1. **Exercise 3.1.1** : Compute the Jaccard similarities of each pair of the following three sets: S1 = 1, 2, 3, 4, S2 = 2, 3, 5, 7, and S3 = 2, 4, 6.

   $sim(C1, C2) = |C1 \cap C2|/|C1 \cup C2|$

   sim(S1, S2) = 2/6 = 1/3
   sim(S1, S3) = 2/5
   sim(S2, S3) = 1/6

2. **Exercise 3.2.1** : What are the first ten 3-shingles in the first sentence of Sec- tion 3.2? "The most effective way to represent documents as sets, for the purpose of iden- tifying lexically similar documents is to construct from the document the set of short strings that appear within it."

   1. "The"
   2. "he "
   3. ë m"
   4. "mo"
   5. "mos"
   6. öst"
   7. "st "
   8. "t e"
   9. ëf"
   10. ëff"

3. **Exercise 3.3.3** : In Fig. 3.5 is a matrix with six rows. (a) Compute the minhash signature for each column if we use the following three hash functions: h1(x) = 2x + 1 mod 6; h2(x) = 3x + 2 mod 6; h3(x)=5x+2 mod6.

Elements	S1	S2	S3	S4	2x+1 mod 6	3x+2 mod 6	5x+2 mod 6
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

	S1	S2	S3	S4
h1(0)	/	1	/	1
h2(0)	/	2	/	2
h3(0)	/	2	/	2
h1(1)	/	1	/	1
h2(1)	/	2	/	2
h3(1)	/	2	/	2
h1(2)	5	1	/	1
h2(2)	2	2	/	2
h3(2)	0	1	/	0
h1(3)	5	1	1	1
h2(3)	2	2	5	2
h3(3)	0	1	5	0
h1(4)	5	1	1	1
h2(4)	2	2	2	2
h3(4)	0	1	4	0
h1(5)	5	1	1	1
h2(5)	2	2	2	2
h3(5)	0	1	4	0

Bereken voor iedere cel de verschillende functies en zet in een tabel. De finale matrix is:

S1	S2	S3	S4
5	1	1	1
2	2	2	2
0	1	4	0

(b) Which of these hash functions are true permutations?

h3 is een permutatie. h1 kan geen permutatie zijn omdat deze als eerste waarde 5 heeft terwijl voor S1 er 2 keer een 1 voorkomt dus er zal altijd een rij eerder een 1 tegenkomen. h2 is ook geen permutatie omdat geen enkele rij 4 keer een 1 heeft. 3 heeft als permutatie 2, 1, 0, 4, 3, 5.

(c) How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?

similarities	1-2	1-3	1-4	2-3	2-4	3-4
col/col	0	0	0.25	0	0.25	0.25
sig/sig	0.33	0.33	0.67	0.67	0.67	0.67

De waarden liggen helemaal niet in elkaars buurt.

4. **Exercise 3.3.5** : Prove that if the Jaccard similarity of two columns is 0, then minhashing always gives a correct estimate of the Jaccard similarity.

5. **Exercise 3.4.1** : Evaluate the S-curve 1 - (1 - sr)b for s = 0.1, 0.2, . . . , 0.9, for the following values of r and b:

   zie oplossingen is gewoon uitrekenen en plotten

6. **Exercise 3.4.2** : For each of the (r,b) pairs in Exercise 3.4.1, compute the threshold, that is, the value of s for which the value of $1 - (1 - s^r)^b$ is exactly 1/2. How does this value compare with the estimate of $(1/b)^{1/r}$ that was suggested in Section 3.4.2?

   De functie invullen met r en b en vervolgens gelijkstellen aan 0.5, dit geeft ons:

   - s = 0.406088
   - s = 0.569353
   - s = 0.424394

   voor $(1/b)^{1/r}$ krijgen we:

   - s = 0.464158883
   - s = 0.606962
   - s = 0.4573050