COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

EECS E6893 Big Data Analytics

# Homework #0 Report

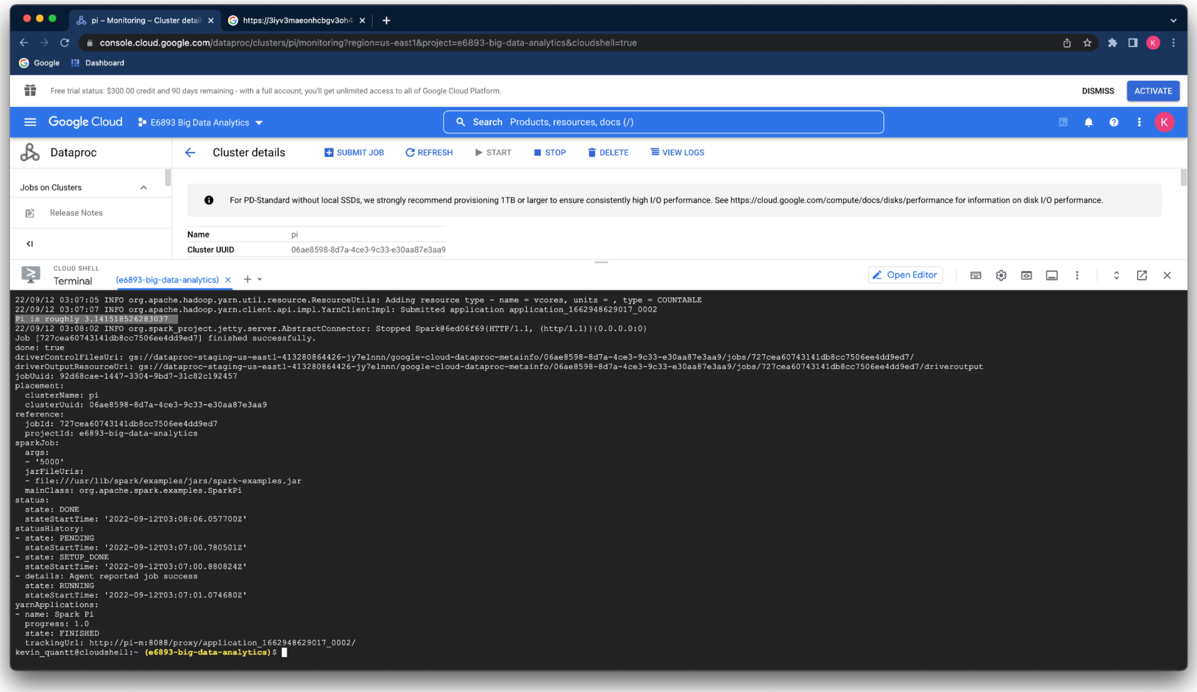
Name: Yi Zhang

Uni: yz4140

September 2022

# Question 1

(1) As shown below, the pi is successfully estimated via the cluster called pi and the value is correct.



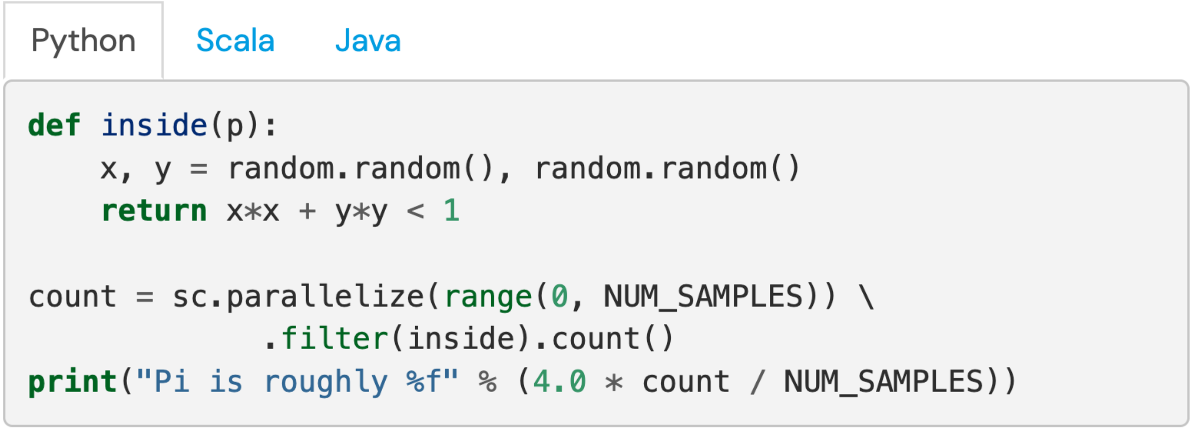
pi – Monitoring – Cluster detail
https://3iyv3maeonhcbgv3oh4
console.cloud.google.com/dataproc/clusters/pi/monitoring?region=us-east1&project=e6893-big-data-analytics&cloudshell=true
Google
Dashboard
Free trial status: $300.00 credit and 90 days remaining - with a full account, you'll get unlimited access to all of Google Cloud Platform.
DISMISS
ACTIVATE
Google Cloud
E6893 Big Data Analytics
Search Products, resources, docs (/)
Dataproc
Cluster details
SUBMIT JOB
REFRESH
START
STOP
DELETE
VIEW LOGS
Jobs on Clusters
Release Notes
For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See https://cloud.google.com/compute/docs/disks/performance for information on disk I/O performance.
CLOUD SHELL
Terminal
(e6893-big-data-analytics)
Open Editor

