Principles of Big Data Management

Phase 1 Report

Team :

Murali krishna sai Chukka (16272588)

Sushma Manne (16271112)

Pujitha Koppanati (16273485)

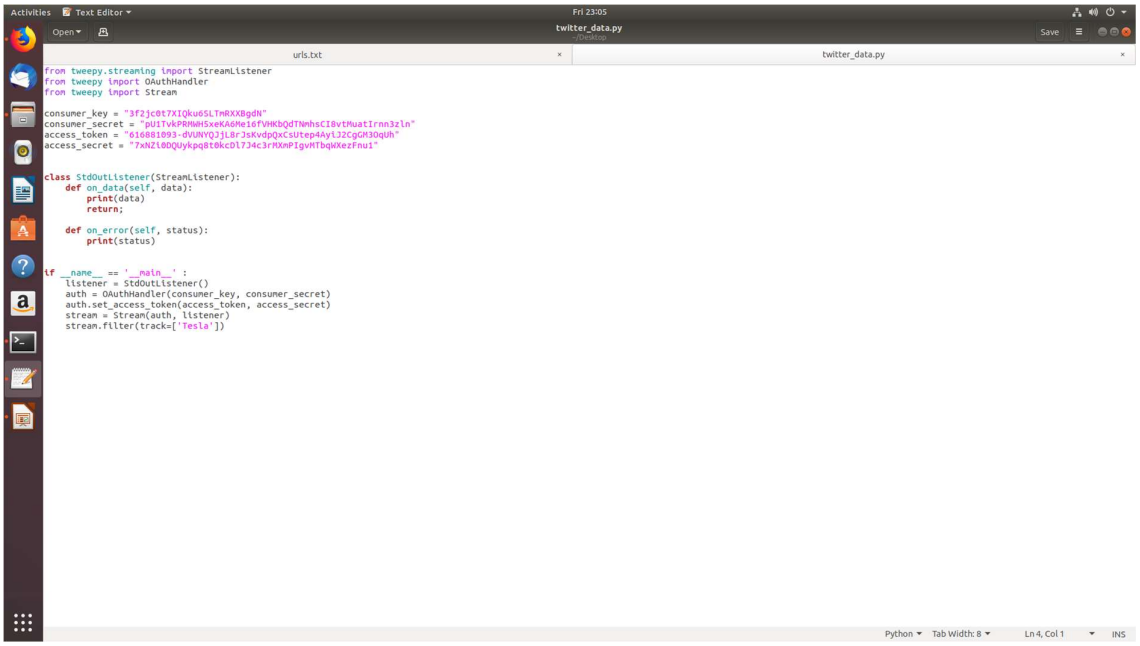
Links :

https://github.com/chkrish9/PB_Project/tree/master/

Task 1

Collect tweets using twitters streaming API(tweepy)

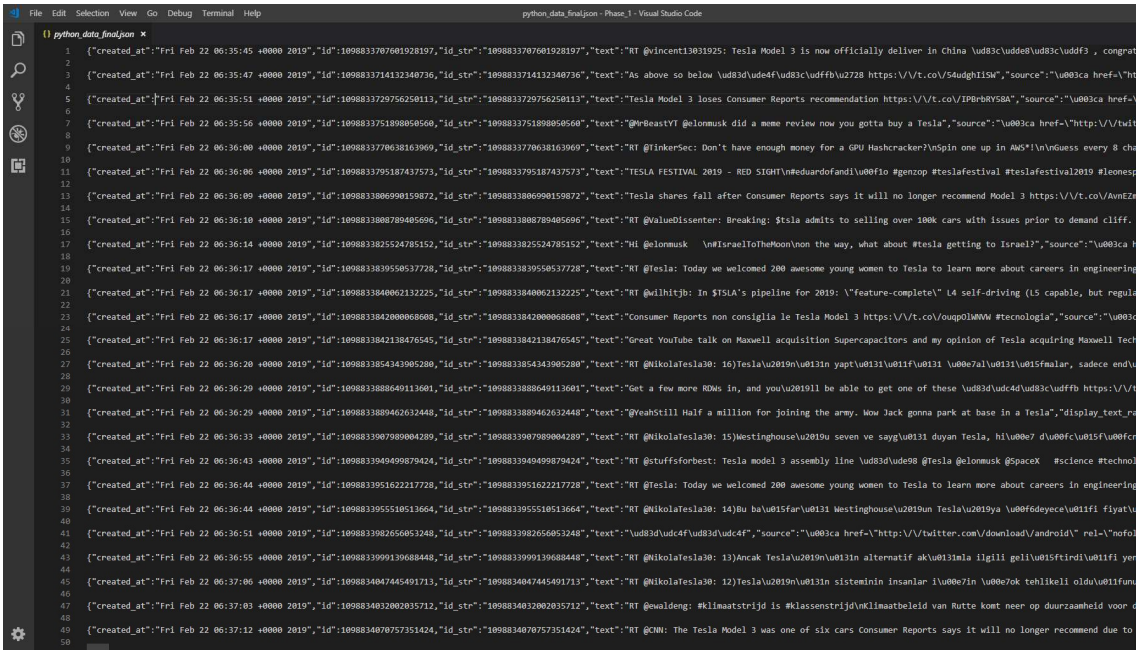
Python code for collecting tweets :



```
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream

consumer_key = "3f2jc0t7XIQku6SLTmRXXBgdN"
consumer_secret = "pU1TvkPRMWH5xeKA6Me16fVHKbQdTNwhscI8vtMuatIrnn3zln"
access_token = "616881093-dVUNYQJjL8rJsKvdpQxCsUtep4AylJ2CgGM3OqUh"
access_secret = "7xNZl0DQUykpq8t0kcDl7J4c3rMXmPIgvMTbqWXezFnu1"

class StdOutListener(StreamListener):
    def on_data(self, data):
        print(data)
        return;

    def on_error(self, status):
        print(status)

if __name__ == '__main__' :
    listener = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_secret)
    stream = Stream(auth, listener)
    stream.filter(track=['Tesla'])
```

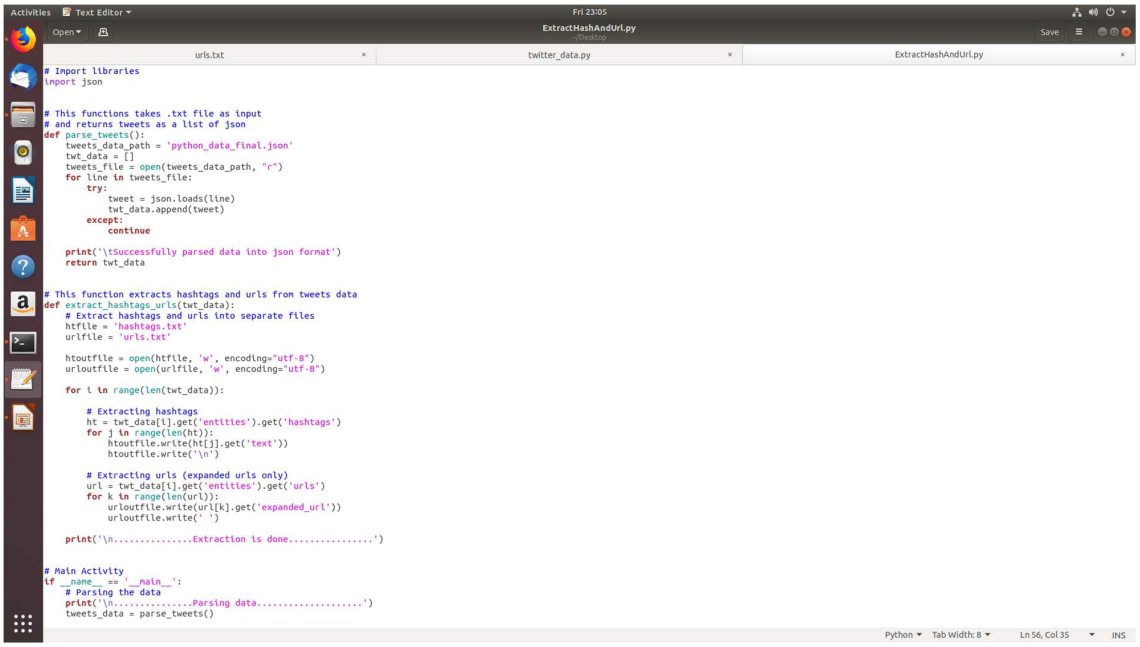
The above python code collects tweets into a text file by using Twitter's streaming API Tweepy. In order to connect to the Twitter's Streaming API, we need to authenticate using credentials from our twitter developer account. The program filters data by "Tesla" Keyword. The data is appended to file 'twitter_data.py'. We have collected around tweets.

