MLSD: Assignment 3
Streams

– Due date: June 5, 2024 –

What to submit

For each exercise, submit a documented Jupyter notebook or a python script to run through spark-submit, and the results of the algorithm. If the results are too large, submit a download link instead.

The comments should explain the main steps of the solution with sufficient detail.

1. Stream processing

   In these exercises you should make use of Spark Structured Streaming: Structured Streaming Programming Guide

   You may use the code provided ('binary_multi_stream_generator.py') as a starting point to generate the data to be consumed by the streaming engine.

   2.1 Implement the DGIM method to estimate the number of 1s within a window of size $k \leq N$, where k and N are parameters. Test it by generating various synthetic bit streams and estimate the number of 1s in each stream at pre-defined intervals. You should also show the correct number of 1s, which is known.

   2.2 Using the streaming dataset provided, apply the exponentially decaying window approach to keep smoothed counts of occurring events. Inspect the stream at one second intervals (to check which events occurred), and display the 5 most frequent events at pre-defined intervals.

2. TBD