Name pi

Cluster UUID 06ae8598-8d7a-4ce3-9c33-e30aa87e3aa9

```
22/09/12 03:07:05 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
22/09/12 03:07:07 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1662948629017_0002
Pi is roughly 3.141518526283037
22/09/12 03:08:02 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@6ed06f69{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [727cea60743141db8cc7506ee4dd9ed7] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-east1-413280864426-jy7elnnn/google-cloud-dataproc-metainfo/06ae8598-8d7a-4ce3-9c33-e30aa87e3aa9/jobs/727cea60743141db8cc7506ee4dd9ed7/
driverOutputResourceUri: gs://dataproc-staging-us-east1-413280864426-jy7elnnn/google-cloud-dataproc-metainfo/06ae8598-8d7a-4ce3-9c33-e30aa87e3aa9/jobs/727cea60743141db8cc7506ee4dd9ed7/driveroutput
jobUuid: 92d68cae-1447-3304-9bd7-31c82c192457
placement:
  clusterName: pi
  clusterUuid: 06ae8598-8d7a-4ce3-9c33-e30aa87e3aa9
reference:
  jobId: 727cea60743141db8cc7506ee4dd9ed7
  projectId: e6893-big-data-analytics
sparkJob:
  args:
  - '5000'
  jarFileUris:
  - file:///usr/lib/spark/examples/jars/spark-examples.jar
  mainClass: org.apache.spark.examples.SparkPi
status:
  state: DONE
  stateStartTime: '2022-09-12T03:08:06.057700Z'
statusHistory:
- state: PENDING
  stateStartTime: '2022-09-12T03:07:00.780501Z'
- state: SETUP_DONE
  stateStartTime: '2022-09-12T03:07:00.880824Z'
- details: Agent reported job success
  state: RUNNING
  stateStartTime: '2022-09-12T03:07:01.074680Z'
yarnApplications:
- name: Spark Pi
  progress: 1.0
  state: FINISHED
  trackingUrl: http://pi-m:8088/proxy/application_1662948629017_0002/
kevin_quantt@cloudshell:~ (e6893-big-data-analytics)$
```

(2) According to the source code provided by the Spark below

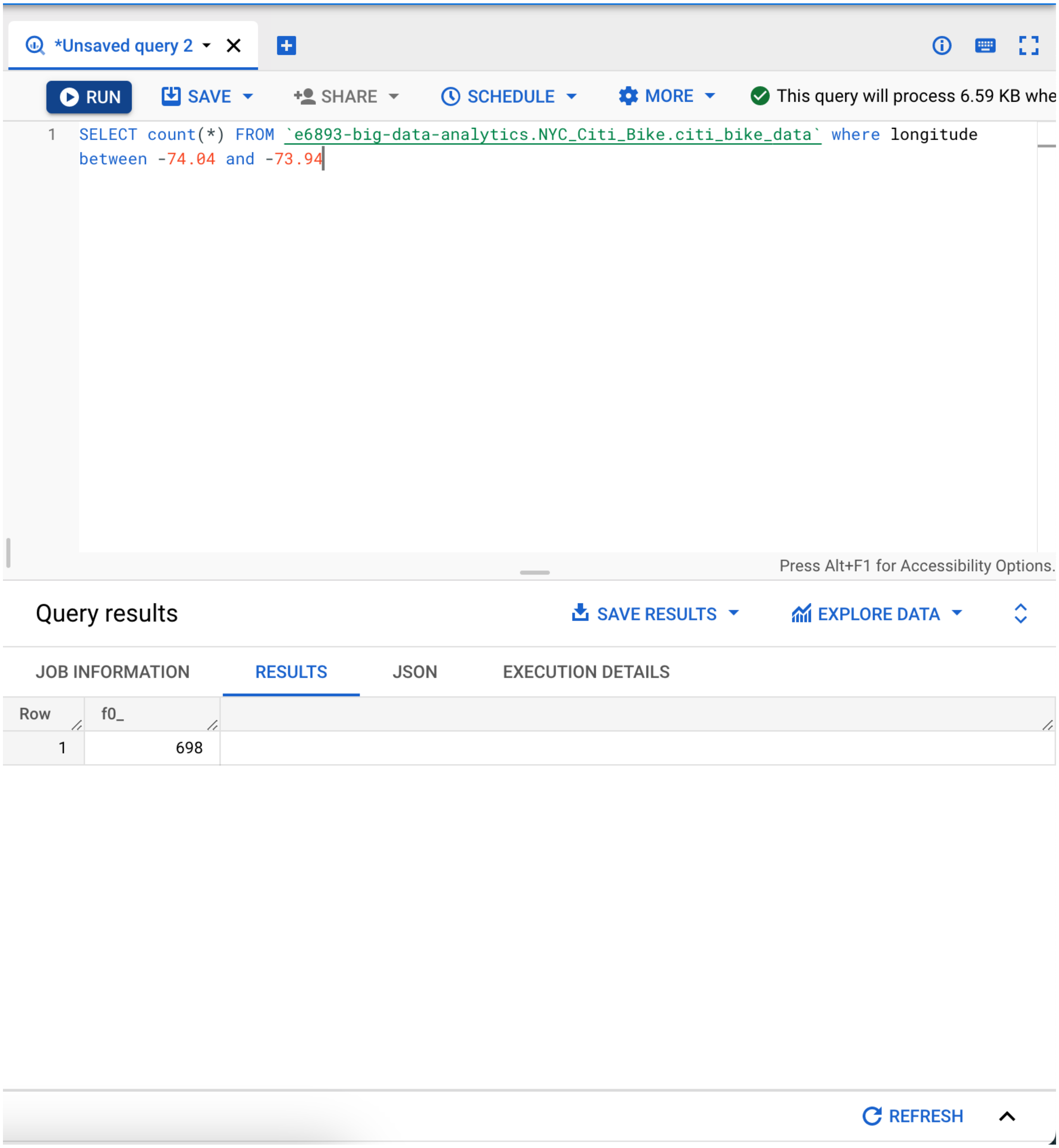


```
def inside(p):
    x, y = random.random(), random.random()
    return x*x + y*y < 1

count = sc.parallelize(range(0, NUM_SAMPLES)) \
             .filter(inside).count()
print("Pi is roughly %f" % (4.0 * count / NUM_SAMPLES))
```