Output file generated is of size 53MB



Task 2

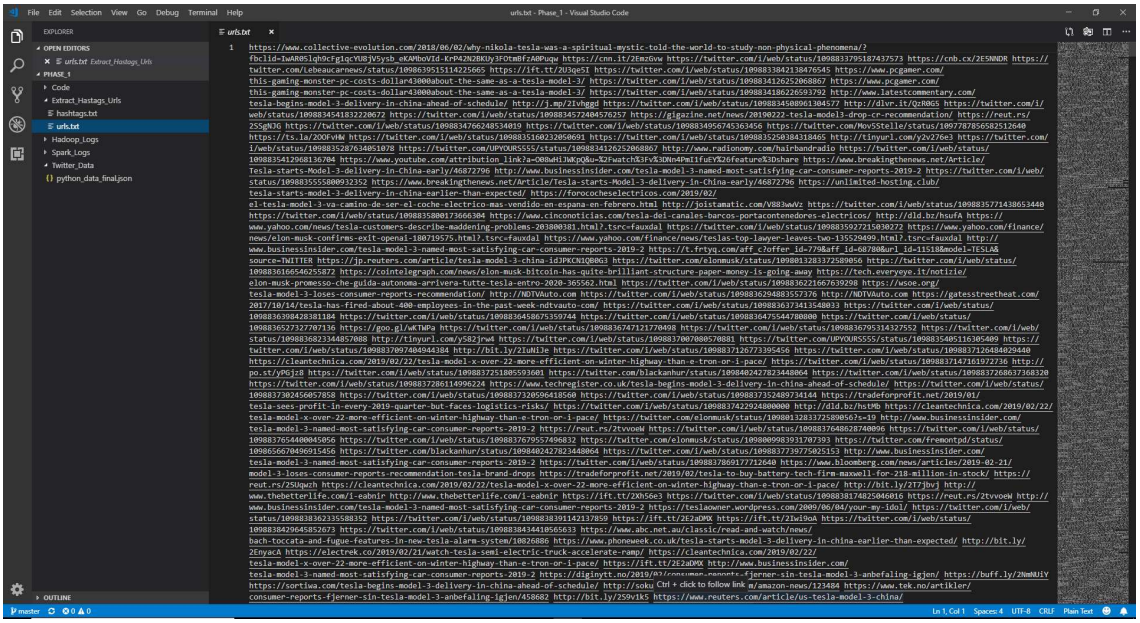
Hashtags and URLs extraction from the twitter data



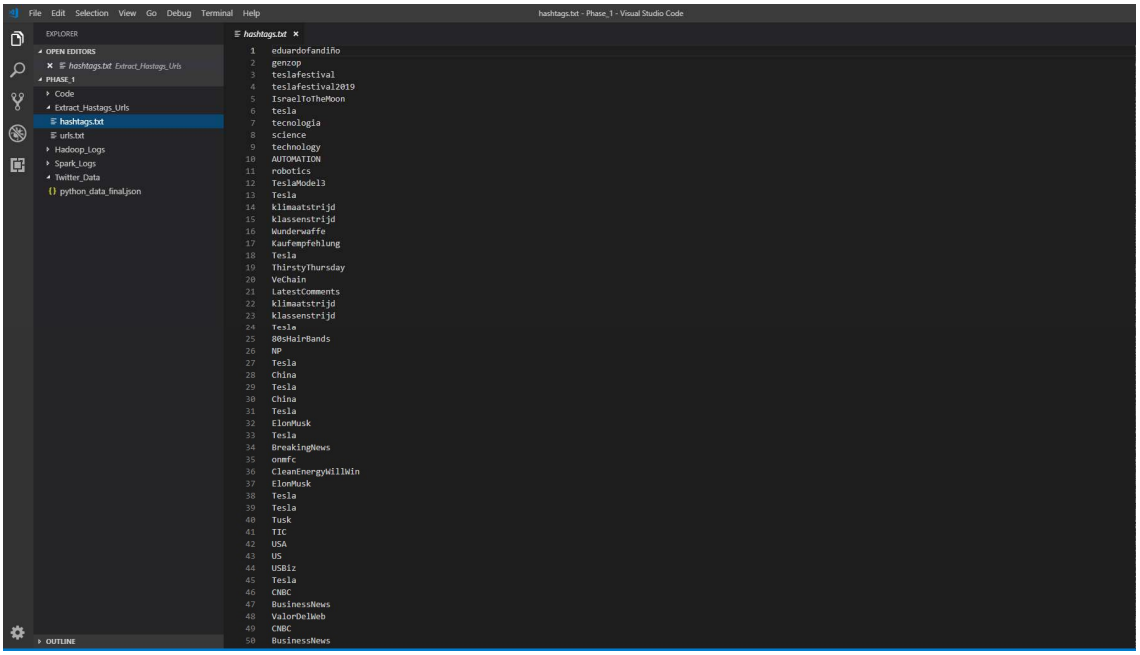
The above program parses tweets file into json format and extracts required Hashtags and urls and writes to two different files.

Outputfiles :

