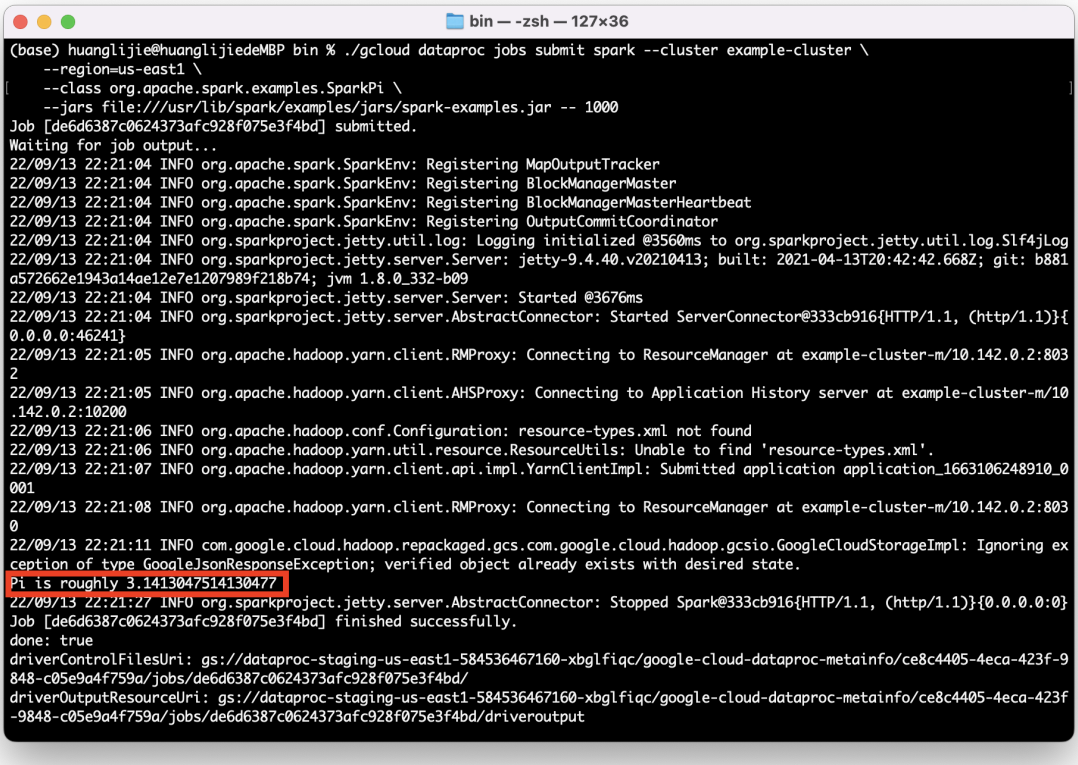
EECS E6893 Big Data Analytics HW0

Lijie Huang, UNI: lh3158
Sep 14, 2022

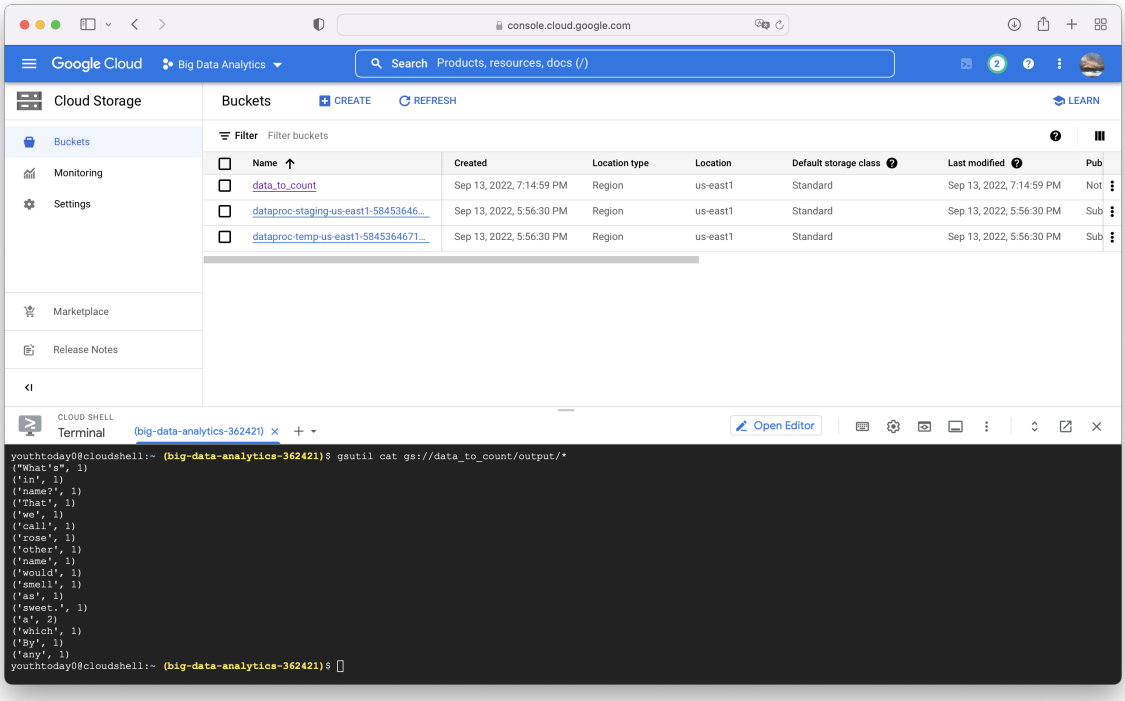
1. Warm-up Exercises

(1) Provide screenshots to prove you've completed the exercises

Exercise 3 Pi Calculation:



Exercise 4 Word Count Job:



(2) List the Spark transformations and actions involved in each exercise. Identify the RDD operation that triggers the program to execute.

- Pi Calculation Job:

SparkPi.scala

```scala
// scalastyle:off println
package org.apache.spark.examples
import scala.math.random
import org.apache.spark.sql.SparkSession

/** Computes an approximation to pi */
object SparkPi {
  def main(args: Array[String]): Unit = {
    val spark = SparkSession
      .builder
      .appName("Spark Pi")
      .getOrCreate()
    val slices = if (args.length > 0) args(0).toInt else 2
    val n = math.min(100000L * slices, Int.MaxValue).toInt // avoid overflow
    val count = spark.sparkContext.parallelize(1 until n, slices).map { i =>