The transformation is ‘**filter**’, which helps to create a new dataset containing the (x,y) within the circle. The action is **‘count’**, which counts the number of data points in the filtered dataset.

## Question 2

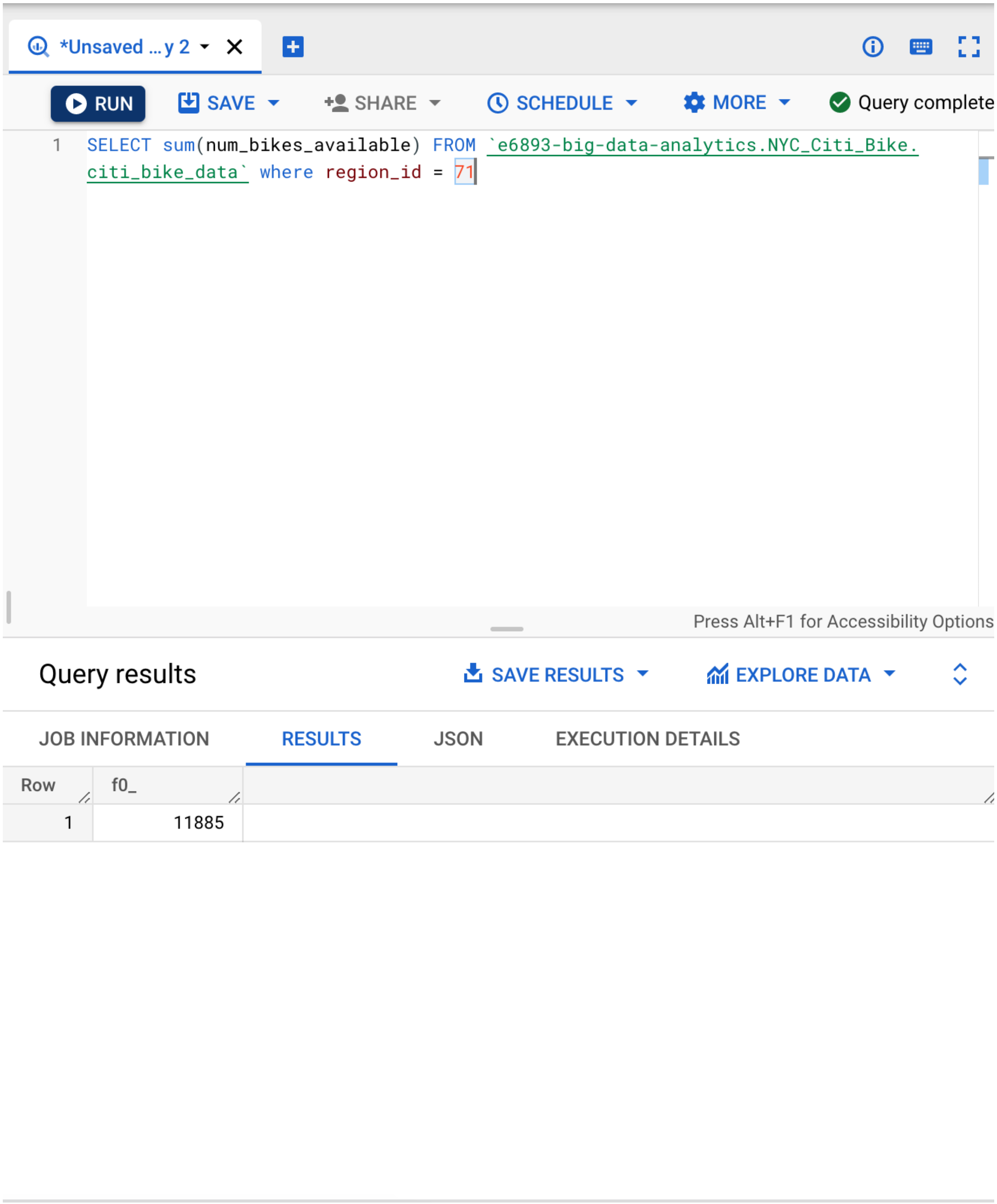
(1) There are 698 stations whose longitude is between -74.04 and -73.94



```
SELECT count(*) FROM `e6893-big-data-analytics.NYC_Citi_Bike.citi_bike_data` where longitude
between -74.04 and -73.94
```