urlsoutput



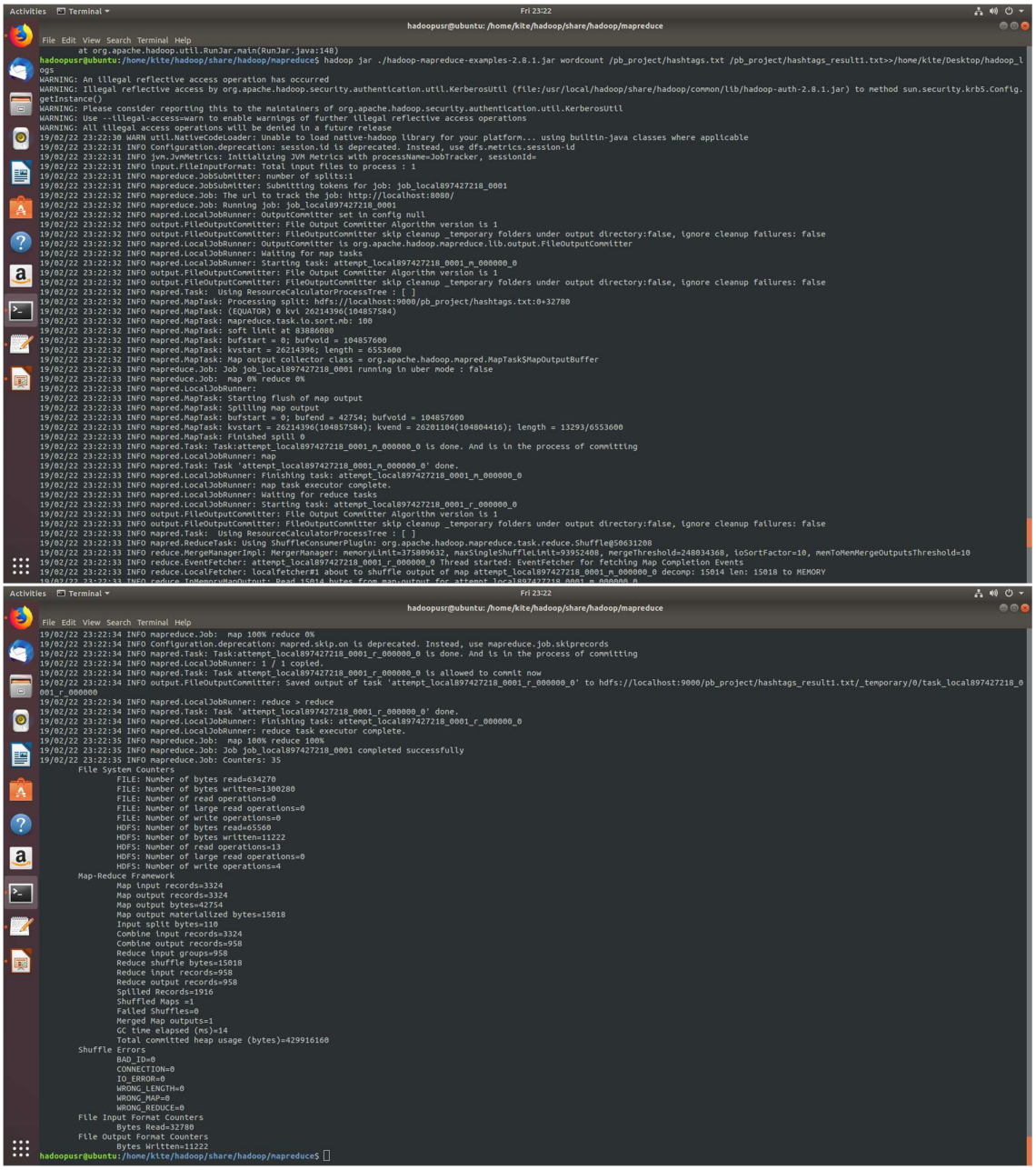
Hashtagsoutput

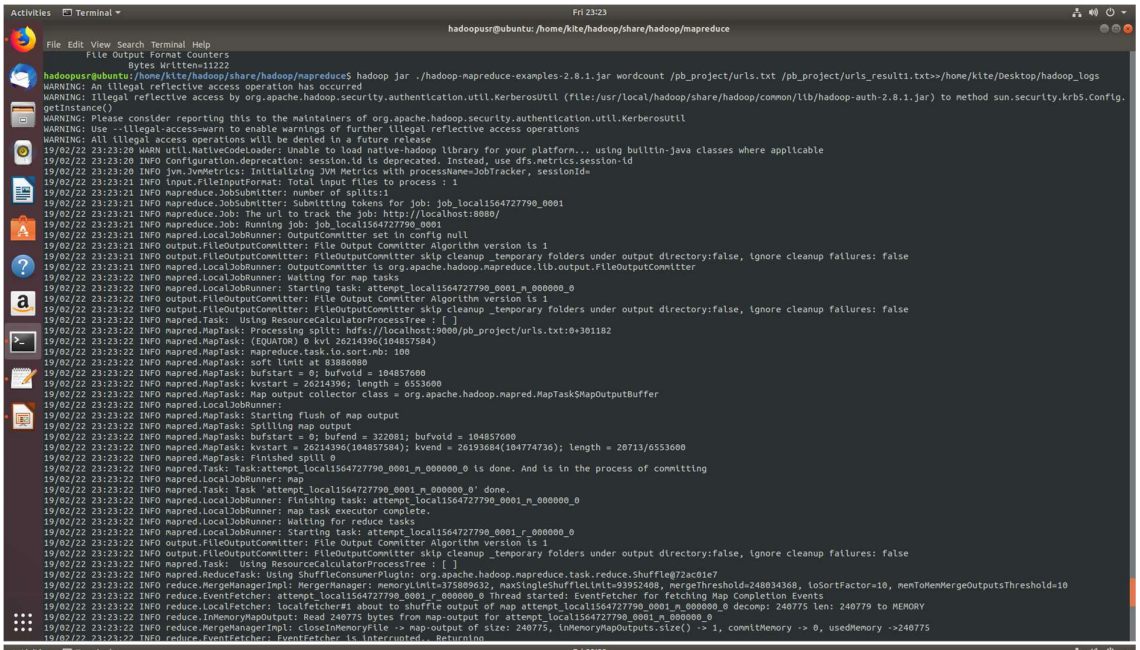


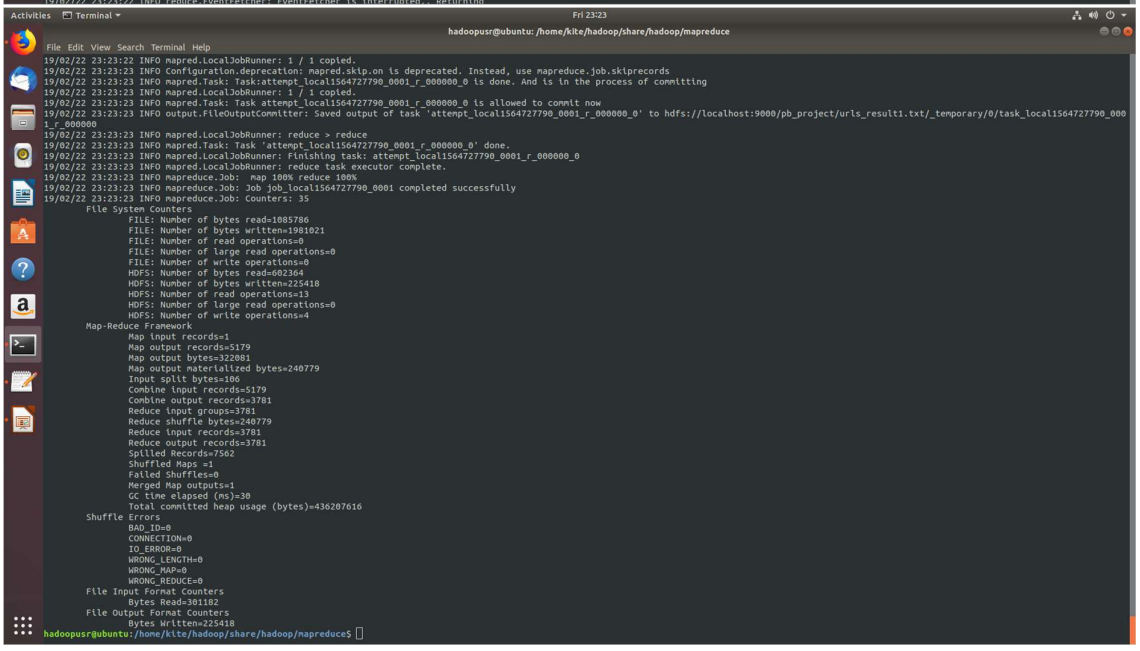
Task 3

WordCount Program Using Apache Hadoop.

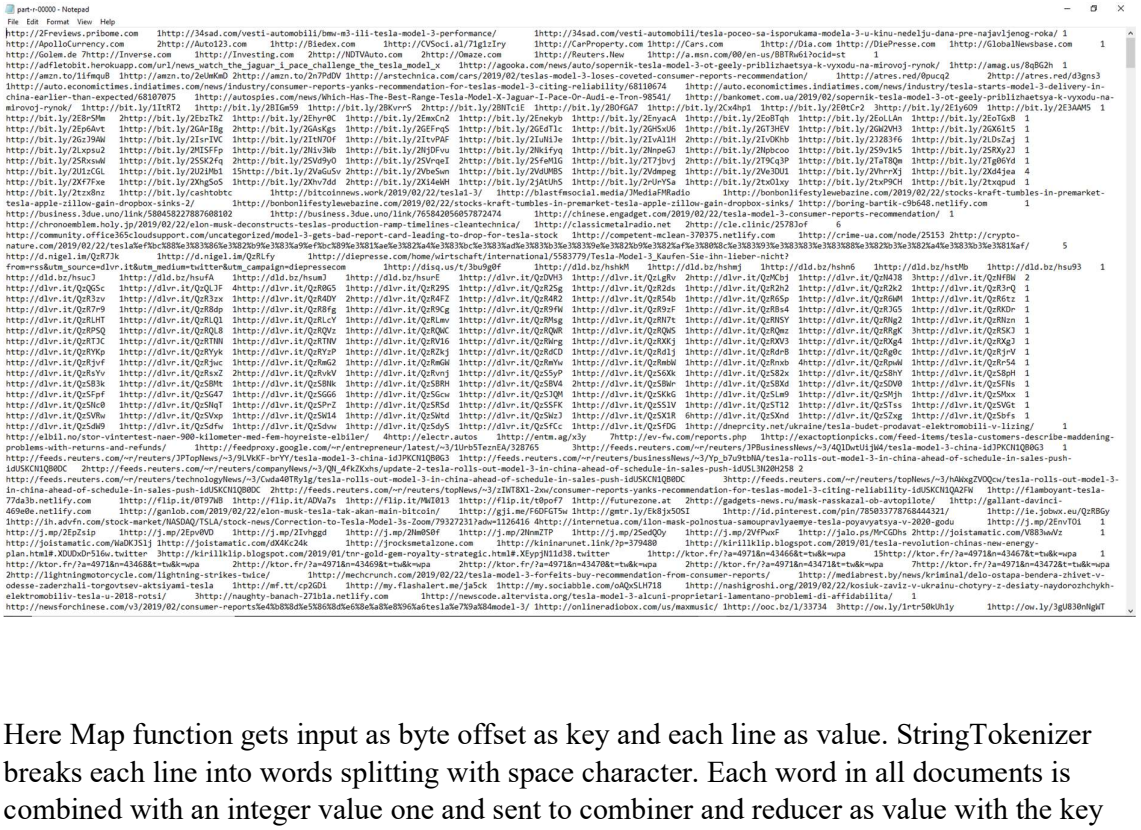
Log generated by running above program on both hashtags file and url file