      val x = random * 2 - 1
      val y = random * 2 - 1
      if (x*x + y*y <= 1) 1 else 0
    }.reduce(_ + _)
    println(s"Pi is roughly ${4.0 * count / (n - 1)}")
    spark.stop()
  }
}
// scalastyle:on println
```

On line 32, a RDD is created by parallelizing the data randomly generated. From line 32 to line 36, a transformation “map” and an action “reduce” are processed to count how many points are inside the circle, whose radius is 1. Finally, Pi can be estimated by calculating the probability.

- Word Count Job:

word-count.py

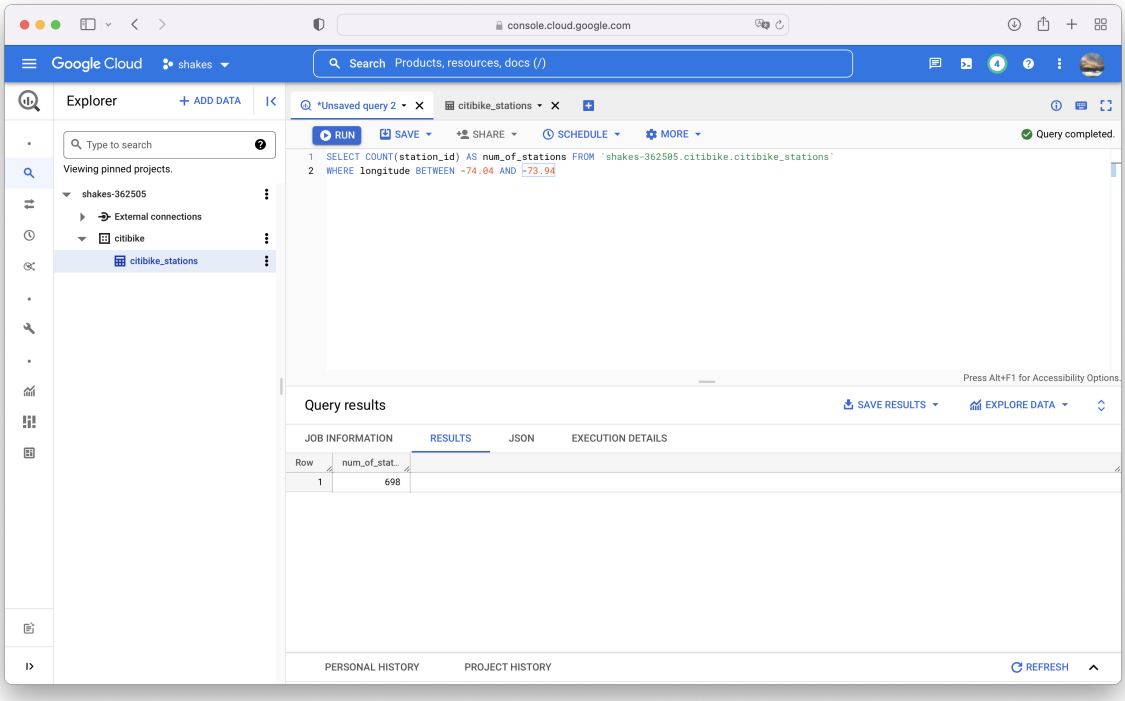
```python
#!/usr/bin/env python

import ...

if len(sys.argv) != 3:
  raise Exception("Exactly 2 arguments are required: <inputUri> <outputUri>")

inputUri=sys.argv[1]
outputUri=sys.argv[2]

sc = pyspark.SparkContext()
lines = sc.textFile(sys.argv[1])
words = lines.flatMap(lambda line: line.split())
wordCounts = words.map(lambda word: (word, 1)).reduceByKey(lambda count1, count2: count1 + count2)
wordCounts.saveAsTextFile(sys.argv[2])
```

