется, если в `DataFrame` используются операции, предусматривающие shuffle данных, например, `union()`, `groupBy()`, `join()` и т.д.

Изменить это значение, например, на 50 можно командой

```
spark.conf.set("spark.sql.shuffle.partitions", "50")
```

([источник](#))