Activities
Terminal
Fri 23:23
hadoopusr@ubuntu: /home/kite/hadoop/share/hadoop/mapreduce
File Edit View Search Terminal Help
File Output Format Counters
Bytes Written=11222
hadoopusr@ubuntu:/home/kite/hadoop/share/hadoop/mapreduce$ hadoop jar ./hadoop-mapreduce-examples-2.8.1.jar wordcount /pb_project/urls.txt /pb_project/urls_result1.txt>>/home/kite/Desktop/hadoop_logs
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.8.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
19/02/22 23:23:20 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/02/22 23:23:20 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
19/02/22 23:23:20 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
19/02/22 23:23:21 INFO input.FileInputFormat: Total input files to process : 1
19/02/22 23:23:21 INFO mapreduce.JobSubmitter: number of splits:1
19/02/22 23:23:21 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1564727790_0001
19/02/22 23:23:21 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
19/02/22 23:23:21 INFO mapreduce.Job: Running job: job_local1564727790_0001
19/02/22 23:23:21 INFO mapred.LocalJobRunner: OutputCommitter set in config null
19/02/22 23:23:21 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
19/02/22 23:23:21 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
19/02/22 23:23:21 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
19/02/22 23:23:22 INFO mapred.LocalJobRunner: Waiting for map tasks
19/02/22 23:23:22 INFO mapred.LocalJobRunner: Starting task: attempt_local1564727790_0001_m_000000_0
19/02/22 23:23:22 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
19/02/22 23:23:22 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
19/02/22 23:23:22 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
19/02/22 23:23:22 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/pb_project/urls.txt:0+301182
19/02/22 23:23:22 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
19/02/22 23:23:22 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
19/02/22 23:23:22 INFO mapred.MapTask: soft limit at 83886080
19/02/22 23:23:22 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
19/02/22 23:23:22 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
19/02/22 23:23:22 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
19/02/22 23:23:22 INFO mapred.LocalJobRunner:
19/02/22 23:23:22 INFO mapred.MapTask: Starting flush of map output
19/02/22 23:23:22 INFO mapred.MapTask: Spilling map output
19/02/22 23:23:22 INFO mapred.MapTask: bufstart = 0; bufend = 322081; bufvoid = 104857600
19/02/22 23:23:22 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26193684(104774736); length = 20713/6553600
19/02/22 23:23:22 INFO mapred.MapTask: Finished spill 0
19/02/22 23:23:22 INFO mapred.Task: Task:attempt_local1564727790_0001_m_000000_0 is done. And is in the process of committing
19/02/22 23:23:22 INFO mapred.LocalJobRunner: map
19/02/22 23:23:22 INFO mapred.Task: Task 'attempt_local1564727790_0001_m_000000_0' done.
19/02/22 23:23:22 INFO mapred.LocalJobRunner: Finishing task: attempt_local1564727790_0001_m_000000_0
19/02/22 23:23:22 INFO mapred.LocalJobRunner: map task executor complete.
19/02/22 23:23:22 INFO mapred.LocalJobRunner: Waiting for reduce tasks
19/02/22 23:23:22 INFO mapred.LocalJobRunner: Starting task: attempt_local1564727790_0001_r_000000_0
19/02/22 23:23:22 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
19/02/22 23:23:22 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
19/02/22 23:23:22 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
19/02/22 23:23:22 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@72ac01e7
19/02/22 23:23:22 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=375809632, maxSingleShuffleLimit=93952408, mergeThreshold=248034368, ioSortFactor=10, memToMemMergeOutputsThreshold=10
19/02/22 23:23:22 INFO reduce.EventFetcher: attempt_local1564727790_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
19/02/22 23:23:22 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local1564727790_0001_m_000000_0 decomp: 240775 len: 240779 to MEMORY
19/02/22 23:23:22 INFO reduce.InMemoryMapOutput: Read 240775 bytes from map-output for attempt_local1564727790_0001_m_000000_0
19/02/22 23:23:22 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 240775, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->240775
19/02/22 23:23:22 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning

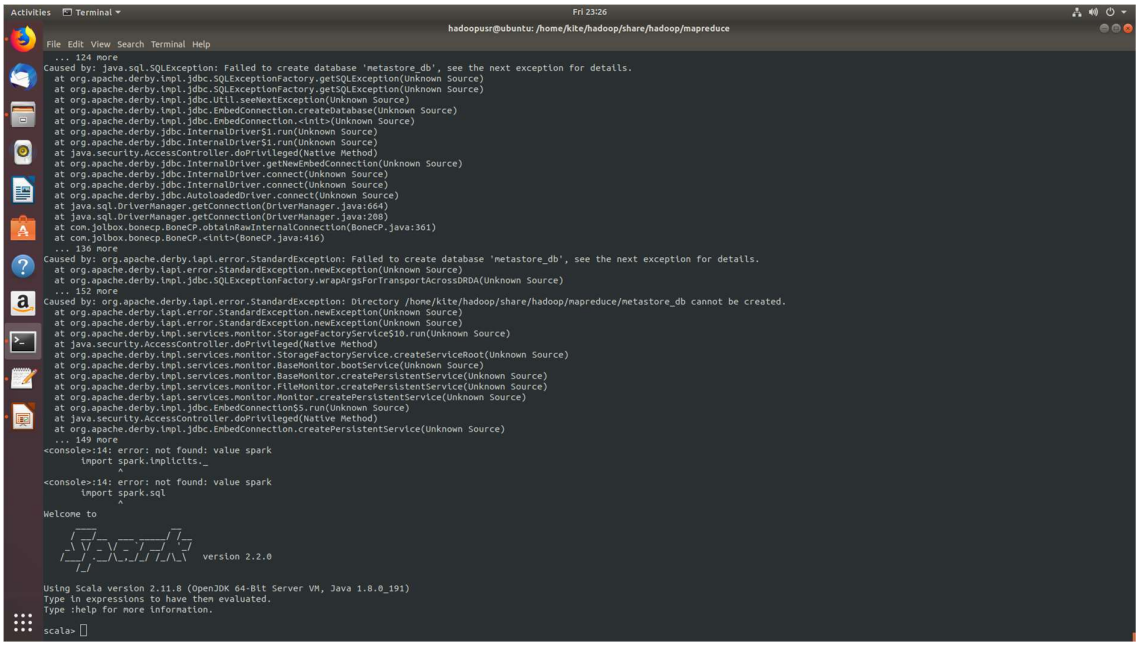

Activities
Terminal
Fri 23:23
hadoopusr@ubuntu: /home/kite/hadoop/share/hadoop/mapreduce
File Edit View Search Terminal Help
19/02/22 23:23:22 INFO mapred.LocalJobRunner: 1 / 1 copied.
19/02/22 23:23:23 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
19/02/22 23:23:23 INFO mapred.Task: Task:attempt_local1564727790_0001_r_000000_0 is done. And is in the process of committing
19/02/22 23:23:23 INFO mapred.LocalJobRunner: 1 / 1 copied.
19/02/22 23:23:23 INFO mapred.Task: Task attempt_local1564727790_0001_r_000000_0 is allowed to commit now
19/02/22 23:23:23 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1564727790_0001_r_000000_0' to hdfs://localhost:9000/pb_project/urls_result1.txt/_temporary/0/task_local1564727790_0001_r_000000
19/02/22 23:23:23 INFO mapred.LocalJobRunner: reduce > reduce
19/02/22 23:23:23 INFO mapred.Task: Task 'attempt_local1564727790_0001_r_000000_0' done.
19/02/22 23:23:23 INFO mapred.LocalJobRunner: Finishing task: attempt_local1564727790_0001_r_000000_0
19/02/22 23:23:23 INFO mapred.LocalJobRunner: reduce task executor complete.
19/02/22 23:23:23 INFO mapreduce.Job: map 100% reduce 100%
19/02/22 23:23:23 INFO mapreduce.Job: Job job_local1564727790_0001 completed successfully
19/02/22 23:23:23 INFO mapreduce.Job: Counters: 35
File System Counters
FILE: Number of bytes read=1085786
FILE: Number of bytes written=1981021
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=602364
HDFS: Number of bytes written=225418
HDFS: Number of read operations=13
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
Map-Reduce Framework
Map input records=1
Map output records=5179
Map output bytes=322081
Map output materialized bytes=240779
Input split bytes=106
Combine input records=5179
Combine output records=3781
Reduce input groups=3781
Reduce shuffle bytes=240779
Reduce input records=3781
Reduce output records=3781
Spilled Records=7562
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=30
Total committed heap usage (bytes)=436207616
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=301182
File Output Format Counters
Bytes Written=225418
hadoopusr@ubuntu:/home/kite/hadoop/share/hadoop/mapreduce$

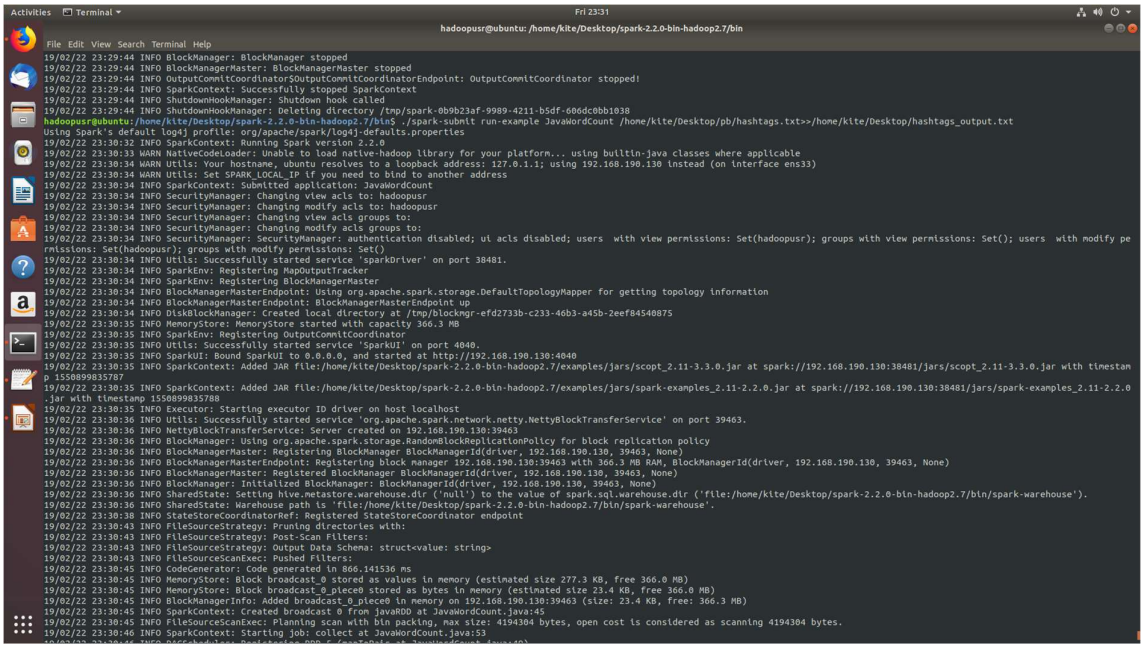
Output file showing hastags, urls and their counts