On line 13, a RDD is created from the text file by function “textFile”. on line 14 and 15, 3 transformations “flatMap”, “map” and “reduceByKey” are processed to count the frequency of each word. On line 16, an action (”saveAsTextFile”) is used to save the count result.

2. NYC Bike Expert

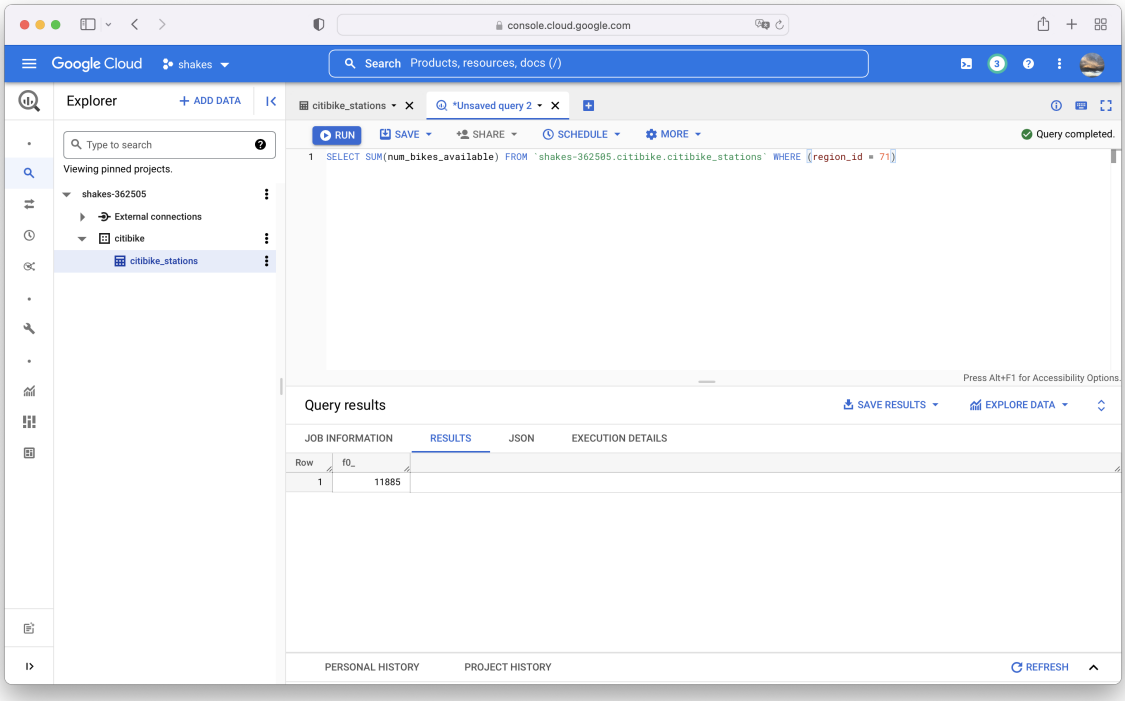
(1) How many stations with longitude between -73.94 and -74.04?

