2018W02.text-anal...

```
println("Spark version " + sc.version)
println("Scala version " + util.Properties.versionString)
```

FINISHED

```
Spark version 2.1.0
Scala version version 2.11.8
```

Took 0 sec. Last updated by anonymous at January 11 2018, 11:04:47 AM.

```
// Global variables
val dir_data = "data/dev-test-exploratory-analysis"
val fcomm_clean = dir_data + "/224564804326967_facebook_comments_clean/part-*"
// End of Global variables
```

FINISHED

```
dir_data: String = data/dev-test-exploratory-analysis
fcomm_clean: String = data/dev-test-exploratory-analysis/224564804326967_facebook_comments_cl
ean/part-*
```

Took 0 sec. Last updated by anonymous at January 11 2018, 11:04:47 AM.

```
// Read text file
val commentsDF = sc.textFile(fcomm_clean)
                    .filter(_.nonEmpty)
                   .filter(l => !l.contains("null"))
                   .filter(l => !l.contains("0"))
                   .filter(s => s.length > 1)
```

FINISHED

```
commentsDF: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[1080] at filter at <console>:
133
```

Took 1 sec. Last updated by anonymous at January 11 2018, 11:04:48 AM.

```
// View data
commentsDF.take(7).map(s => println(s))
```

FINISHED

```
byt zaujimavy cr kupovat vajicko 89 9 korunk d nevediet stat sk z sliepkama d d d byt kriz d
 d d
naozaj pomaly velmi pomaly zvysovanie cena mat rychly tempo
julius holz zaujimavy byt fakt vajickovy kriz nekonat nemecky
konecna byt cas
jak kriz kriz sty urobit dotedy nekupit vajce kto nebyt povodny cena
buda asi ta ochutene holandsky slovensky asi nic nechovat treba vozit hnoj eu
rychlo cena hora dol pomalicky taky slovensky vajce dostatok teraz vyzmykat clovek dat nenazr
anec
res625: Array[Unit] = Array((), (), (), (), (), (), ())
```

Took 0 sec. Last updated by anonymous at January 11 2018, 11:04:48 AM.

```
import org.apache.spark.sql.functions.desc
// Count words
val wordCountMap = commentsDF
      .flatMap(_.split(" "))
      .map(w => (w, 1))
```

FINISHED

```
      .countByKey()
// Convert Map to Dataset
val wordCountDF = sc.parallelize(wordCountMap.toSeq)
                        .toDF("word", "count")
                        .sort(desc("count"))
wordCountDF.show(50)
```

```
import org.apache.spark.sql.functions.desc
wordCountMap: scala.collection.Map[String,Long] = Map(zubac -> 1, klasicky -> 1, oba -> 1,
 fenomen -> 1, htm -> 1, malatin -> 1, 45 -> 1, vlhkost -> 1, chybajuci -> 1, formulacia ->
 3, zobudit -> 2, vacsina -> 3, chybat -> 1, usmrtenou -> 1, ocenovat -> 1, vyborny -> 2, n
enechat -> 1, najst -> 6, spravanie -> 4, teplice -> 1, vyskusat -> 2, teplaky -> 1, zdravi
t -> 4, dodatocny -> 2, predajny -> 33, huraa -> 1, e -> 4, vona -> 1, praca -> 2, trcat ->
 1, setrit -> 1, zviera -> 1, vysoka -> 8, pavuk -> 1, nehadzat -> 2, deravy -> 1, nemislit
e -> 1, politika -> 2, nedostatok -> 1, poistka -> 1, evicka -> 1, macka -> 1, daky -> 2, b
abatko -> 1, inak -> 4, vcas -> 1, pocitit -> 1, suseda -> 2, hora -> 5, obalenie -> 2, sve
tovladcou -> 1, vzatych -> 1, vlacik -> 5, zlacniet -> 1, odstavovac -...
wordCountDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [word: string, count:
 bigint]
+---------+-----+
|     word|count|
+---------+-----+
|      byt|  554|
|     cena|  158|
|       on|  122|
```

Took 1 sec. Last updated by anonymous at January 11 2018, 11:04:49 AM.

FINISHED

```
import org.apache.spark.sql.Encoders
// Sum of all word counts
val sumWordsDF = wordCountDF.agg(sum("count"))
sumWordsDF.show()
val sumWordsDouble = sumWordsDF.as(Encoders.DOUBLE).collect()
println("Number of words " + sumWordsDouble(0))
```

```
import org.apache.spark.sql.Encoders
sumWordsDF: org.apache.spark.sql.DataFrame = [sum(count): bigint]
+----------+
|sum(count)|
+----------+
|      9710|
+----------+
sumWordsDouble: Array[Double] = Array(9710.0)
Number of words 9710.0
```

Took 1 sec. Last updated by anonymous at January 11 2018, 11:04:50 AM.

READY