

# Exercises on Block2:

- Finding Frequent Item Sets
- Finding Similar Items
- Searching in Data Streams

Advanced Search Techniques for Large Scale Data Analytics

Pavel Zezula and Jan Sedmidubsky

Masaryk University

<http://disa.fi.muni.cz>

# Frequent Item Sets (1) – 15min

- Suppose 100 items (numbered 1 to 100) and 100 baskets (numbered 1 to 100)
  - Item  $i$  is in basket  $b$  if and only if  $i$  divides  $b$  with no remainder, i.e., item 1 is in all the baskets, item 2 is in all fifty of the even-numbered baskets, etc.
- Tasks:
  - 1) Identify the frequent items when the support threshold is set to 5
  - 2) Compute the confidence of these association rules
    - a)  $\{5, 7\} \rightarrow 2$
    - b)  $\{2, 3, 4\} \rightarrow 5$

# Frequent Item Sets (2) – 20min

- Consider the following twelve baskets, each of them contains 3 of 6 items (1 through 6):
  - {1, 2, 3} {2, 3, 4} {3, 4, 5} {4, 5, 6}
  - {1, 3, 5} {2, 4, 6} {1, 3, 4} {2, 4, 5}
  - {3, 5, 6} {1, 2, 4} {2, 3, 5} {3, 4, 6}
- Suppose the support threshold is 4. On the first pass of the PCY algorithm, a hash table with 11 buckets is used, and the set  $\{i, j\}$  is hashed to bucket  $ixj \bmod 11$ :
  - 1) Compute the support for each item and each pair of items
  - 2) Which pairs hash to which buckets?
  - 3) Which buckets are frequent?
  - 4) Which pairs are counted on the second pass?

# Finding Similar Items (1) – 5min

- Compute the Jaccard similarities of each pair of the following three sets:
  - $A = \{1, 2, 3, 4\}$
  - $B = \{2, 3, 5, 7\}$
  - $C = \{2, 4, 6\}$

# Finding Similar Items (2) – 5min

- Consider two documents A and B
  - If their 3-shingle resemblance is 1 (using Jaccard similarity), does that mean that A and B are identical?
    - If so, prove it. If not, give a counterexample.

# Finding Similar Items (3) – 10min

- Consider two documents A and B
  - Each document's number of token is  $O(n)$
  - What is the runtime complexity of computing A and B's  $k$ -shingle resemblance (using Jaccard similarity)?
    - Assume that comparison of two  $k$ -shingles to assess their equivalence is  $O(k)$
    - Express your answer in terms of  $n$  and  $k$ , where  $n \gg k$

# Finding Similar Items (4) – 25min

- For the matrix

| Element | S <sub>1</sub> | S <sub>2</sub> | S <sub>3</sub> | S <sub>4</sub> |
|---------|----------------|----------------|----------------|----------------|
| 0       | 0              | 1              | 0              | 1              |
| 1       | 0              | 1              | 0              | 0              |
| 2       | 1              | 0              | 0              | 1              |
| 3       | 0              | 0              | 1              | 0              |
| 4       | 0              | 0              | 1              | 1              |
| 5       | 1              | 0              | 0              | 0              |

- 1) Compute the minhash signature for each column using the following hash functions:
  - $h_1(x) = 2x + 1 \pmod 6$
  - $h_2(x) = 3x + 2 \pmod 6$
  - $h_3(x) = 5x + 2 \pmod 6$
- 2) Which of these hash functions are true permutations?
- 3) How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?

# Data Streams (1) – 20min

- Suppose we are maintaining a count of 1s using the DGIM method
  - Each bucket is represented by  $(i, t)$ 
    - $i$  – the number of 1s in the bucket
    - $t$  – the bucket timestamp (time of the most recent 1)
- Consider the following properties:
  - Current time is 200
  - Window size is 60
  - Current buckets are:
    - $(16, 148)$   $(8, 162)$   $(8, 177)$   $(4, 183)$   $(2, 192)$   $(1, 197)$   $(1, 200)$
  - At the next ten clocks (201 through 210), the stream has 0101010101
- What will the sequence of buckets be at the end of these ten inputs?