Big Data Analytics Assignment - 4

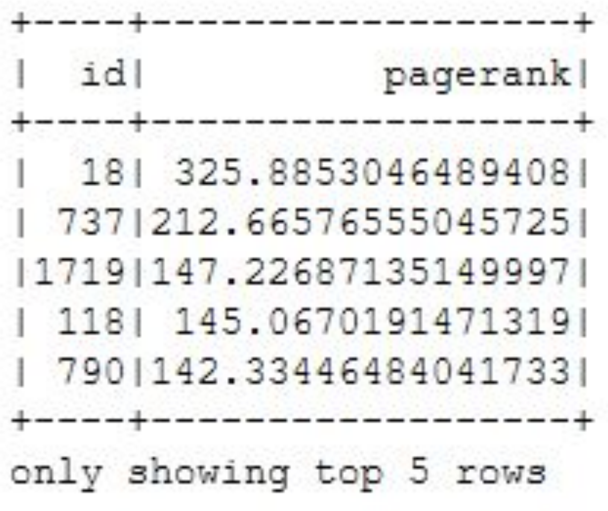
Abhishek Agarwal - 2016126
Kaustav Vats - 2016048

Learnings

- Learnt how to apply PageRank algorithm
- Various applications of PageRank
- Shortcomings of PageRank
- HITS algorithm and its applications
- Learned about usage of graphFrames library and various api calls for graph like problems. Learned how to create graph and run various algorithms on the graph.
- In this assignment, we learned how we can implement the HITS algorithm using Map-Reduce APIs present in the apache storm.
- Most importantly we learned about, how various graph problems can be solved using libraries like GraphFrame, Networkx, etc.

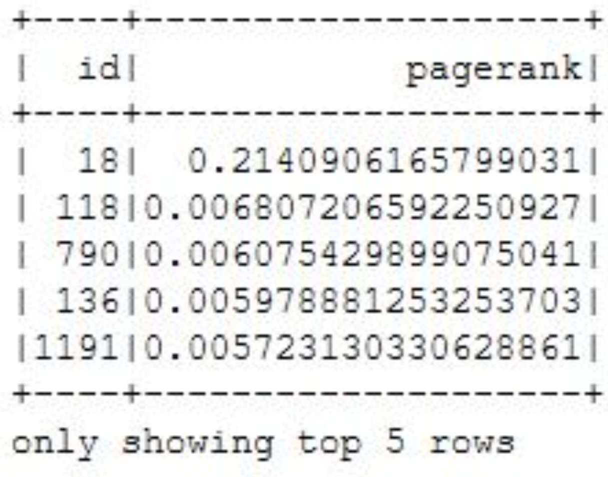
Results - 1

PageRank Results



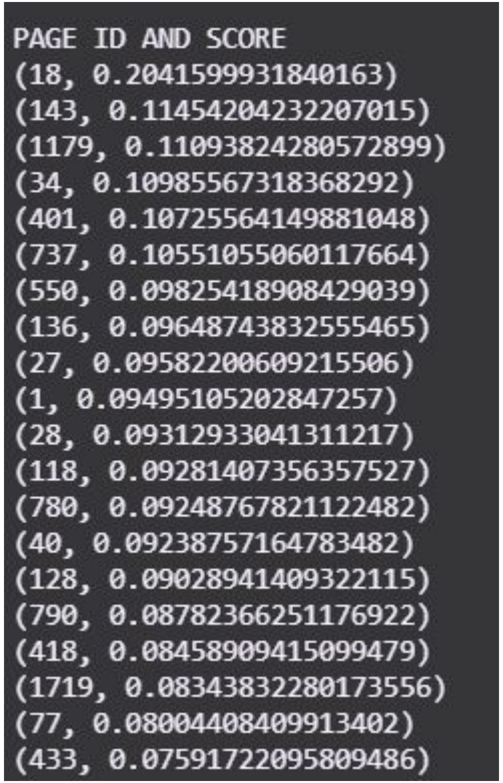
```
+----+------------------+
|  id|          pagerank|
+----+------------------+
|  18| 325.8853046489408|
| 737|212.66576555045725|
|1719|147.22687135149997|
| 118| 145.0670191471319|
| 790|142.33446484041733|
+----+------------------+
only showing top 5 rows
```