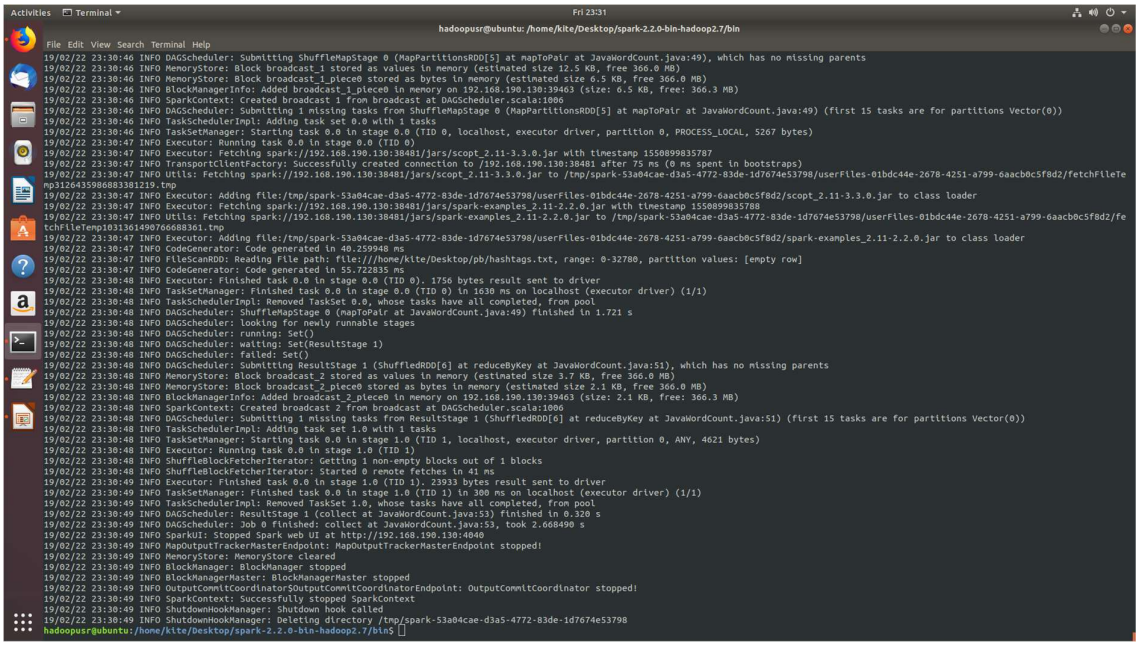


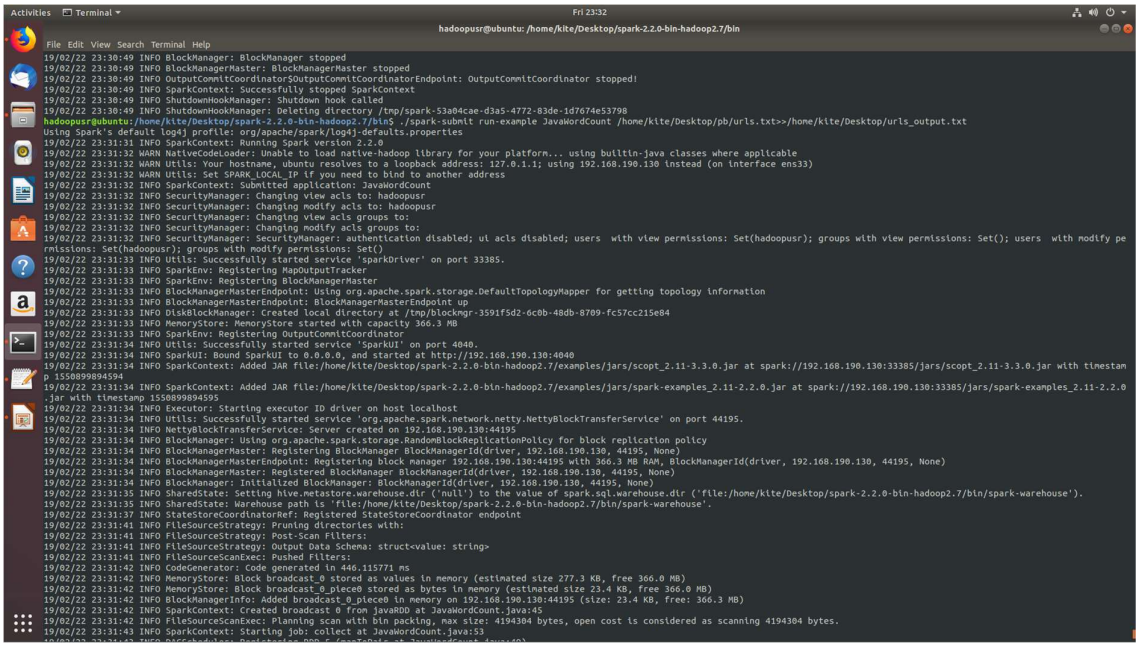
Here Map function gets input as byte offset as key and each line as value. StringTokenizer breaks each line into words splitting with space character. Each word in all documents is combined with an integer value one and sent to combiner and reducer as value with the key as word. Each Reducer takes Similar key values from the mapper, and counts all the values of unique key and send it as output value. The result from the reducer is word and the number of its occurrences from all the documents.

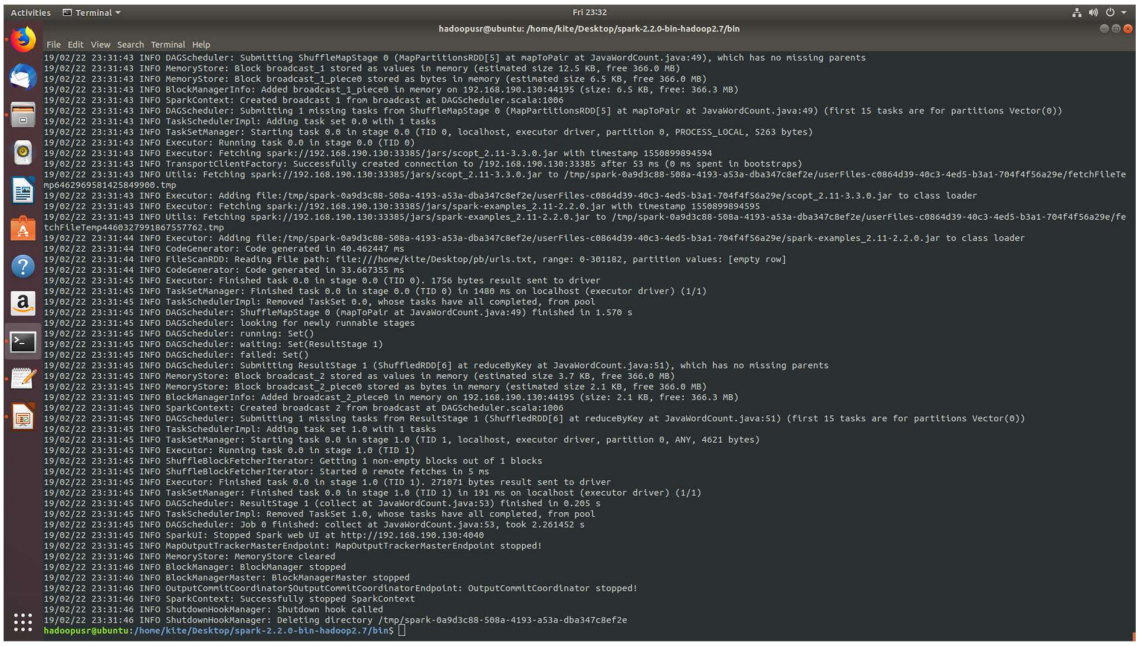
WordCount Program Using Apache Spark.



