Question 2. LSH.

ci) document A ("the sky is blue the sun is bright")

B ("the sun in the sky is bright")

a) compute 2-shingles for A and B

b) compute their Jaccard similarity based on their 2-shingles.

Answer: a) $k=2$, set of 2-shingles

2-singles for A: {(the, sky), (sky, is), (is, blue), (blue, the), (the, sun), (sun, is), (is, bright)}

2-singles for B: {(the, sun), (sun, in), (in, the), (the, sky), (sky, is), (is, bright)}

the Jaccard similarity is $sim(A,B) = |A \cap B| / |A \cup B|$

$= 4/8$

(ii) $h_1(n) = 5n - 1 \mod M$

$h_2(n) = 2n + 1 \mod M$