| Row | f0_ |
|---|---|
| 1 | 698 |

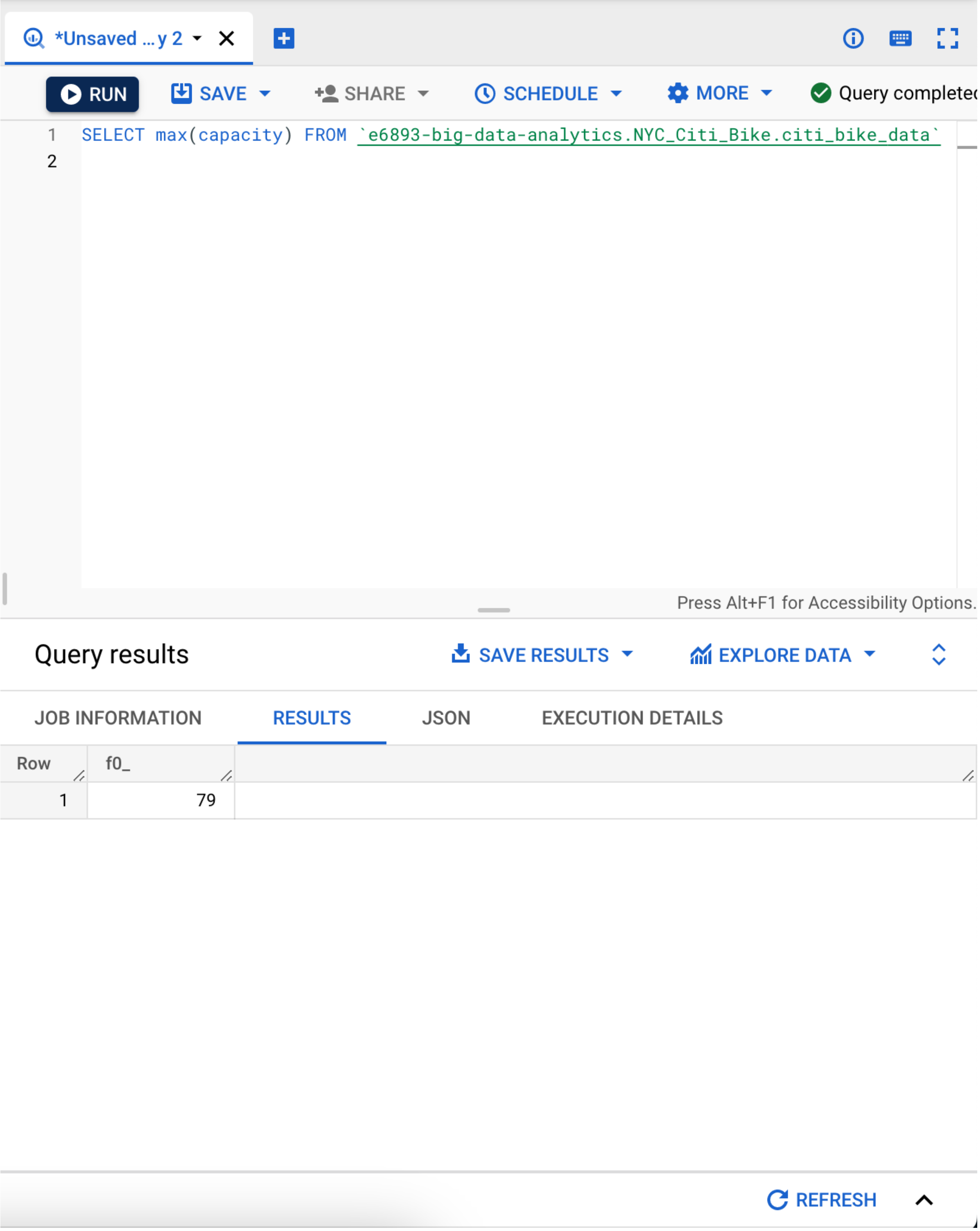
(2) There are 11885 bikes available in region 71



```sql
SELECT sum(num_bikes_available) FROM `e6893-big-data-analytics.NYC_Citi_Bike.
citi_bike_data` where region_id = 71
```

| Row | f0_ |
|---|---|
| 1 | 11885 |

(3) The largest capacity is 79 and station 445, 422, and 501 attain the largest capacity.

