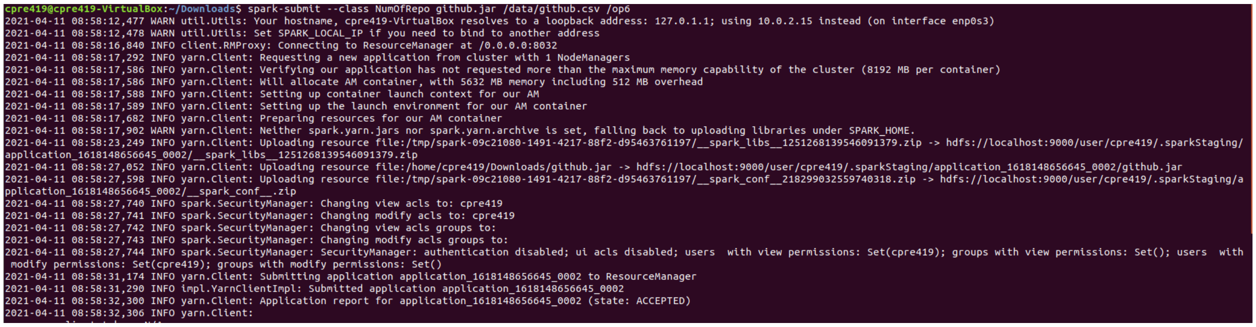
CPR E 419 - Lab 6
Spark

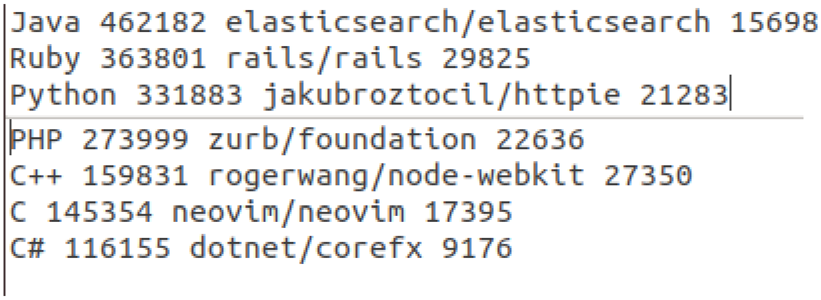
1.

Below are screenshots of the commands run and successful completion of program execution.