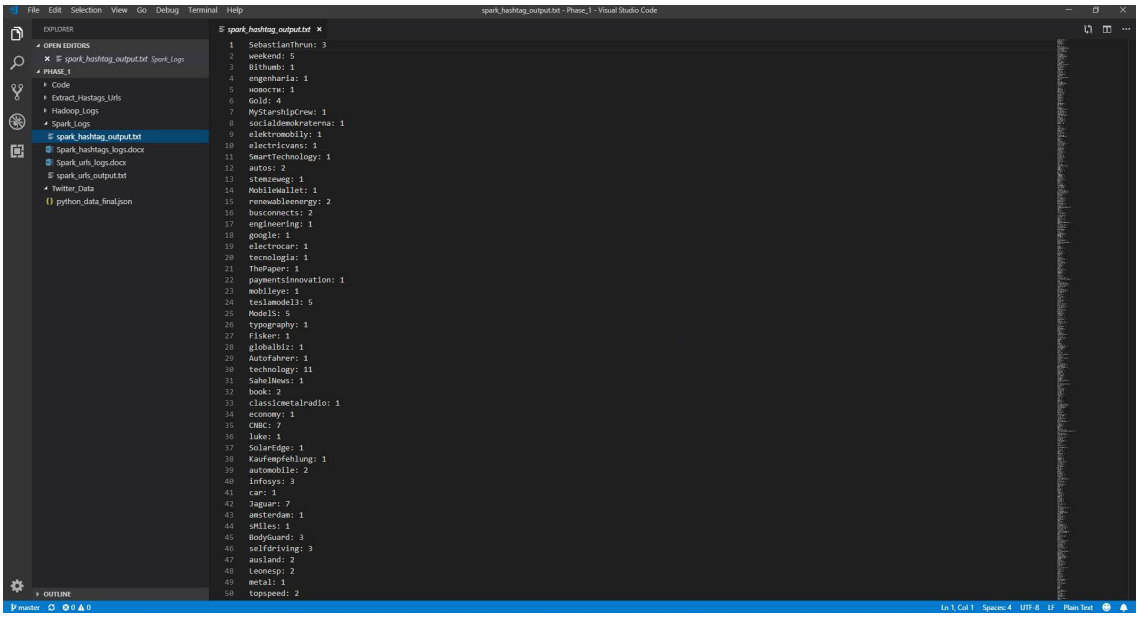


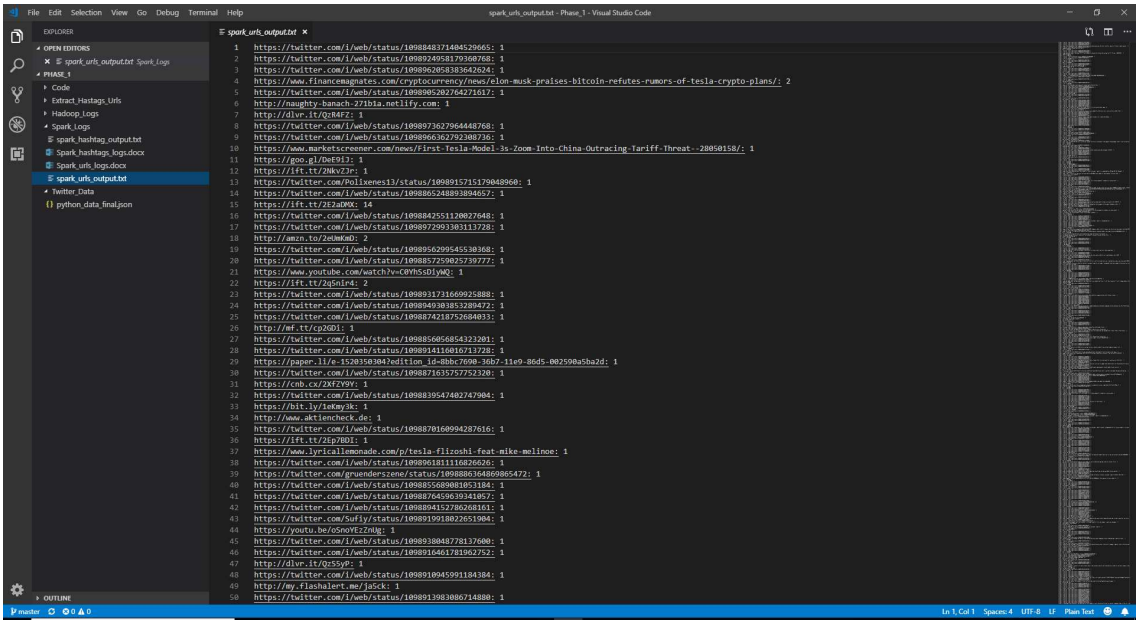

Activities
Terminal
Fri 23:31
hadoopusr@ubuntu: /home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/bin
File Edit View Search Terminal Help
19/02/22 23:30:46 INFO DAGScheduler: Submitting ShuffleMapStage 0 (MapPartitionsRDD[5] at mapToPair at JavaWordCount.java:49), which has no missing parents
19/02/22 23:30:46 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 12.5 KB, free 366.0 MB)
19/02/22 23:30:46 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 6.5 KB, free 366.0 MB)
19/02/22 23:30:46 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on 192.168.190.130:39463 (size: 6.5 KB, free: 366.3 MB)
19/02/22 23:30:46 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1006
19/02/22 23:30:46 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 0 (MapPartitionsRDD[5] at mapToPair at JavaWordCount.java:49) (first 15 tasks are for partitions Vector(0))
19/02/22 23:30:46 INFO TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
19/02/22 23:30:46 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, executor driver, partition 0, PROCESS_LOCAL, 5267 bytes)
19/02/22 23:30:47 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
19/02/22 23:30:47 INFO Executor: Fetching spark://192.168.190.130:38481/jars/scopt_2.11-3.3.0.jar with timestamp 1550899835787
19/02/22 23:30:47 INFO TransportClientFactory: Successfully created connection to /192.168.190.130:38481 after 75 ms (0 ms spent in bootstraps)
19/02/22 23:30:47 INFO Utils: Fetching spark://192.168.190.130:38481/jars/scopt_2.11-3.3.0.jar to /tmp/spark-53a04cae-d3a5-4772-83de-1d7674e53798/userFiles-01bdc44e-2678-4251-a799-6aacb0c5f8d2/fetchFileTemp3314359868838121219.tmp
19/02/22 23:30:47 INFO Executor: Adding file:/tmp/spark-53a04cae-d3a5-4772-83de-1d7674e53798/userFiles-01bdc44e-2678-4251-a799-6aacb0c5f8d2/scopt_2.11-3.3.0.jar to class loader
19/02/22 23:30:47 INFO Executor: Fetching spark://192.168.190.130:38481/jars/spark-examples_2.11-2.2.0.jar with timestamp 1550899835788
19/02/22 23:30:47 INFO Utils: Fetching spark://192.168.190.130:38481/jars/spark-examples_2.11-2.2.0.jar to /tmp/spark-53a04cae-d3a5-4772-83de-1d7674e53798/userFiles-01bdc44e-2678-4251-a799-6aacb0c5f8d2/fetchFileTemp1013614907668836l.tmp
19/02/22 23:30:47 INFO Executor: Adding file:/tmp/spark-53a04cae-d3a5-4772-83de-1d7674e53798/userFiles-01bdc44e-2678-4251-a799-6aacb0c5f8d2/spark-examples_2.11-2.2.0.jar to class loader
19/02/22 23:30:47 INFO CodeGenerator: Code generated in 40.259948 ms
19/02/22 23:30:47 INFO FileScanRDD: Reading File path: file:///home/kite/Desktop/pb/hashtags.txt, range: 0-32788, partition values: [empty row]
19/02/22 23:30:47 INFO CodeGenerator: Code generated in 55.722835 ms
19/02/22 23:30:48 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1756 bytes result sent to driver
19/02/22 23:30:48 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 1630 ms on localhost (executor driver) (1/1)
19/02/22 23:30:48 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
19/02/22 23:30:48 INFO DAGScheduler: ShuffleMapStage 0 (mapToPair at JavaWordCount.java:49) finished in 1.721 s
19/02/22 23:30:48 INFO DAGScheduler: looking for newly runnable stages
19/02/22 23:30:48 INFO DAGScheduler: running: Set()
19/02/22 23:30:48 INFO DAGScheduler: waiting: Set(ResultStage 1)
19/02/22 23:30:48 INFO DAGScheduler: failed: Set()
19/02/22 23:30:48 INFO DAGScheduler: Submitting ResultStage 1 (ShuffledRDD[6] at reduceByKey at JavaWordCount.java:51), which has no missing parents
19/02/22 23:30:48 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 3.7 KB, free 366.0 MB)
19/02/22 23:30:48 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 2.1 KB, free 366.0 MB)
19/02/22 23:30:48 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on 192.168.190.130:39463 (size: 2.1 KB, free: 366.3 MB)
19/02/22 23:30:48 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1006
19/02/22 23:30:48 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (ShuffledRDD[6] at reduceByKey at JavaWordCount.java:51) (first 15 tasks are for partitions Vector(0))
19/02/22 23:30:48 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
19/02/22 23:30:48 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, executor driver, partition 0, ANY, 4621 bytes)
19/02/22 23:30:48 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
19/02/22 23:30:48 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
19/02/22 23:30:48 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 41 ms
19/02/22 23:30:49 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 23933 bytes result sent to driver
19/02/22 23:30:49 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 309 ms on localhost (executor driver) (1/1)
19/02/22 23:30:49 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
19/02/22 23:30:49 INFO DAGScheduler: ResultStage 1 (collect at JavaWordCount.java:53) finished in 0.320 s
19/02/22 23:30:49 INFO DAGScheduler: Job 0 finished: collect at JavaWordCount.java:53, took 2.668490 s
19/02/22 23:30:49 INFO SparkUI: Stopped Spark web UI at http://192.168.190.130:4040
19/02/22 23:30:49 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
19/02/22 23:30:49 INFO MemoryStore: MemoryStore cleared
19/02/22 23:30:49 INFO BlockManager: BlockManager stopped
19/02/22 23:30:49 INFO BlockManagerMaster: BlockManagerMaster stopped
19/02/22 23:30:49 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
19/02/22 23:30:49 INFO SparkContext: Successfully stopped SparkContext
19/02/22 23:30:49 INFO ShutdownHookManager: Shutdown hook called
19/02/22 23:30:49 INFO ShutdownHookManager: Deleting directory /tmp/spark-53a04cae-d3a5-4772-83de-1d7674e53798
hadoopusr@ubuntu:/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/bin$