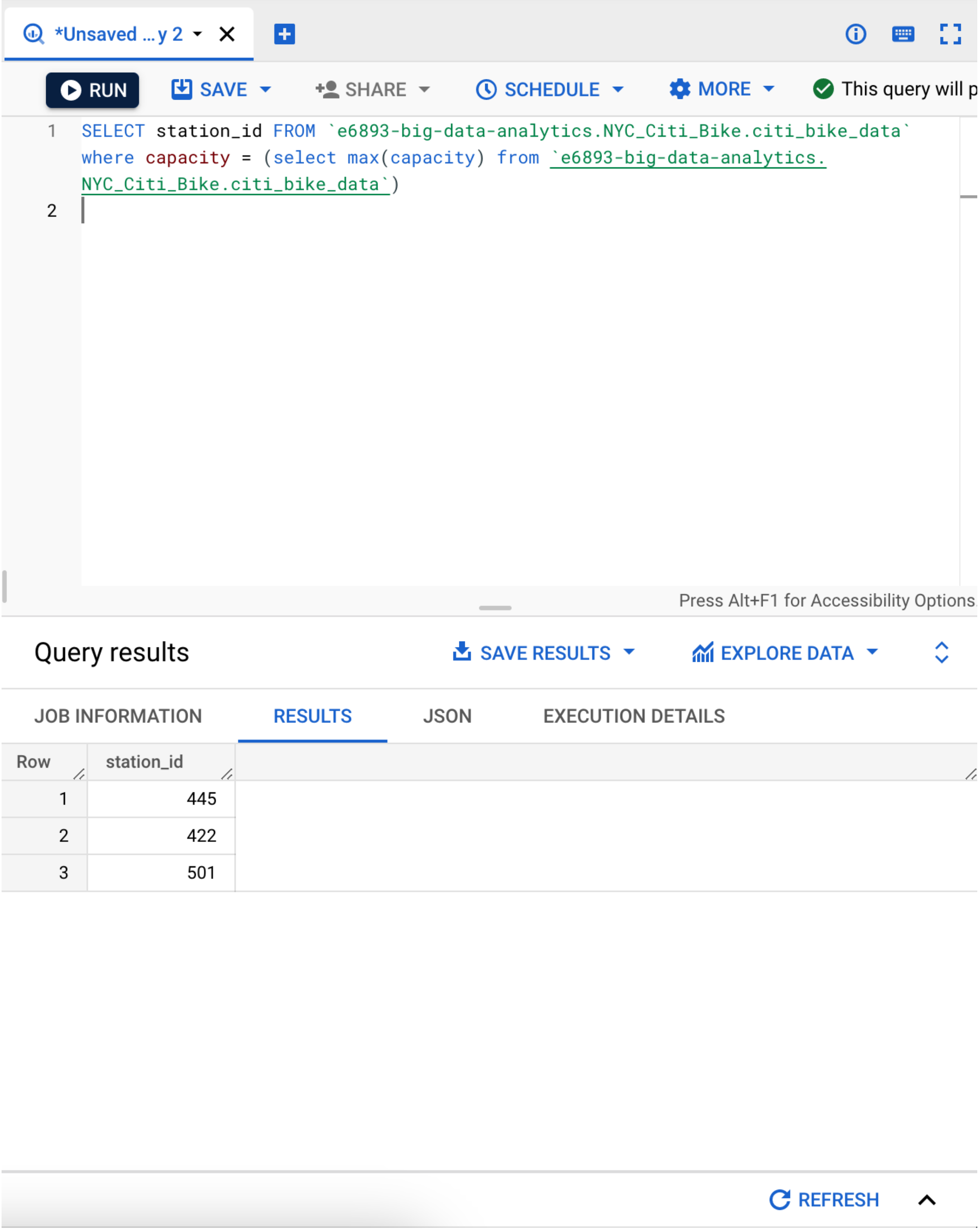


Query complete

```
SELECT max(capacity) FROM `e6893-big-data-analytics.NYC_Citi_Bike.citi_bike_data`

```

Press Alt+F1 for Accessibility Options.

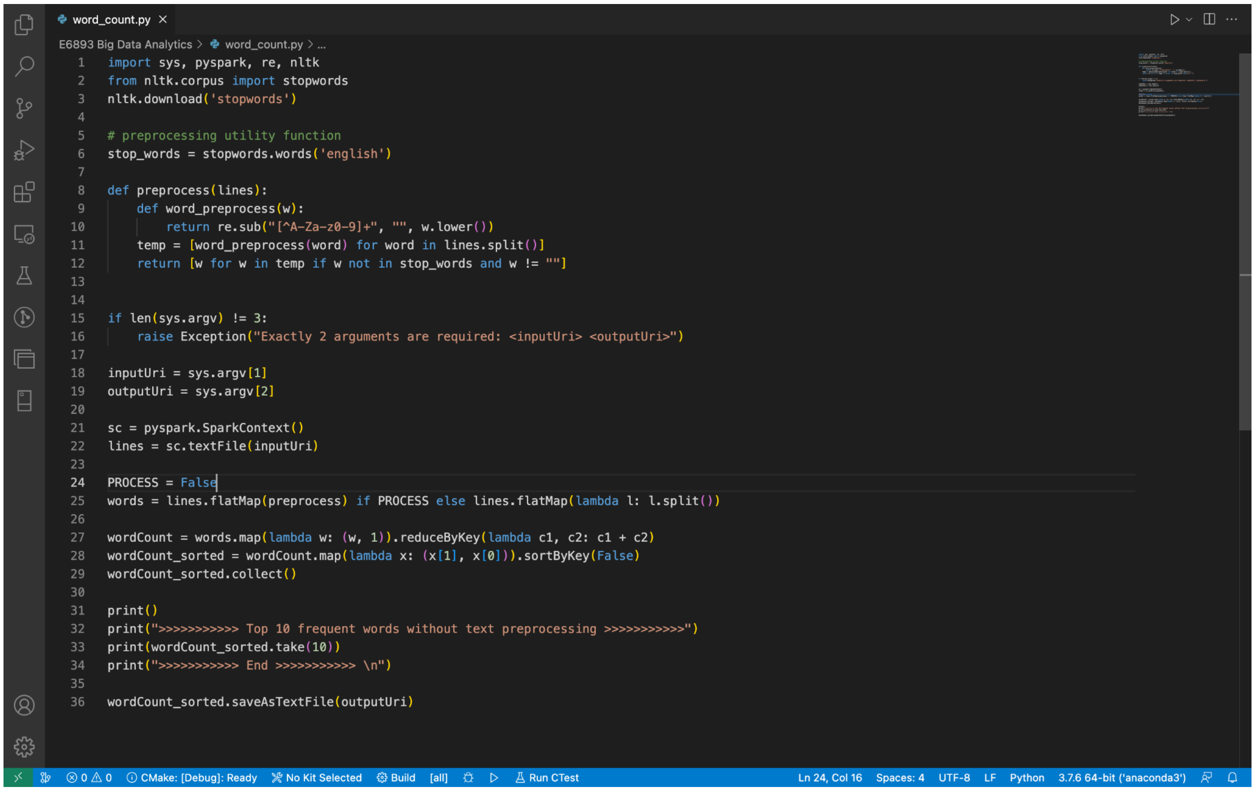
| Row | f0_ |
|---|---|
| 1 | 79 |



```sql
SELECT station_id FROM `e6893-big-data-analytics.NYC_Citi_Bike.citi_bike_data`
where capacity = (select max(capacity) from `e6893-big-data-analytics.
NYC_Citi_Bike.citi_bike_data`)
```