698



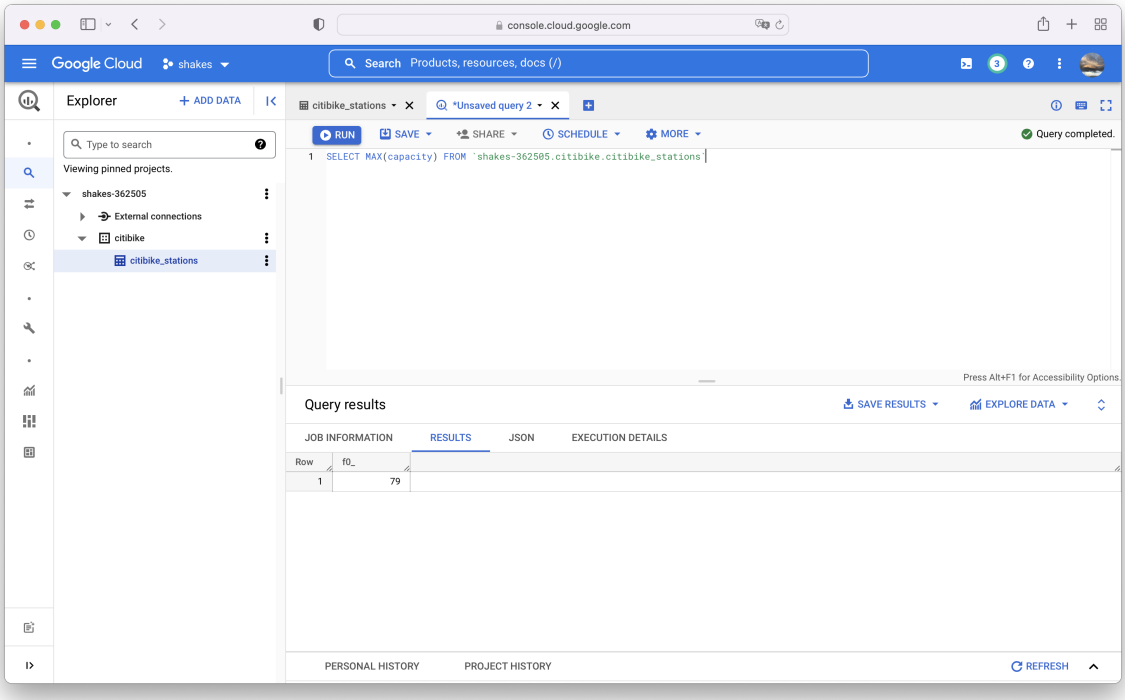
(2) What's the total number of bikes available in region_id 71?

