```
from pyspark.sql import SparkSession
from graphframes import *
from pyspark.sql.functions import split, col
from pyspark.sql.types import IntegerType
import time
from datetime import datetime

appName = "Amazon0601_cluster_mode"

# Create Spark session
spark = SparkSession.builder \
    .appName(appName) \
    .getOrCreate()


sc = spark.sparkContext
print(sc)

# Start Time: "%H:%M:%S.%f"
# ========================
now = datetime.now()
start_time = now.strftime("%H:%M:%S.%f")
print("Start Time =", start_time)
# Start Time in seconds since the epoch as a floating point number.
start = time.time()
# =============
edges_df = spark.read.text('jba-datasets/amazon0601/amazon0601.txt')

edges_df=edges_df.filter(~col("value").startswith("#")).replace("#*", None)

edges_df = edges_df.withColumn("src", split("value",
"\t").getItem(0).cast(IntegerType())).withColumn("dst", split("value",
"\t").getItem(1).cast(IntegerType())).drop("value")

# edges_df.show()

# edges_df.printSchema()

# edges_df.count()

vertices_dst_df = edges_df.select("dst").withColumn("id",
col("dst")).drop("dst").distinct()

# vertices_dst_df.count()

vertices_src_df = edges_df.select("src").withColumn("id",
col("src")).drop("src").distinct()

# vertices_src_df.count()
```

```
46
47 vertices_df = vertices_dst_df.union(vertices_src_df).distinct()
48
49 # vertices_df.show(10)
50
51 # vertices_df.printSchema()
52
53 graph_df=GraphFrame(vertices_df,edges_df)
54
55
56 # maxIter: [3,4,5,8,10]
57 results = graph_df.pageRank(resetProbability=0.15, maxIter=5)
58
59 # End Time in seconds since the epoch as a floating point number.
60 # =============
61 end = time.time()
62 #Elapsed time in seconds as a floating point number.
63 # =============
64 print(f"Time elapsed for pageRank completion: {end - start:0.4f} seconds")
65 # End Time "%H:%M:%S.%f"
66 now = datetime.now()
67 end_time = now.strftime("%H:%M:%S.%f")
68 print("End Time =", end_time)
69 # =============
70 pagerank_results_df = results.vertices.sort("pagerank", ascending=False)
71
72 pagerank_results_df.show(20, False)
73
74 pagerank_results_df.coalesce(1) \
75       .write \
76       .option("header","true") \
77       .option("sep",",") \
78       .mode("overwrite") \
79       .csv("jba-datasets/amazon0601/pagerank_amazon0601.csv")
80
81 # pagerank_results_df= spark.read \
82 #       .option("header","true") \
83 #       .option("sep",",") \
84 #       .option("inferSchema", "true") \
85 #       .csv("jba-datasets/amazon0601/pagerank_amazon0601.csv")
86
87 # pagerank_results_df.printSchema()
88
89 # pagerank_results_df=pagerank_results_df.sort("pagerank", ascending=False)
90 # pagerank_results_df.show(20, False)
91
92 print("Bye")
93 sc.stop()
```