| Row | station_id |
|---|---|
| 1 | 445 |
| 2 | 422 |
| 3 | 501 |

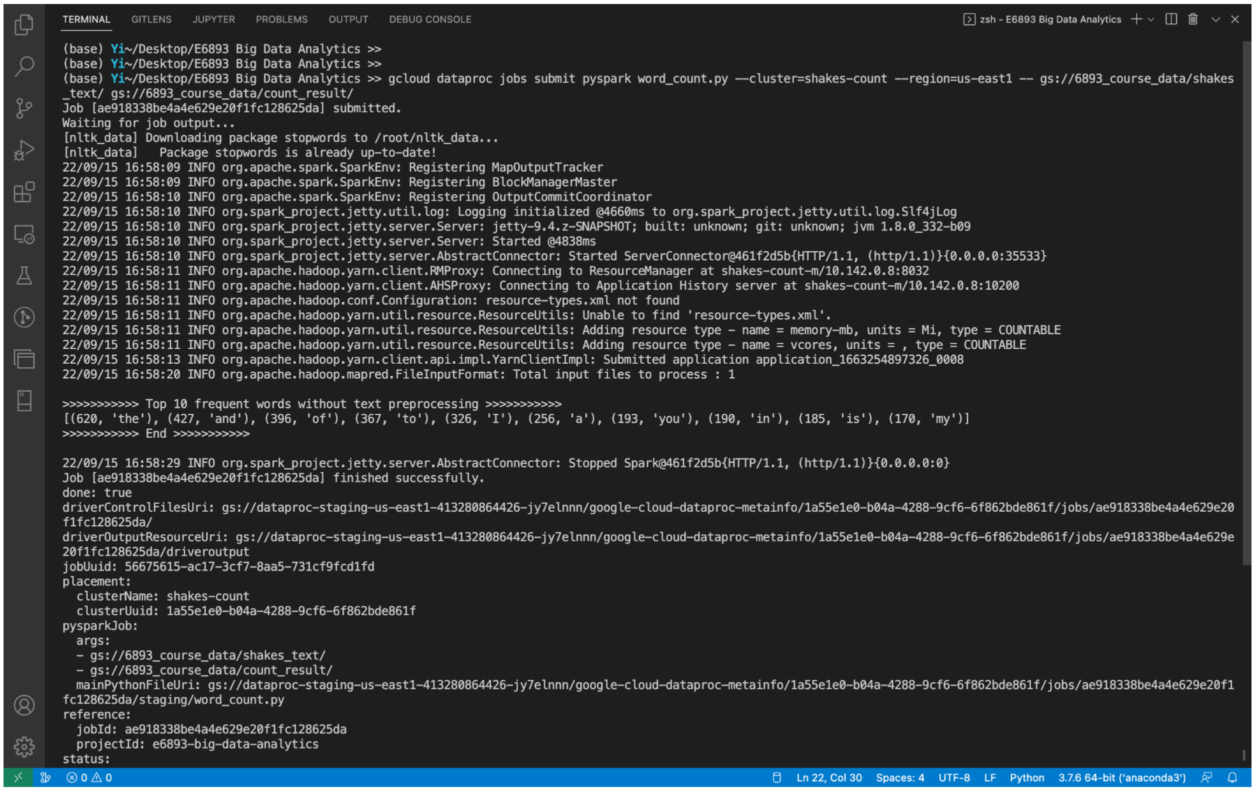
# Question 3



```python
import sys, pyspark, re, nltk
from nltk.corpus import stopwords
nltk.download('stopwords')

# preprocessing utility function
stop_words = stopwords.words('english')

def preprocess(lines):
    def word_preprocess(w):
        return re.sub("[^A-Za-z0-9]+", "", w.lower())
    temp = [word_preprocess(word) for word in lines.split()]
    return [w for w in temp if w not in stop_words and w != ""]


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(inputUri)

PROCESS = False
words = lines.flatMap(preprocess) if PROCESS else lines.flatMap(lambda l: l.split())

wordCount = words.map(lambda w: (w, 1)).reduceByKey(lambda c1, c2: c1 + c2)
wordCount_sorted = wordCount.map(lambda x: (x[1], x[0])).sortByKey(False)
wordCount_sorted.collect()

print()
print(">>>>>>>>>>> Top 10 frequent words without text preprocessing >>>>>>>>>>>")
print(wordCount_sorted.take(10))
print(">>>>>>>>>>> End >>>>>>>>>>> \n")

wordCount_sorted.saveAsTextFile(outputUri)
```

