IR (Chapter 19) Classwork Problem

Dell Zhang
Birkbeck, University of London

1. Suppose that we need to perform near-duplicate detection in a collection of three documents. Their sets of shingle fingerprints are as follows.
   $D_1 : \{0, 1, 2\}$
   $D_2 : \{1, 3, 4\}$
   $D_3 : \{0, 2, 3\}$
   Please estimate their pairwise Jaccard coefficients using MinHash with the following two hash functions.
   $h_1(x) = (2x + 1) \mod 5$
   $h_2(x) = (3x + 1) \mod 5$