Лабораторная работа №5 «Применение методов машинного обучения к данным из хранилища больших данных»

Выполнили студенты группы ИТ-50916: Кураженкова О.С., Ставских А.Д., Тюленева Е.М.

Получим список файлов в папке lab4 на удалённом сервере в hdfs. В директории test1out находится выход задачи подсчёта количества каждого слова в книге «Alice.txt».

Просмотрим файл «part-r-00000» командой hdfs dfs -cat

```
administrator@clickhousestudent:~$ hdfs dfs -cat hdfs://192.168.121.147:9000/lab4/test1out/part-r-00000
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:
/home/administrator/hadoop-2.8.5/share/hadoop/common/lib/hadoop-auth-2.8.5.jar) to method sun.security.k
rb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.
util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
"Defects,"      1
"Information    1
"Plain  2
"Project        5
"Right  1
#11]    1
$5,000) 1
'AS-IS',        1
("the   1
($1     1
(801)   1
(Alice  4
(And,   1
(As     1
(Before 1
(Dinah  1
(For,   1
(He     1
(If     1
(In     1
(It     1
(Sounds 1
(The    3
(We     1
(Which  2
(_with  1
(a      1
(a)     1
(and    2
(any    1
(as     2
(b)     1
(c)     1
(does   1
(for    1
(for,   1
(if     1
(it     4
(look   1
```

Каждому слову соответствует количество повторений в книге. Слова упорядочены в алфавитном порядке.

С помощью команды awk '{print $1":"$2}' разобьём содержимое файла «part-r-00000» на колонки с разделителем «:», добавив с помощью grep фильтр «the»

```
administrator@clickhousestudent: ~                                    —    □    ✕

administrator@clickhousestudent:~$ hdfs dfs -cat hdfs://192.168.121.147:9000/lab4/test1out/part-r-00000
| awk '{print $1":"$2}' | grep the
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:
/home/administrator/hadoop-2.8.5/share/hadoop/common/lib/hadoop-auth-2.8.5.jar) to method sun.security.k
rb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.
util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
("the:1
(the:1
("the:1
Atheling:1
Drawling—the:1
Either:1
Father:4
Jack-in-the-box,:1
Panther:3
_their_:1
_then_—she:1
_there_:1
_these?_":1
altogether:1
altogether,:2
altogether.:1
altogether;:1
another:21
another!:1
baby—the:1
bank—the:1
bother:1
breathe:1
breathe.":1
breathe'!":1
brother's:1
by—the:1
either:6
either!":1
either,:2
farther:1
father,:1
father;:1
feather:1
feathers,:1
first—they:1
further:2
further.:1
further::1
gather:1
```

Отсортируем полученные данные по второму столбцу с помощью команды sort:

hdfs dfs -cat hdfs://192.168.121.147:9000/lab4/test1out/part-r-00000 | awk '{print $1":"$2}' | grep the | sort -t: -nk2

```
other:40
there:44
them:49
their:50
then:50
they:110
the:1677
administrator@clickhousestudent:~$
```

Найдём количество слов «Alice». Корректный ответ мешают получить знаки препинания перед и после имени.


```
administrator@clickhousestudent:~$ hdfs dfs -cat hdfs://192.168.121.147:9000/lab4/test1out/part-r-00000
| awk '{print $1":"$2}' | grep Alice
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:
/home/administrator/hadoop-2.8.5/share/hadoop/common/lib/hadoop-auth-2.8.5.jar) to method sun.security.k
rb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.
util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
(Alice:4
Alice:221
Alice!:3
Alice!":1
Alice)—"and:1
Alice,:76
Alice,):2
Alice.:54
Alice::7
Alice;:16
Alice's:14
Alice's,:1
_Alice's:1
"Alice!":1
```

Модифицируем код awk для удаления знаков препинания и суммирования количеств слов Alice: awk '{ sum +=$2 } END {print "Alice:"sum}'


```
administrator@clickhousestudent:~$ hdfs dfs -cat hdfs://192.168.121.147:9000/lab4/test1out/part-r-00000
| grep Alice | awk '{ sum+=$2 } END {print "Alice:"sum}'
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:
/home/administrator/hadoop-2.8.5/share/hadoop/common/lib/hadoop-auth-2.8.5.jar) to method sun.security.k
rb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.
util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Alice:402
```

Создадим файл «.509579.tmp»

```
administrator@clickhousestudent:~$ cat .509579.tmp
ИТ-50916
Кураженкова О.С.
https://github.com/Teytress/BigData
Ставских А.Д.
https://github.com/KarenKrill/BigData
Тюленева Е.М.
https://github.com/H0mura/BigData
```

Запишем файл в hdfs с помощью команды put

```
administrator@clickhousestudent:~$ hdfs dfs -put .509579.tmp hdfs://192.168.121.147:9000/exam/509579
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/hom
krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
administrator@clickhousestudent:~$ hdfs dfs -ls hdfs://192.168.121.147:9000/exam/
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/hom
krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Found 7 items
-rw-r--r--   3 administrator supergroup        387 2020-12-16 14:55 hdfs://192.168.121.147:9000/exam/11962
-rw-r--r--   3 administrator supergroup        351 2020-12-17 09:50 hdfs://192.168.121.147:9000/exam/409007
-rw-r--r--   3 administrator supergroup        448 2020-12-22 10:56 hdfs://192.168.121.147:9000/exam/5091614
-rw-r--r--   3 administrator supergroup        291 2020-12-16 07:53 hdfs://192.168.121.147:9000/exam/50916ke
-rw-r--r--   3 administrator supergroup        197 2020-12-23 05:34 hdfs://192.168.121.147:9000/exam/509579
-rw-r--r--   3 administrator supergroup        351 2020-12-17 09:51 hdfs://192.168.121.147:9000/exam/lab4
-rw-r--r--   3 administrator supergroup        411 2020-12-16 13:03 hdfs://192.168.121.147:9000/exam/shalopa
administrator@clickhousestudent:~$ hdfs dfs -cat hdfs://192.168.121.147:9000/exam/509579
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/hom
krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
ИТ-50916
Кураженкова О.С.
https://github.com/Teytress/BigData
Ставских А.Д.
https://github.com/KarenKrill/BigData
Тюленева Е.М.
https://github.com/H0mura/BigData
administrator@clickhousestudent:~$
```

Записать файл в «/lab4» не удалось, т.к. владельцем этой директории является
«hadoop».

```
administrator@clickhousestudent:~$ hdfs dfs -put .509579.tmp hdfs://192.168.121.147:9000/lab4/509579
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file
/home/administrator/hadoop-2.8.5/share/hadoop/common/lib/hadoop-auth-2.8.5.jar) to method sun.security.
rb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication
util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
put: Permission denied: user=administrator, access=WRITE, inode="/lab4":hadoop:supergroup:drwxr-xr-x
administrator@clickhousestudent:~$
```

```
administrator@clickhousestudent:~$ hdfs dfs -ls hdfs://192.168.121.147:9000/ | grep -i -e items -e exam -e lab4
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/ad
krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.Ker
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
Found 18 items
drwxr-xr-x   - hadoop        supergroup          0 2020-08-19 12:24 hdfs://192.168.121.147:9000/copy_from_local_
drwxr-xr-x   - administrator supergroup          0 2020-12-23 05:34 hdfs://192.168.121.147:9000/exam
drwxr-xr-x   - hadoop        supergroup          0 2020-12-21 14:22 hdfs://192.168.121.147:9000/lab4
```