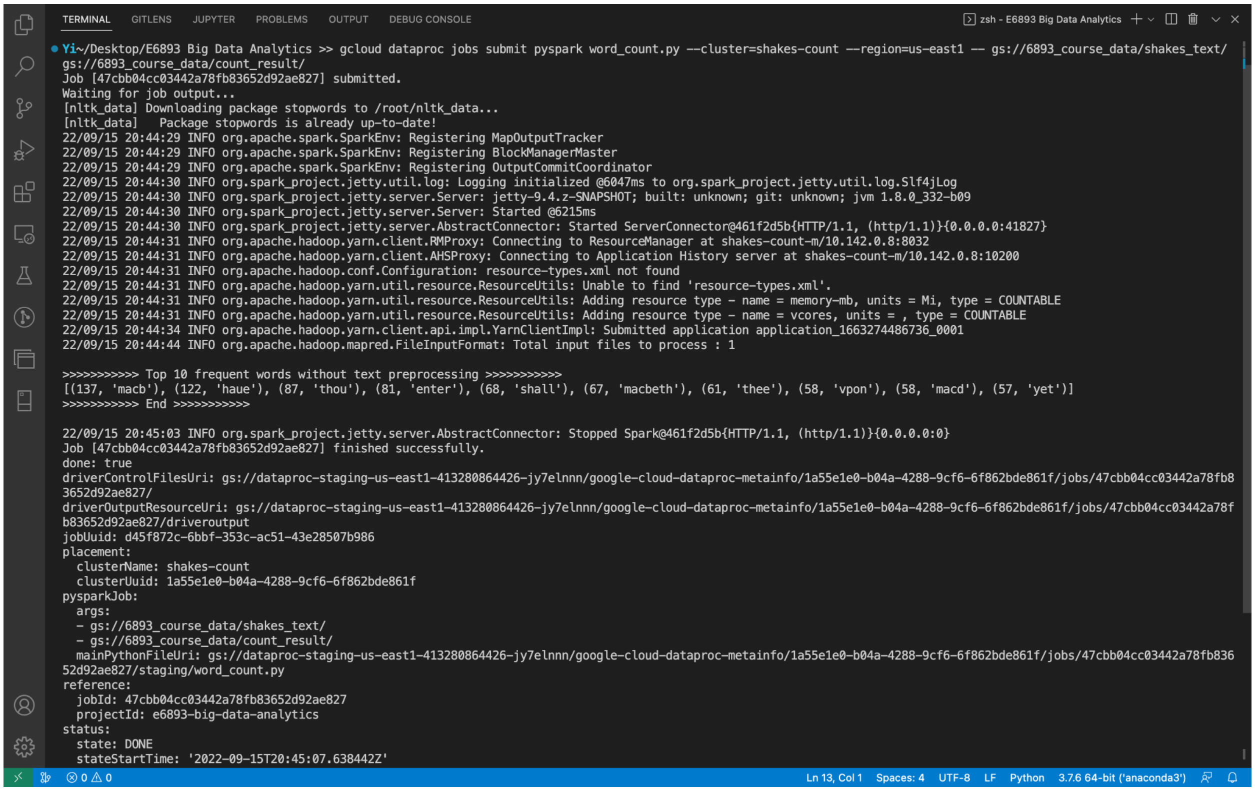
The code above shows the overall workflow of the program. In this exercise, the transformation operations are **‘flatMap’, ‘map’**. The action operations are **‘reduceByKey’, ‘sortByKey’, ‘collect’, ‘take’**.

(1) As the result shows, the top 10 frequent words are ‘the’, ‘and’, ‘of’, ‘to’, ‘I’, ‘a’, ‘you’, ‘in’, ‘is’, ‘my’.



```
(base) Yi~/Desktop/E6893 Big Data Analytics >>
(base) Yi~/Desktop/E6893 Big Data Analytics >>
(base) Yi~/Desktop/E6893 Big Data Analytics >> gcloud dataproc jobs submit pyspark word_count.py --cluster=shakes-count --region=us-east1 -- gs://6893_course_data/shakes
_text/ gs://6893_course_data/count_result/
Job [ae918338be4a4e629e20f1fc128625da] submitted.
Waiting for job output...
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
22/09/15 16:58:09 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/09/15 16:58:09 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/09/15 16:58:10 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
22/09/15 16:58:10 INFO org.spark_project.jetty.util.log: Logging initialized @4660ms to org.spark_project.jetty.util.log.Slf4jLog
22/09/15 16:58:10 INFO org.spark_project.jetty.server.Server: jetty-9.4.z-SNAPSHOT; built: unknown; git: unknown; jvm 1.8.0_332-b09
22/09/15 16:58:10 INFO org.spark_project.jetty.server.Server: Started @4838ms
22/09/15 16:58:10 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerConnector@461f2d5b{HTTP/1.1, (http/1.1)}{0.0.0.0:35533}
22/09/15 16:58:11 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at shakes-count-m/10.142.0.8:8032
22/09/15 16:58:11 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at shakes-count-m/10.142.0.8:10200
22/09/15 16:58:11 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
22/09/15 16:58:11 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/09/15 16:58:11 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
22/09/15 16:58:11 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
22/09/15 16:58:13 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1663254897326_0008
22/09/15 16:58:20 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1

>>>>>>>>>>> Top 10 frequent words without text preprocessing >>>>>>>>>>>
[(620, 'the'), (427, 'and'), (396, 'of'), (367, 'to'), (326, 'I'), (256, 'a'), (193, 'you'), (190, 'in'), (185, 'is'), (170, 'my')]
>>>>>>>>>>> End >>>>>>>>>>>

22/09/15 16:58:29 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@461f2d5b{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [ae918338be4a4e629e20f1fc128625da] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-east1-413280864426-jy7elnnn/google-cloud-dataproc-metainfo/1a55e1e0-b04a-4288-9cf6-6f862bde861f/jobs/ae918338be4a4e629e20
f1fc128625da/
driverOutputResourceUri: gs://dataproc-staging-us-east1-413280864426-jy7elnnn/google-cloud-dataproc-metainfo/1a55e1e0-b04a-4288-9cf6-6f862bde861f/jobs/ae918338be4a4e629e
20f1fc128625da/driveroutput
jobUuid: 56675615-ac17-3cf7-8aa5-731cf9fcd1fd
placement:
  clusterName: shakes-count
  clusterUuid: 1a55e1e0-b04a-4288-9cf6-6f862bde861f
pysparkJob:
  args:
  - gs://6893_course_data/shakes_text/
  - gs://6893_course_data/count_result/
  mainPythonFileUri: gs://dataproc-staging-us-east1-413280864426-jy7elnnn/google-cloud-dataproc-metainfo/1a55e1e0-b04a-4288-9cf6-6f862bde861f/jobs/ae918338be4a4e629e20f1
fc128625da/staging/word_count.py
reference:
  jobId: ae918338be4a4e629e20f1fc128625da
  projectId: e6893-big-data-analytics
status:
```

