Raj Kamal Yadav
2016076

Rajat Bansal
2016260

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Programming Assignment-4, BDA

1. Analyze the Epinions Social Network dataset.
2. Run PageRank on data for most trusted user.
3. Run HITS algorithm on data for most trusted user.
4. Run Simrank(Personalized PageRank) for most trusted users of the top Pagerank user.

Reading Data from Txt

```
import org.apache.spark._
import org.apache.spark.graphx._
import org.apache.spark.rdd.RDD
import java.util.Calendar
import org.apache.spark.sql.{DataFrame, Row, SQLContext}
import org.apache.spark.sql.functions._
import org.apache.spark.{SparkConf, SparkContext}
import org.apache.spark.graphx.{Graph, VertexRDD, Edge => GXEdge}
import org.apache.spark.sql.types.{IntegerType, LongType}
import org.apache.spark.graphx.GraphLoader

val graph = GraphLoader.edgeListFile(sc, "/home/ajkamal/Desktop/zzz/Rbans/soc-Epinions1.txt")
```

1. **GraphX is a Scala Library that is used for iterative graph computation within a single system and data visualization as both graphs and collections.**

2. **The format of the .txt file is kept as provided at the time downloads without any alteration.**
   # Directed Epinions social network
   # Nodes: 75879 Edges: 508837
   # FromNodeId ToNodeId

Terminology and Methodology

- **Reset Probability:** Damping parameter for PageRank; 1-alpha.
- **Convergence Rate:** Error tolerance used to check convergence in power method solver.
- **Authority Score:** Estimates the value of the content of the page.
- **Hub Score:** Estimates the value of its links to other pages.
- **Personalized Pagerank:** Personalization vector" consisting of a dictionary with a key for every graph node and nonzero personalization value for each node also known as SimRank.

OUTPUT SCORES OF PAGERANK ALGORITHM ON THE EPINIONS DATASET

With Convergence = 0.0001

Rank	User-Id	PageRank Score of the User(Un-Normalized)
1	18	325.77
2	737	212.54
3	1719	147.164
4	118	144.99
5	790	142.29

OUTPUT SCORES OF HITS ALGORITHM ON THE EPINIONS DATASET based on AUTHORITY SCORE with 25 Iterations.

Rank	User-Id	HITS Score of the user (Normalized)
1	**645**	**0.003396**
2	**634**	**0.00244**
3	**34**	**0.001734**
4	**763**	**0.00161**
5	**44**	**0.00157**

OUTPUT SCORES OF HITS ALGORITHM ON THE EPINIONS DATASET based on HUB UPDATE SCORE with 25 Iterations.

Rank	User-Id	HITS Score of the user (Normalized)
1	**18**	**0.261**
2	**118**	**0.0081**
3	**790**	**0.0073**
4	**136**	**0.0070**
5	**1191**	**0.0069**

OUTPUT SCORES OF THE SIMRANK ALGORITHM ON USER-ID “18” (Highest PageRank)

Rank	User-Id	Simrank Score with User “18”
1	18	0.261
2	118	0.0081
3	790	0.0073
4	136	0.0070
5	1191	0.00698
6	128	0.00692

Learnings from the Assignment

- We learnt about links and their significances.
- Learnt the pros and cons of using different twitter API's for streaming tweets and rate limits.
- Learnt about GraphX library, its setup, usage(Graph+Collection) and ease of access to iterate and infer Graph edges and vertices.
- Hands on experience of PageRank Algorithm, Personalized PageRank Algorithm(SimRank) and HITS Algorithm.
- Learnt the significance and difference b/w Authority and Hub Scores in HITS algorithm.

Challenges faced in solving the assignment

- Version mismatch. Often specific versions of Scala run specific versions of GraphX or if you are running Spark, that also runs specific versions of GraphX, Scala and the streaming api's(Apache Spark/GraphX/Graphframe) making it difficult to integrate.

- We are relatively beginners to Scala as a Programming Language, hence had to look over lots of resources for syntax for even very benign statements/tasks.

- Really scarce detailed well explained documentation/blogs online to refer from for GraphX Library. Also GraphX only integrates with Scala.

- Normalization in HITS algorithm to join and update old graph with new attributes.

Resources Used

1) For PageRank

https://www.infoq.com/articles/apache-spark-graphx/

2) For HITS

https://medium.com/@gangareddy619/advanced-graph-algorithms-in-spark-using-graphx-aggregated-messages-and-collective-communication-f3396c7be4aa

3) For SimRank

https://livebook.manning.com/book/spark-graphx-in-action/chapter-5/50