Bielefeld University
Faculty of Technology

Genome Data Science
Johannes Schlüter
Maren Knop

Big Data Analytics

Summer 2023

Number 01, Submission Deadline: April 17 2023, 11:59 PM

1. **Shingles:** Consider two documents: *ABRACADABRA* and *BRICABRAC* (8 P)
   (a) How many 2-shingles does *ABRACADABRA* have? List the shingles.
   (b) How many 2-shingles does *BRICABRAC* have? List the shingles.
   (c) How many 2-shingles do they have in common? List the shingles.
   (d) What is the Jaccard similarity between the two documents? Please note single steps.

2. **Minhashing:**
   (a) Perform a minhashing of the following matrix: (4 P)

Row	C_1	C_2	C_3	C_4
0	0	1	1	0
1	1	0	1	1
2	0	1	0	1
3	0	0	1	0
4	1	0	1	0
5	0	1	0	0

   using the following row permutation vectors $(4, 5, 0, 2, 3, 1)^T$ and $(3, 1, 0, 5, 2, 4)^T$. [1]

   (b) Consider the following matrix: (12 P)

[1] In the task you are suppose to use the row permutation vectors (4,5,0,2,3,1)T and (3,1,0,5,2,4)T, which are transposed. This means you have to assign row 4 to row 0, row 5 to row 1. $4 \to 0$, $5 \to 1$, $0 \to 2$, $2 \to 3$, $3 \to 4$, $1 \to 5$, etc.

Row	S_1	S_2	S_3	S_4
0	0	1	0	1
1	0	0	0	1
2	1	1	0	0
3	0	0	1	0
4	0	1	1	0
5	1	0	0	0
6	1	0	1	0

a) Compute the minhash signature for each column using the following three hash functions:
$h_1(x) = 2x + 1 \ mod \ 7$
$h_2(x) = 3x + 2 \ mod \ 7$
$h_3(x) = 5x + 2 \ mod \ 7$

b) Describe how to check if these hash functions are permutations?

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

d) What is the benefit of using hash functions instead of permutations?

3. **LSH:** Consider the following signature matrix:

C_1	C_2	C_3	C_4	C_5	C_6	C_7
1	2	1	1	2	5	4
2	3	4	2	3	2	2
3	1	2	3	1	3	2
4	1	3	1	2	4	4
5	2	5	1	1	5	1
6	1	6	4	1	1	4

Perform locality-sensitive hashing as introduced in the lecture with three bands of two rows each. Assume we have enough buckets available so that the hash function for each band can be the identity function (i.e. columns hash to the same bucket if and only if they are identical in the band).

a) Find all candidate pairs and give a textual explanation how you identified the candidates. (4 P)

4. **LS Families:** What is the effect of probability of stating with the family of minhash functions and applying: (8 P)

1) A 2-way AND construction followed by a 3-way OR construction.

2) A 3-way OR construction followed by a 2-way AND construction.

3) A 2-way AND construction followed by a 2-way OR construction, followed by a 2-way AND construction.

4) A 2-way OR construction followed by a 2-way AND construction, followed by a 2-way OR construction, followed by a 2-way AND construction.

Show how the probability can be calculated and how the s-curves could look like for the 4 cases.

Important:
Please submit your group solution via Moodle. You are free to hand solutions in as PDFs or Jupyter Notebooks.