Çankaya University
Department of Computer Engineering

CENG 595 Distributed Data Processing and Analysis
Spring 2017

Homework 2

Due by Apr 4 (Tuesday) 6pm

Subject: Spark and graph problem

1) Download the following graph data graph.tsv from course web page.
   Each line includes a triple of the form <*source_url destination_url weight*> (tab separated).
2) For each url compute the out degree (number of outgoing edges) and output (node, count) pairs in sorted order by node (**outdegree**.scala or .py program).
3) For each url compute the sum of weights of incoming edges and output (node, weight_sum) pairs in sorted order by node (**weight**.scala or .py program).
4) For each node X, find a list of all other nodes Y such that there is an (X,Y) edge in the graph and a (Y,X) edge in the graph, and output (X, [Y1, Y2, …, Yn]) pairs in order sorted by X (**pairs**.scala or .py program).

   Develop the above graph processing algorithms in Apache Spark using Scala or Python languages.

   Submit your programs and output files as a single zipped file with your name (**asg2_lastname_firstname.zip**) via webonline. Output files should be named the same as the program, for example **outdegree.txt**.

Resources:

- Developing Spark programs in Eclipse: http://freecontent.manning.com/wp-content/uploads/how-to-start-developing-spark-applications-in-eclipse.pdf
- Spark download and quick start: http://spark.apache.org/docs/latest/index.html
- Spark programming guide: http://spark.apache.org/docs/latest/programming-guide.html
- Spark SQL: http://spark.apache.org/docs/latest/sql-programming-guide.html
- Mastering Apache Spark 2 Book: https://jaceklaskowski.gitbooks.io/mastering-apache-spark/
  - https://github.com/jaceklaskowski/mastering-apache-spark-book