**Big Data (A, I)**

**Midterm Exam**

**Wednesday, April 6, 2022**

**Frequent Itemset/Pattern Mining**

The dataset consists of two files, “movies.csv,” which contains the movie lists (movieId, title), and “watchlist.csv,” which has (userId, movieId) to describe that the user has already seen the movie.

* Based on that dataset, your job is to create an FP-growth recommendation model. Create step-by-step big data analytics process using CRISP-DM methodology. Explain each step you do as clearly as possible.
* During the modeling step, try the hyperparameter tuning (minSupport and minConfidence) using several combination values. Then decide what the best value for minSupport and minConfidence is.
* During the deployment step, please use new data (several userIds and movieIds) and use your model to predict the recommendation for the users. You must display the original movie title in the recommendation result instead of movieId.

References:

* <https://colab.research.google.com/drive/1JTJ6vHlYhQy77_X_gtwBpTHl5W3vlebg?usp=sharing>
* <https://spark.apache.org/docs/latest/ml-frequent-pattern-mining.html>

**Link Analysis, Page Rank**

The dataset, “us\_edgelist.csv,” consists of the relationships between US presidents or candidates and their spouses. Meanwhile, the “us\_age.csv” contains their respective age. You need to provide the IDs for the data.

Based on that data, you are required to do graph analysis to answer these questions:

* Find the top 3 most influential persons based on the PageRank result.
* Find the person who has: a) highest in degree values, b) highest out-degree values
* Find the shortest path from “Ivanka” to “G. Bush.”

References:

* <https://docs.databricks.com/spark/latest/graph-analysis/graphframes/user-guide-python.html>
* <https://colab.research.google.com/drive/15pv0ovKvZyE9ovoDP0hXweJI9A9Zn_YG?usp=sharing>
* <https://graphframes.github.io/graphframes/docs/_site/user-guide.html#shortest-paths>

**Assignment Submission**

* Write your notes and code on a Jupyter Notebook.

Please explain each step as clearly as possible.

Create a separate notebook for each task (one for frequent itemset, one for graph analysis)

* Download both of your Jupyter Notebooks in .ipynb format and compress them using .zip format.
* Upload your zipped file.