Algorithmic Data Analysis (ADA)
Spring 2024
Coding Assignment #2: Mining temporal data – Frequent sequences
Initial submission on 31.01.24 at 23:59, then revisions weekly on Wednesdays

UNIVERSITY OF
EASTERN FINLAND

Please carefully read and follow the general instructions regarding coding assignments. Failing to meet the requirements might lead to penalties. `https://elearn.uef.fi/mod/page/view.php?id=248672`

If you suspect that something is wrong with some task instructions, please contact the lecturer.

If you face persistent issues while working on a task, do ask for help, e.g. during a course meeting or by contacting the lecturer via email.

Dataset

– *CRSW* originally from `https://en.ilmatieteenlaitos.fi/download-observations#!/`, `2019-01-03_itemsets-CRSW.txt` contains two months worth of weather data in Kuopio (January–February 2019). Each line represents weather events during one hour as an itemset. The first column contains the contextual attribute, i.e. indication of time in `Year-Month-Day_Hour` format. The second column contains the itemset. `C`, `R` `S` and `W` stand respectively for clouds, precipitation, sunshine and wind.

Task 1. Implement the generalized sequential pattern mining (GSP) algorithm for mining frequent patterns from discrete sequences.

Apply the algorithm on the *CRSW* dataset.

Try different values for the mininum support threshold, different constraints such as max gap and max span, and considering the data either as one long sequence or with each day as a separate short sequence. You might also try replacing repeated occurrences of the same itemset by a single copy, i.e. represent constant weather during successive hours with only one itemset.

Report on the type and number of patterns obtained under different conditions.