Exercise 2

Window Models

1. What is the main difference between a Sliding and Landmark windows?

2. Consider the scenario of Exercise 1, Task 2. We still consider the reviews as a stream with the timepoint being the 1st of each month. Let the window size be 3 months. Calculate
   a. For timepoint t_i, the number of reviews in the window
   b. For timepoint t_i, the number of positive/negative reviews in the window
   c. For timepoint t_i, the number of users in the window

   Use a Sliding window and Landmark window for the above calculations.

3. In your implementation have you maintained any sufficient statistics? If yes, can you define them and any operators that you have used? If no, are there any disadvantages to your implementation?

4. For a damped window model, consider the fading function $f(t) = 2^{-\lambda t}$ where t is the timepoint and λ is a user-defined parameter. What is the weight of an instance x observed at timepoint T ($T > t$)? Calculate the weight of the instance x at t_0, t_1, t_2, t_3, t_4 since t_0. Plot a graph of weight vs timepoint.

CluStream Algorithm

5. With $q = 5$, determine the initial micro-clusters for the following set of initial points.

T	X	y
1	6	2
2	7	3
3	6.5	1
4	1	1
5	2	2
6	3	1
7	3	2.5
8	2	8
9	2	6
10	2.5	7
11	4	7

   Use the following points for the initial cluster centers: (1, 1), (2.5, 2), (2, 7), (4, 7), (6, 2)

6. Perform the online step of the CluStream Algorithm for the following stream of points: (2, 7), (2.5, 3), (3.5, 7), (7, 8), (6, 7), (2.5, 2), (5, 5).

T	X	y
12	2	7
13	2.5	3
14	3.5	7
15	7	8
16	6	7
17	2.5	2
18	5	5

The maximum boundary of the micro-cluster is defined as follows:
For micro-cluster with more than 1 point: It is 2 times the micro-cluster radius
For a single point micro-cluster: distance to the closest micro-cluster
To calculate relevancy of the micro-cluster use a time window of 10 timepoints.

7. Perform the offline step to form 3 clusters. (HINT: No need to perform Step 1 of the of the Offline step)