ID2221: Data-Intensive Computing
Lab4 – GraphX Hands on

In this lab you will first practice some basic analytic tasks using GraphX. Next you will implement iterative PageRank algorithm.

1. **Work Environment:**
   The good news is you already have the required frameworks installed if you have done lab1 and lab2. Otherwise, follow the same instructions to install Docker, Spark, and Jupyter.

2. **Datasets:**
   The assignment folder already contains a wiki-Vote dataset that is downloaded from Stanford data repository (available via https://snap.stanford.edu/data/index.html). Wiki-Vote dataset is directed network that is not too big as I'm not sure of the computational capacity of your computers. Yet, the repository contains much larger datasets in case you are going to need some for your master thesis.

3. **Lab Guide:**
   Unzip lab4.zip and copy its contents into the mywork folder you created before for lab2. Then the files should be available in jupyter. You should have **two notebooks** and a **data folder**:
   1. Graphx_basics
   2. Iterative_PageRank

   The notebooks are self-explanatory, such that with each part you can find description of used functions.

4. **Final Report:**

In your final report copy only your solution code (should NOT be much) and explain with a few words how the values in each required RDD (i.e., transitions and ranks) are updated. Also, include in the report the nodeIDs of the top 10 highly ranked nodes. Please use filename format "Lab4_yourName.pdf".