Assignment 8-1  \textit{CluStream}  

Given the following series of data points.

<table>
<thead>
<tr>
<th>Time t</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data point p</td>
<td>(1 1)</td>
<td>(2 1)</td>
<td>(4 9)</td>
<td>(4 8)</td>
<td>(10 4)</td>
<td>(9 3)</td>
<td>(2 3)</td>
<td>(11 3)</td>
<td>(12 12)</td>
<td>(12 11)</td>
<td>(11 12)</td>
<td>(4 2)</td>
</tr>
</tbody>
</table>

Perform the online steps of the CluStream algorithm on the data point series with the following settings:

- \textit{initPoints} = 6
- \( q = 3 \)
- factor of clu radius \( t = 5 \)

Assignment 8-2  \textit{Finding similar items}  

Suppose that the universal set is given by \{1, \ldots, 10\}. Construct minhash signatures for the following sets:

(a) \( S_1 = \{3, 6, 9\} \)

(b) \( S_2 = \{2, 4, 6, 8\} \)

(c) \( S_3 = \{2, 3, 4\} \)

1. Construct the signatures for the sets using the following list of permutations:
   
   - (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
   - (10, 8, 6, 4, 2, 9, 7, 5, 3, 1)
   - (4, 7, 2, 9, 1, 5, 3, 10, 6, 8)

2. Suppose that instead of using particular permutations to construct signatures for the threse sets, we use hash functions. The three hash functions we use are:

   - \( h_1(x) = x \mod 10 \)
   - \( h_2(x) = (2x + 1) \mod 10 \)
   - \( h_3(x) = (3x + 2) \mod 10 \)

3. How does the estimated Jaccard similarity, derived from (1.) and (2.) compare with the true Jaccard similarity of the original data? How to reduce deviations in the approximated Jaccard similarities?