(2) Setting the **PROCESS** variable to **True**, the program will first do preprocessing such as converting all capital letters to lowercase and filtering out stopwords in English. The top 10 after preprocessing becomes **'macb'**, **'haue'**, **'thou'**, **'enter'**, **'shall'**, **'macbeth'**, **'thee'**, **'vpon'**, **'macd'**, **'yet'**.



```
● Yi~/Desktop/E6893 Big Data Analytics >> gcloud dataproc jobs submit pyspark word_count.py --cluster=shakes-count --region=us-east1 -- gs://6893_course_data/shakes_text/
gs://6893_course_data/count_result/
Job [47cbb04cc03442a78fb83652d92ae827] submitted.
Waiting for job output...
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
22/09/15 20:44:29 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/09/15 20:44:29 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/09/15 20:44:29 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
22/09/15 20:44:30 INFO org.spark_project.jetty.util.log: Logging initialized @6047ms to org.spark_project.jetty.util.log.Slf4jLog
22/09/15 20:44:30 INFO org.spark_project.jetty.server.Server: jetty-9.4.z-SNAPSHOT; built: unknown; git: unknown; jvm 1.8.0_332-b09
22/09/15 20:44:30 INFO org.spark_project.jetty.server.Server: Started @6215ms
22/09/15 20:44:30 INFO org.spark_project.jetty.server.AbstractConnector: Started ServerConnector@461f2d5b{HTTP/1.1, (http/1.1)}{0.0.0.0:41827}
22/09/15 20:44:31 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at shakes-count-m/10.142.0.8:8032
22/09/15 20:44:31 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at shakes-count-m/10.142.0.8:10200
22/09/15 20:44:31 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
22/09/15 20:44:31 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/09/15 20:44:31 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
22/09/15 20:44:31 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
22/09/15 20:44:34 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1663274486736_0001
22/09/15 20:44:44 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1

>>>>>>>>>>> Top 10 frequent words without text preprocessing >>>>>>>>>>>
[(137, 'macb'), (122, 'haue'), (87, 'thou'), (81, 'enter'), (68, 'shall'), (67, 'macbeth'), (61, 'thee'), (58, 'vpon'), (58, 'macd'), (57, 'yet')]
>>>>>>>>>>> End >>>>>>>>>>>

22/09/15 20:45:03 INFO org.spark_project.jetty.server.AbstractConnector: Stopped Spark@461f2d5b{HTTP/1.1, (http/1.1)}{0.0.0.0:0}
Job [47cbb04cc03442a78fb83652d92ae827] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-east1-413280864426-jy7elnnn/google-cloud-dataproc-metainfo/1a55e1e0-b04a-4288-9cf6-6f862bde861f/jobs/47cbb04cc03442a78fb8
3652d92ae827/
driverOutputResourceUri: gs://dataproc-staging-us-east1-413280864426-jy7elnnn/google-cloud-dataproc-metainfo/1a55e1e0-b04a-4288-9cf6-6f862bde861f/jobs/47cbb04cc03442a78f
b83652d92ae827/driveroutput
jobUuid: d45f872c-6bbf-353c-ac51-43e28507b986
placement:
  clusterName: shakes-count
  clusterUuid: 1a55e1e0-b04a-4288-9cf6-6f862bde861f
pysparkJob:
  args:
  - gs://6893_course_data/shakes_text/
  - gs://6893_course_data/count_result/
  mainPythonFileUri: gs://dataproc-staging-us-east1-413280864426-jy7elnnn/google-cloud-dataproc-metainfo/1a55e1e0-b04a-4288-9cf6-6f862bde861f/jobs/47cbb04cc03442a78fb836
52d92ae827/staging/word_count.py
reference:
  jobId: 47cbb04cc03442a78fb83652d92ae827
  projectId: e6893-big-data-analytics
status:
  state: DONE
  stateStartTime: '2022-09-15T20:45:07.638442Z'
```