SimRank Results



```
+----+--------------------+
|  id|            pagerank|
+----+--------------------+
|  18|  0.2140906165799031|
| 118|0.006807206592250927|
| 790|0.006075429899075041|
| 136|0.005978881253253703|
|1191|0.005723130330628861|
+----+--------------------+
only showing top 5 rows
```

Results - 2 (HITS algorithm results)

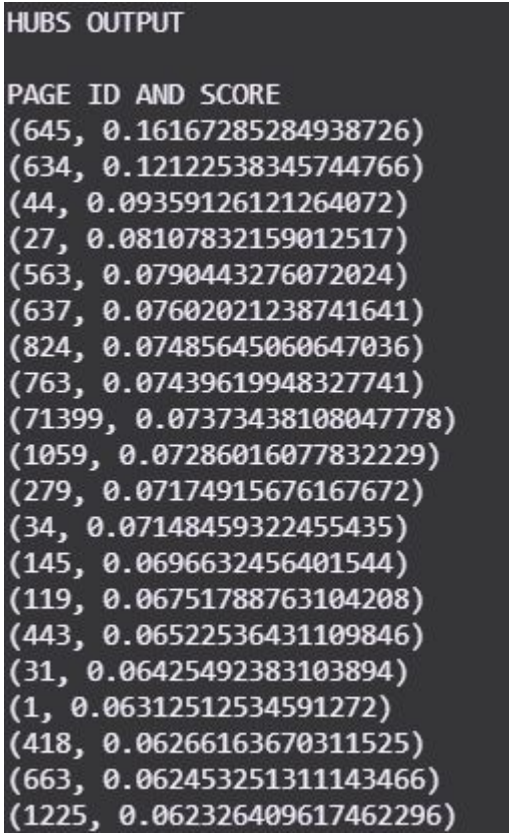


```
PAGE ID AND SCORE
(18, 0.2041599931840163)
(143, 0.11454204232207015)
(1179, 0.11093824280572899)
(34, 0.10985567318368292)
(401, 0.10725564149881048)
(737, 0.10551055060117664)
(550, 0.09825418908429039)
(136, 0.09648743832555465)
(27, 0.09582200609215506)
(1, 0.09495105202847257)
(28, 0.09312933041311217)
(118, 0.09281407356357527)
(780, 0.09248767821122482)
(40, 0.09238757164783482)
(128, 0.09028941409322115)
(790, 0.08782366251176922)
(418, 0.08458909415099479)
(1719, 0.08343832280173556)
(77, 0.08004408409913402)
(433, 0.07591722095809486)
```

Left
Top 20 result of HITS algorithm with ID and Authority Score in decreasing order.

Right
Top 20 result of HITS algorithm with ID and Hubs Score in decreasing order.

Top 5 Most trusted users are - 18, 143, 1179, 34, 401.



```
HUBS OUTPUT

PAGE ID AND SCORE
(645, 0.16167285284938726)
(634, 0.12122538345744766)
(44, 0.09359126121264072)
(27, 0.08107832159012517)
(563, 0.0790443276072024)
(637, 0.07602021238741641)
(824, 0.07485645060647036)
(763, 0.07439619948327741)
(71399, 0.07373438108047778)
(1059, 0.07286016077832229)
(279, 0.07174915676167672)
(34, 0.07148459322455435)
(145, 0.0696632456401544)
(119, 0.06751788763104208)
(443, 0.06522536431109846)
(31, 0.06425492383103894)
(1, 0.06312512534591272)
(418, 0.06266163670311525)
(663, 0.062453251311143466)
(1225, 0.062326409617462296)
```

Challenges

- Took some time to implement and understand all the resources mentioned online. Implementation of HITS algorithm using Map-Reduce.

References

- https://graphframes.github.io/graphframes/docs/_site/api/python/graphframes.lib.html
- HITS Algorithm Reference - https://raw.githubusercontent.com/jaimeps/hits-algorithm/master/spark/hits_spark.py