11885

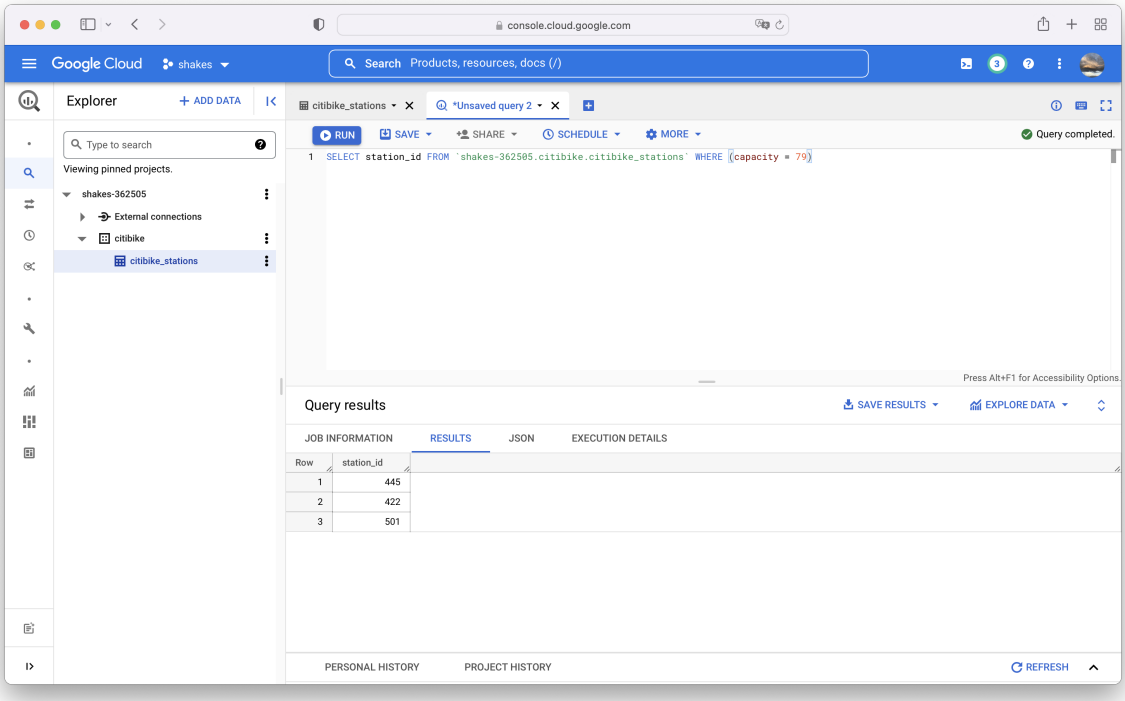


(3) What's the largest capacity for a station? List all the station_id of the stations that have the largest capacity.

largest capacity: 79



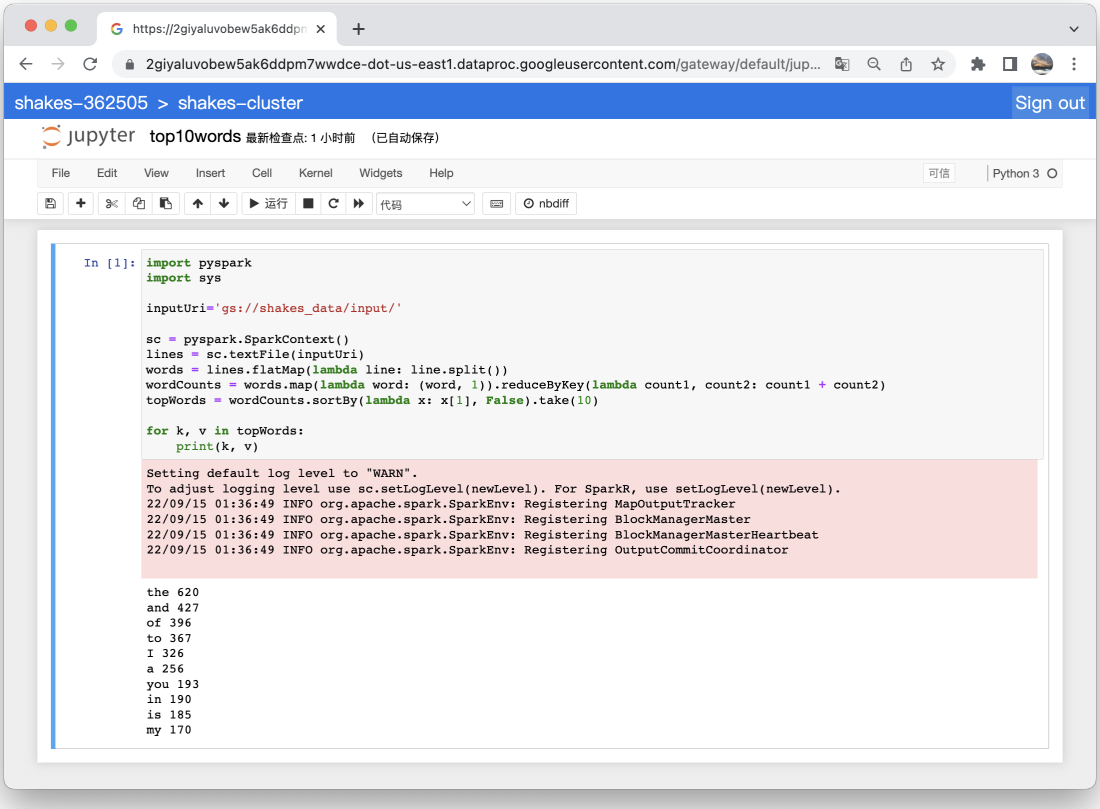
stations with the largest capacity:



3. Understanding William Shakespeare

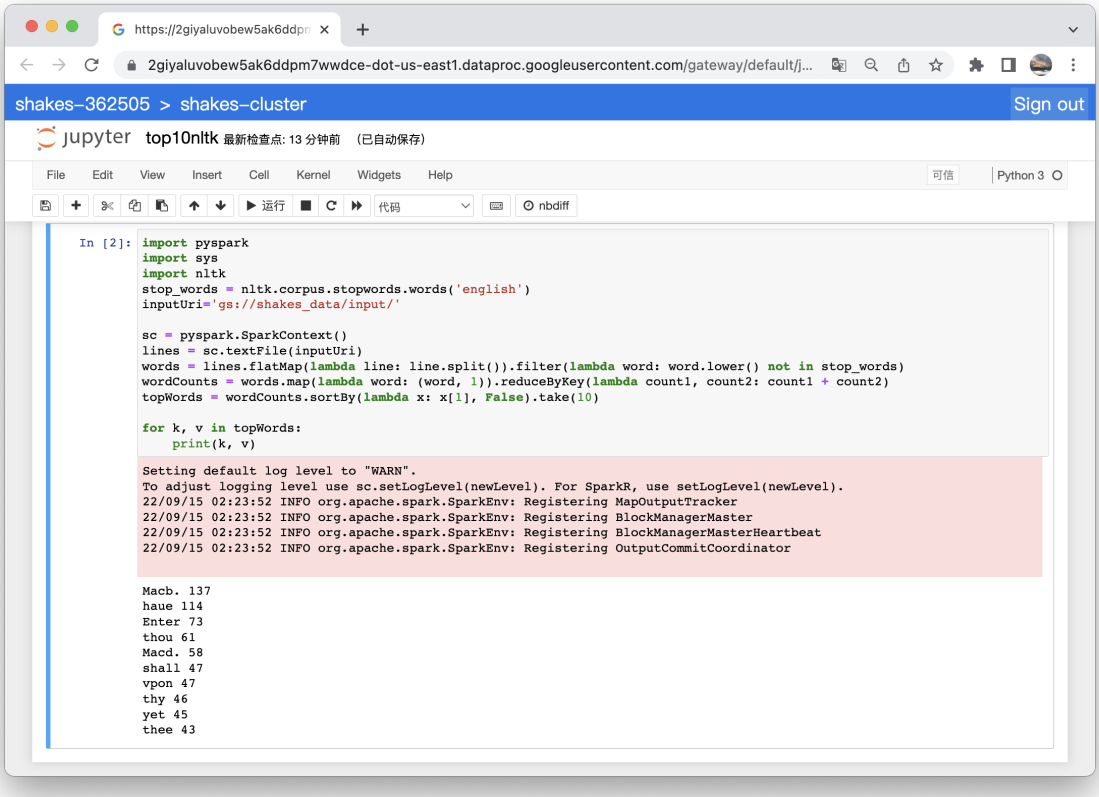
(1) Find top-10 frequent words without any text preprocessing.

```
[(the,620),(and,427),(of,396),(to,367),(I,326),(a,256),(you,193),(in,190),(is,185),(my,170)]
```



(2) Find top-10 frequent words by first filtering out stop words provides by NLTK packages.

```
[(Macb.,137),(haue,114),(Enter,73),(thou,61),(Macd.,58),(shall,47),(vpon,47),(thy,46),(yet,45),(thee,43)]
```


https://2giyaluvobew5ak6ddpn
2giyaluvobew5ak6ddpm7wwdce-dot-us-east1.dataproc.googleusercontent.com/gateway/default/j...
shakes-362505 > shakes-cluster
Sign out
jupyter
top10nltk 最新检查点: 13 分钟前 (已自动保存)
File Edit View Insert Cell Kernel Widgets Help
可信
Python 3
运行
代码
nbdiff
In [2]:
import pyspark
import sys
import nltk
stop_words = nltk.corpus.stopwords.words('english')
inputUri='gs://shakes_data/input/'
sc = pyspark.SparkContext()
lines = sc.textFile(inputUri)
words = lines.flatMap(lambda line: line.split()).filter(lambda word: word.lower() not in stop_words)
wordCounts = words.map(lambda word: (word, 1)).reduceByKey(lambda count1, count2: count1 + count2)
topWords = wordCounts.sortBy(lambda x: x[1], False).take(10)
for k, v in topWords:
print(k, v)
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/09/15 02:23:52 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/09/15 02:23:52 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/09/15 02:23:52 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
22/09/15 02:23:52 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Macb. 137
haue 114
Enter 73
thou 61
Macd. 58
shall 47
vpon 47
thy 46
yet 45
thee 43