Activities
Terminal
Fri 23:32
hadoopusr@ubuntu: /home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/bin
File Edit View Search Terminal Help
19/02/22 23:30:49 INFO BlockManager: BlockManager stopped
19/02/22 23:30:49 INFO BlockManagerMaster: BlockManagerMaster stopped
19/02/22 23:30:49 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
19/02/22 23:30:49 INFO SparkContext: Successfully stopped SparkContext
19/02/22 23:30:49 INFO ShutdownHookManager: Shutdown hook called
19/02/22 23:30:49 INFO ShutdownHookManager: Deleting directory /tmp/spark-53a04cae-d3a5-4772-83de-1d7674e53798
hadoopusr@ubuntu:/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/bin$ ./spark-submit run-example JavaWordCount /home/kite/Desktop/pb/urls.txt>>/home/kite/Desktop/urls_output.txt
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
19/02/22 23:31:31 INFO SparkContext: Running Spark version 2.2.0
19/02/22 23:31:32 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/02/22 23:31:32 WARN Utils: Your hostname, ubuntu resolves to a loopback address: 127.0.1.1; using 192.168.190.130 instead (on interface ens33)
19/02/22 23:31:32 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
19/02/22 23:31:32 INFO SparkContext: Submitted application: JavaWordCount
19/02/22 23:31:32 INFO SecurityManager: Changing view acls to: hadoopusr
19/02/22 23:31:32 INFO SecurityManager: Changing modify acls to: hadoopusr
19/02/22 23:31:32 INFO SecurityManager: Changing view acls groups to:
19/02/22 23:31:32 INFO SecurityManager: Changing modify acls groups to:
19/02/22 23:31:32 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(hadoopusr); groups with view permissions: Set(); users with modify permissions: Set(hadoopusr); groups with modify permissions: Set()
19/02/22 23:31:33 INFO Utils: Successfully started service 'sparkDriver' on port 33385.
19/02/22 23:31:33 INFO SparkEnv: Registering MapOutputTracker
19/02/22 23:31:33 INFO SparkEnv: Registering BlockManagerMaster
19/02/22 23:31:33 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
19/02/22 23:31:33 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
19/02/22 23:31:33 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-3591f5d2-6c0b-48db-8709-fc57cc215e84
19/02/22 23:31:33 INFO MemoryStore: MemoryStore started with capacity 366.3 MB
19/02/22 23:31:33 INFO SparkEnv: Registering OutputCommitCoordinator
19/02/22 23:31:34 INFO Utils: Successfully started service 'SparkUI' on port 4040.
19/02/22 23:31:34 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.190.130:4040
19/02/22 23:31:34 INFO SparkContext: Added JAR file:/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/examples/jars/scopt_2.11-3.3.0.jar at spark://192.168.190.130:33385/jars/scopt_2.11-3.3.0.jar with timestamp 1550899894594
19/02/22 23:31:34 INFO SparkContext: Added JAR file:/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/examples/jars/spark-examples_2.11-2.2.0.jar at spark://192.168.190.130:33385/jars/spark-examples_2.11-2.2.0.jar with timestamp 1550899894595
19/02/22 23:31:34 INFO Executor: Starting executor ID driver on host localhost
19/02/22 23:31:34 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 44195.
19/02/22 23:31:34 INFO NettyBlockTransferService: Server created on 192.168.190.130:44195
19/02/22 23:31:34 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
19/02/22 23:31:34 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 192.168.190.130, 44195, None)
19/02/22 23:31:34 INFO BlockManagerMasterEndpoint: Registering block manager 192.168.190.130:44195 with 366.3 MB RAM, BlockManagerId(driver, 192.168.190.130, 44195, None)
19/02/22 23:31:34 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 192.168.190.130, 44195, None)
19/02/22 23:31:34 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.190.130, 44195, None)
19/02/22 23:31:35 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('file:/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/bin/spark-warehouse').
19/02/22 23:31:35 INFO SharedState: Warehouse path is 'file:/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/bin/spark-warehouse'.
19/02/22 23:31:37 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
19/02/22 23:31:41 INFO FileSourceStrategy: Pruning directories with:
19/02/22 23:31:41 INFO FileSourceStrategy: Post-Scan Filters:
19/02/22 23:31:41 INFO FileSourceStrategy: Output Data Schema: struct<value: string>
19/02/22 23:31:41 INFO FileSourceScanExec: Pushed Filters:
19/02/22 23:31:42 INFO CodeGenerator: Code generated in 446.115771 ms
19/02/22 23:31:42 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 277.3 KB, free 366.0 MB)
19/02/22 23:31:42 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 23.4 KB, free 366.0 MB)
19/02/22 23:31:42 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on 192.168.190.130:44195 (size: 23.4 KB, free: 366.3 MB)
19/02/22 23:31:42 INFO SparkContext: Created broadcast 0 from javaRDD at JavaWordCount.java:45
19/02/22 23:31:42 INFO FileSourceScanExec: Planning scan with bin packing, max size: 4194304 bytes, open cost is considered as scanning 4194304 bytes.
19/02/22 23:31:43 INFO SparkContext: Starting job: collect at JavaWordCount.java:53



In above spark code, first rdd is created by using textFile function of the sparkcontext from the hashtags and urls textfile. By using flatMap function, we generated words from each line by taking space as delimiter and appended one as value with the word as key. Using reduceByKey function we are adding values by the key and the output generated is the words and their sum is saved to a textfile.




File Edit Selection View Go Debug Terminal Help
spark_urls_output.txt - Phase_1 - Visual Studio Code
EXPLORER
OPEN EDITORS
spark_urls_output.txt Spark_Logs
PHASE_1
Code
Extract_Hastags_Urls
Hadoop_Logs
Spark_Logs
spark_hashtag_output.txt
Spark_hashtags_logs.docx
Spark_urls_logs.docx
spark_urls_output.txt
Twitter_Data
python_data_final.json
OUTLINE
https://twitter.com/i/web/status/1098848371404520665: 1
https://twitter.com/i/web/status/1098924958179360768: 1
https://twitter.com/i/web/status/1098962058383642624: 1
https://www.financemagnates.com/cryptocurrency/news/elon-musk-praises-bitcoin-refutes-rumors-of-tesla-crypto-plans/: 2
https://twitter.com/i/web/status/1098905202764271617: 1
http://naughty-banach-271bia.netlify.com: 1
http://dlvr.it/Qz84F2: 1
https://twitter.com/i/web/status/1098973627964448768: 1
https://twitter.com/i/web/status/1098966362792308736: 1
https://www.marketscreener.com/news/First-Tesla-Model-3s-Zoom-Into-China-Outracing-Tariff-Threat--28050158/: 1
https://goo.gl/DeF913: 1
https://ift.tt/2NKvZ7m: 1
https://twitter.com/Polixenes13/status/1098915715179048960: 1
https://twitter.com/i/web/status/1098865248893894657: 1
https://ift.tt/2E2aDMX: 14
https://twitter.com/i/web/status/1098842551120027648: 1
https://twitter.com/i/web/status/1098972993303113728: 1
http://amzn.to/2eUmKmD: 2
https://twitter.com/i/web/status/1098956299545530368: 1
https://twitter.com/i/web/status/1098857250025739777: 1
https://www.youtube.com/watch?v=C0YhSsDlyNQ: 1
https://ift.tt/2qSnlr4: 2
https://twitter.com/i/web/status/1098931731669925888: 1
https://twitter.com/i/web/status/1098849303853289472: 1
https://twitter.com/i/web/status/1098874218752684033: 1
http://mf.tt/cp2GDl: 1
https://twitter.com/i/web/status/1098856056854323201: 1
https://twitter.com/i/web/status/1098914116016713728: 1
https://paper.li/e-1520350304?edition_id=8bbc7690-36b7-11e9-86d5-002590a5ba2d: 1
https://twitter.com/i/web/status/1098871635757752320: 1
https://cnb.cx/2XfZY9Y: 1
https://twitter.com/i/web/status/1098830547402747904: 1
https://bit.ly/1eKmy3k: 1
http://www.aktiencheck.de: 1
https://twitter.com/i/web/status/1098870160994287616: 1
https://ift.tt/2Ep7BDI: 1
https://www.lyricallemonade.com/p/tesla-flizoshi-feat-mike-melinoe: 1
https://twitter.com/i/web/status/1098961811116826626: 1
https://twitter.com/gruenderszene/status/1098886364869865472: 1
https://twitter.com/i/web/status/1098855689081053184: 1
https://twitter.com/i/web/status/1098876459639341057: 1
https://twitter.com/i/web/status/1098894152786268161: 1
https://twitter.com/Sufiy/status/1098919918022651904: 1
https://youtu.be/oSnoYEzZnUg: 1
https://twitter.com/i/web/status/1098938048778137600: 1
https://twitter.com/i/web/status/1098916461781962752: 1
http://dlvr.it/Qz55y9: 1
https://twitter.com/i/web/status/1098910945991184384: 1
http://my.flashalert.me/ja5ck: 1
https://twitter.com/i/web/status/1098913983086714880: 1
master
Ln 1, Col 1
Spaces: 4
UTF-8
LF
Plain Text