Submitted by – Yash Vijaynarayan Gupta

Big Data Management & Analytics (CS6350)

PageRank for Airports

PageRank algorithm can be used to evaluate relative importance of nodes in a connected environment. It is based on the concept of in-links and out-links of a node and is used to rank nodes in a graph in order of their importance. Details about this algorithm and its implementation using MapReduce can be found in Chapter 5 of the reference book Data Intensive Text Processing using MapReduce. You can also look at the slides available at http://lintool.github.io/UMD-courses/bigdata-2015-Spring/slides/session05.pdf.

Note that you cannot use any external library that automatically computes PageRank, including Spark GraphX. The dataset for this project will be a graph that shows connections between various airports. This data is available at: Bureau of Transportation website. You would need to do the following to download the data:

1. Go to https://transtats.bts.gov/

2. On the left menu, click under "Aviation" under the "By Mode" block.

3. On the next page, click Air Carrier Statistics (Form 41 Traffic)- U.S. Carriers

4. On the next page, click download link under T-100 Domestic Segment (U.S. Carriers)

5. On the next page, set Filter Year = Most Recent Year, Filter Period = Any month on which data is available (e.g. July), and select following fields:

- Origin (Origin Airport Code)
- OriginCityName (optional)
- Dst (Destination Airport Code)
- DstCityName (optional)

6. Download and unzip to get a csv file that will used as input file.

You will use the following equation to compute PageRank:

$$\text{PR}(x) = \alpha \times \frac{1}{N} + (1-\alpha) \times \sum_{1}^{n} \frac{PR(t_i)}{C(t_i)}$$

where $\alpha = 0.15$ and x is a page with inlinks from $t_1, t_2, \ldots, t_n$, $C(t)$ is the out-degree of t, and N is the total number of nodes in the graph.

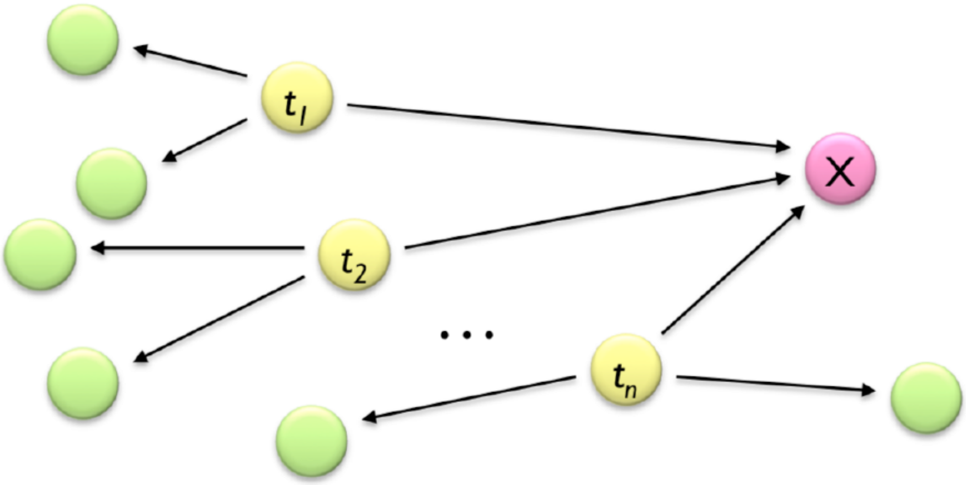

Figure 1: PageRank computation

Compute the page rank of each node (airport) based on number of inlinks and outlinks. There may be multiple connections between two airports, and you should consider them independent of each other to compute the number of inlinks and outlinks. For example, if node A is connected to node B with an out-count of 10 and node C with an out-count of 10, then the total number of outlinks for node A would be 20. Initialize all the PageRank values to be 10.0. The output should contain the airport code and its PageRank, and data should be sorted by the PageRank in a descending order.