CENG 686 Distributed Data Processing and Analysis
Erdogan Dogdu

Assignment 3

Subject: Spark SQL

In this assignment you will use a retweet network data and answer the following question:

1. Download retweet network data here (zipped 29MB): data

   Data file contains retweet counts between two users in each line in the following format:

   ```
   user1,user2,retweetCount
   ```

2. Develop a **Spark SQL** program in Jupyter notebook format to compute the following **queries** on the retweet network file:

   a. How many unique users are in the dataset?
   b. How many total retweets happened in the dataset?
   c. Find the top 50 users who are retweeted (user2) the most in decreasing order.
   d. Find the top 50 users who retweeted (user1) the most in decreasing order.
   e. Find the top 50 users who are retweeted (user2) by different users the most in decreasing order.
   f. Find the top 50 users who are retweeting (user1) different users the most in decreasing order.
   g. Find the common users between (c) and (e), and then between (d) and (f). Report how many users in each list and what is the percentage of common users between these lists?

3. For each query, report **the processing time** on your notebook after each query/program piece.
4. Submit your notebook from Google Colab (share the link) by sending an email to the instructor with the subject "**CENG686** Assignment 3".

Resources:

- Spark SQL Programming Guide (2.4)
- Spark SQL Example on CoLab