UNIVERSITY OF NICE, SOPHIA ANTIPOLIS

Project in Spark 2017

Adrianna Janik

Ion Mosnoi

Lei Guo

October 16, 2017

1 TASK

Firstly we uncompressed the data stored in ling-spam.zip folder with *Extract all* command. Secondly we open Virtual Box machine with Hortonworks, we signed in with maria_dev username and maria_dev password on Ambari available under 127.0.0.1:8080 ip address. We have selected *Files view*, than navigated to */tmp* folder and created directories *tmp/ling-spam/ham* and *ling-spam/spam*. Following that we logged in with ssh credentials to Hortonworks machine

```
$ssh root@127.0.0.1 –p 2222
```

In the meantime upload to the virtual machine ling-spam.zip with:

```
$sudo scp –P 2222 ../ling–spam.zip  root@127.0.0.1:/tmp/
```

We unzipped ling-spam.zip with:

```
$unzip ling–spam.zip –d /tmp/ling–spam
```

We putted files into /tmp/ling-spam/ folder in hdfs with:

```
$hdfs dfs –put ./ling–spam/ham /tmp/ling–spam/ham
$hdfs dfs –put ./ling–spam/spam /tmp/ling–spam/spam
```

2 TASK

Installation of sbt:

```
$wget http://dl.bintray.com/sbt/rpm/sbt−0.13.12.rpm
```

Edit file /etc/yum.repos.d/sandbox.repo:

```
~[sandbox]
~name=Sandbox repository (tutorials)
~gpgcheck=0
~enabled=0
~baseurl=http://dev2.hortonworks.com.s3.amazonaws.com/repo/dev/master/utils/

$yum clean all
$yum update
$sudo yum localinstall sbt−0.13.12.rpm
$sbt −update
$sudo scp −P 2222 −r ../spamTopWords/* root@127.0.0.1:/tmp/spamTopWords/
$sbt package
```

3 Task

4 Task

5 Task

6 Scala useful functions

- holeTextFiles -
- map -
- flatMapValues -
- mapValues -
- filter -
- case -
- reduce -
- reduceByKey -
- Ordering.by -
- fullOuterJoin -
- join -
- leftOuterJoin -

- getOrElse -
- math.log -
- toDouble -
- toSet -
- takeOrdered -
- foreach -
- swap -