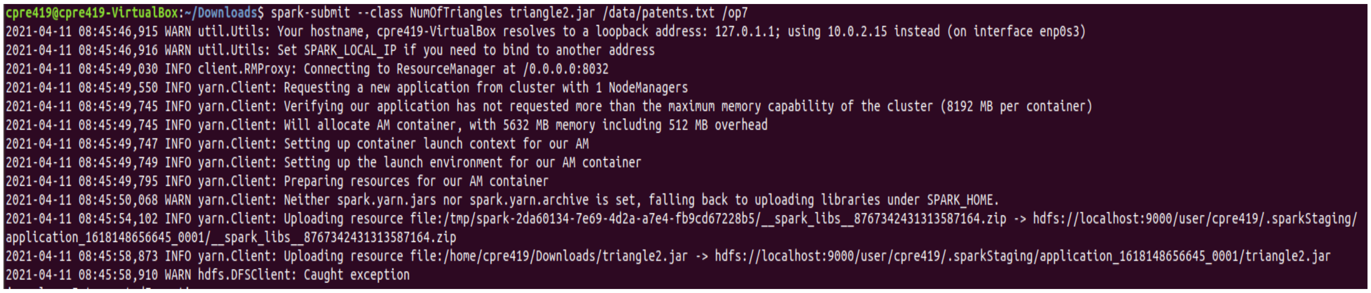
```
cpre419@cpre419-VirtualBox:~/Downloads$ spark-submit --class NumOfRepo github.jar /data/github.csv /op6
2021-04-11 08:58:12,477 WARN util.Utils: Your hostname, cpre419-VirtualBox resolves to a loopback address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
2021-04-11 08:58:12,478 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
2021-04-11 08:58:16,840 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2021-04-11 08:58:17,292 INFO yarn.Client: Requesting a new application from cluster with 1 NodeManagers
2021-04-11 08:58:17,586 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (8192 MB per container)
2021-04-11 08:58:17,586 INFO yarn.Client: Will allocate AM container, with 5632 MB memory including 512 MB overhead
2021-04-11 08:58:17,588 INFO yarn.Client: Setting up container launch context for our AM
2021-04-11 08:58:17,589 INFO yarn.Client: Setting up the launch environment for our AM container
2021-04-11 08:58:17,682 INFO yarn.Client: Preparing resources for our AM container
2021-04-11 08:58:17,902 WARN yarn.Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
2021-04-11 08:58:23,249 INFO yarn.Client: Uploading resource file:/tmp/spark-09c21080-1491-4217-88f2-d95463761197/__spark_libs__1251268139546091379.zip -> hdfs://localhost:9000/user/cpre419/.sparkStaging/
application_1618148656645_0002/__spark_libs__1251268139546091379.zip
2021-04-11 08:58:27,052 INFO yarn.Client: Uploading resource file:/home/cpre419/Downloads/github.jar -> hdfs://localhost:9000/user/cpre419/.sparkStaging/application_1618148656645_0002/github.jar
2021-04-11 08:58:27,598 INFO yarn.Client: Uploading resource file:/tmp/spark-09c21080-1491-4217-88f2-d95463761197/__spark_conf__2182990325597403l8.zip -> hdfs://localhost:9000/user/cpre419/.sparkStaging/a
pplication_1618148656645_0002/__spark_conf__.zip
2021-04-11 08:58:27,740 INFO spark.SecurityManager: Changing view acls to: cpre419
2021-04-11 08:58:27,741 INFO spark.SecurityManager: Changing modify acls to: cpre419
2021-04-11 08:58:27,742 INFO spark.SecurityManager: Changing view acls groups to:
2021-04-11 08:58:27,743 INFO spark.SecurityManager: Changing modify acls groups to:
2021-04-11 08:58:27,744 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(cpre419); groups with view permissions: Set(); users  with
 modify permissions: Set(cpre419); groups with modify permissions: Set()
2021-04-11 08:58:31,174 INFO yarn.Client: Submitting application application_1618148656645_0002 to ResourceManager
2021-04-11 08:58:31,290 INFO impl.YarnClientImpl: Submitted application application_1618148656645_0002
2021-04-11 08:58:32,300 INFO yarn.Client: Application report for application_1618148656645_0002 (state: ACCEPTED)
2021-04-11 08:58:32,306 INFO yarn.Client:
```

Output Screenshot: (from 2 files)



```
Java 462182 elasticsearch/elasticsearch 15698
Ruby 363801 rails/rails 29825
Python 331883 jakubroztocil/httpie 21283
PHP 273999 zurb/foundation 22636
C++ 159831 rogerwang/node-webkit 27350
C 145354 neovim/neovim 17395
C# 116155 dotnet/corefx 9176
```

2. Verified and attached the results of the data on a smaller dataset. The dataset and results are provided in the submissions folder. Will include results of running the experiment on patents dataset in the final submission.



```
cpre419@cpre419-VirtualBox:~/Downloads$ spark-submit --class NumOfTriangles triangle2.jar /data/patents.txt /op7
2021-04-11 08:45:46,915 WARN util.Utils: Your hostname, cpre419-VirtualBox resolves to a loopback address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
2021-04-11 08:45:46,916 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address
2021-04-11 08:45:49,030 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
2021-04-11 08:45:49,550 INFO yarn.Client: Requesting a new application from cluster with 1 NodeManagers
2021-04-11 08:45:49,745 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (8192 MB per container)
2021-04-11 08:45:49,745 INFO yarn.Client: Will allocate AM container, with 5632 MB memory including 512 MB overhead
2021-04-11 08:45:49,747 INFO yarn.Client: Setting up container launch context for our AM
2021-04-11 08:45:49,749 INFO yarn.Client: Setting up the launch environment for our AM container
2021-04-11 08:45:49,795 INFO yarn.Client: Preparing resources for our AM container
2021-04-11 08:45:50,068 WARN yarn.Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
2021-04-11 08:45:54,102 INFO yarn.Client: Uploading resource file:/tmp/spark-2da60134-7e69-4d2a-a7e4-fb9cd67228b5/__spark_libs__8767342431313587164.zip -> hdfs://localhost:9000/user/cpre419/.sparkStaging/
application_1618148656645_0001/__spark_libs__8767342431313587164.zip
2021-04-11 08:45:58,873 INFO yarn.Client: Uploading resource file:/home/cpre419/Downloads/triangle2.jar -> hdfs://localhost:9000/user/cpre419/.sparkStaging/application_1618148656645_0001/triangle2.jar
2021-04-11 08:45:58,910 WARN hdfs.DFSClient: Caught exception
```