FACULTE INFORMATIQUE ET COMMUNICATIONS
SCHOOL OF COMPUTER AND COMMUNICATION SCIENCES

Laboratoire de Systèmes d'Information Repartis
Distributed Information Systems Laboratory

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Distributed Information Systems

Prof. Karl Aberer

Final exam, winter semester 2012 / 2013
January 18th 2013, 12:15 – 15:15

The following materials are allowed: lecture slides, exercise sheets and solutions, past exams with your own solution, and personally written notes. You can use a pocket calculator but any other electronic devices (including mobile phones, laptops, handheld devices, etc.) **must be switched off**. The exam consists of 28 pages including the cover sheet. Please write your answers *only* on the appropriate pages.

- If necessary, you can ask for additional sheets.
- Do not separate the exam sheets (by unstapling).
- Please number the additional pages and do not forget to put your name on them.

Student name: **Karl Aberer** Seat number: **0**

PLEASE HAVE YOUR STUDENT CARD READY FOR CONTROL

GOOD LUCK!

Each question receives a maximum of **25 points**.

Question 1	Question 2	Question 3	Question 4	Total points

Question 1: Chord

Question 1.1 (12 pts)

Assume you are given a Chord network with 10 peers named A, B, ..., J. The keys are generated using a hash function $\mathcal{H} : \Sigma \rightarrow [1,64]$, where Σ is the space of all strings.

Given $\mathcal{H}(A) = 2$, $\mathcal{H}(C) = 7$, $\mathcal{H}(D) = 10$, $\mathcal{H}(E) = 12$, $\mathcal{H}(F) = 16$, $\mathcal{H}(I) = 36$ and the finger tables for all the peers as shown in Table 1, draw the Chord ring and place all the peers at their exact location. You will have to determine $\mathcal{H}(B)$, $\mathcal{H}(G)$, $\mathcal{H}(H)$ and $\mathcal{H}(J)$.

i	d	peer A	peer B	peer C	peer D	peer E	peer F	peer G	peer H	peer I	peer J
1	1	B	C	D	E	F	G	H	I	J	A
2	2	C	C	D	F	F	H	H	I	J	A
3	4	C	D	E	F	G	H	H	I	J	A
4	8	E	F	F	H	H	I	I	I	A	A
5	16	H	H	H	I	I	I	I	J	A	A
6	32	I	J	J	J	A	A	A	A	C	E

Table 1: Finger tables for all the peers.

Question 1.2 (5 pts)

Suppose we have a resource with key 17. Identify where it is stored and, starting from peer *I*, explain all the steps needed to process a search for this resource. Furthermore, provide the names of the peers along the search path and explain why these peers are chosen.

Question 1.3 (8 pts)

A new peer M joins with $\mathcal{H}(M) = 50$. Rewrite Table 1 including also the column for M and, if needed, update the other columns.

Question 2: XML Filtering

Given an XML document:

```
<city>
    <name>A</name>
    <museum>
        <name>B</name>
        <ticket>X</ticket>
    </museum>
    <park>
        <name>C</name>
    </park>
</city>
```

and three XPath queries (in this specific order):

1. `/city/name/park`
2. `//park`
3. `/city//museum/ticket`

Question 2.1 (3 pts)
Determine the path nodes for the 3 queries above (without using load balancing while generating such path nodes).

Question 2.2 (4 pts)
For all elements, provide the candidate lists and waiting lists that are constructed for building the query index (without using load balancing).

Question 2.3 (9 pts)
Provide a detailed description of the processing steps of the XML document filtering engine when processing the XML document given above. For each step, provide in a table:

- the type of event processed (start/end/PCDATA)
- the action preformed for the event (match, promote/remove from candidate list)
- the candidate list after the processing of the event

Question 2.4 (3 pts)
Determine the matched and unmatched queries.

Question 2.5 (6 pts)
When using load balancing:

(a) determine the path nodes for the three queries above (following the specific order)

(b) for all the elements, provide the candidate lists and waiting lists that are constructed

Question 3: Data Mining

Question 3.1 (5pts)

Consider a marketing dataset consisting of 100 transactions. The support for the itemsets $\{X\}, \{Y\}$ and $\{X, Y\}$ are 50%, 80% and 40%, respectively.

a. Compute the confidence of the association rule: $\{X\} \rightarrow \{Y\}$

b. Is the rule $\{X\} \rightarrow \{Y\}$ interesting if the threshold for both support and confidence is 60% ?

Question 3.2 (5pts)

Given the distribution of points in Figure 1, assume that you execute k-Means with $k = 3$. For each of the 5 clusterings below explain (with a single sentence) if and why it can be a valid result from running the algorithm:

1. $\{C1 \cup C2\}$, $\{C3 \cup C4\}$, $\{C5\}$
2. $\{C1 \cup C2\}$, $\{C3\}$, $\{C4 \cup C5\}$
3. $\{C1 \cup C4 \cup C5\}$, $\{C2\}$, $\{C3\}$
4. $\{C1\}$, $\{C2\}$, $\{C3 \cup C4 \cup C5\}$
5. $\{C1\}$, $\{C2 \cup C3\}$, $\{C4 \cup C5\}$

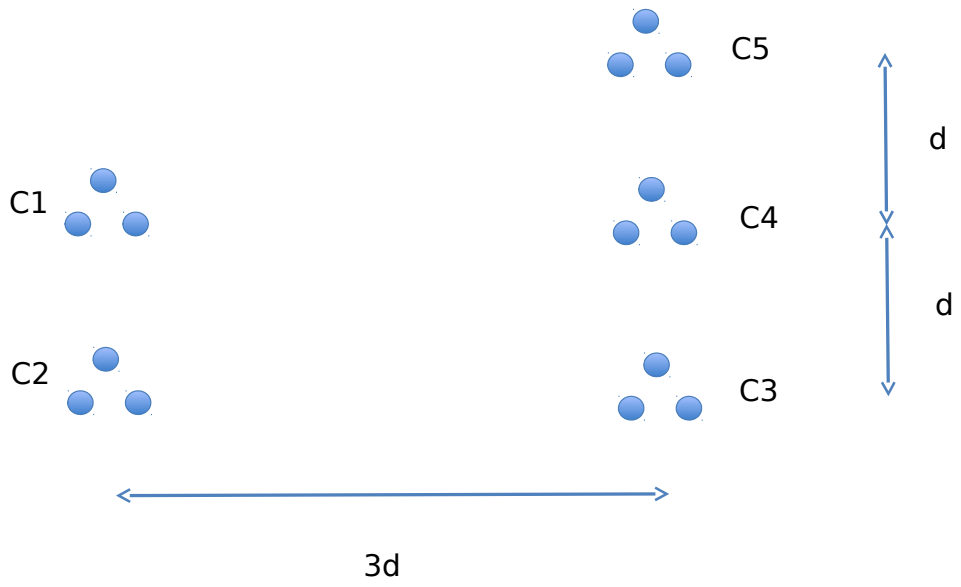

Figure 1: $C1 \ldots C5$ represent sets of points. d denotes a unit distance.

Question 3.3 (15pts)

Consider the training dataset in Table 2, and the binary class attribute **Swimming**. Assume you apply decision tree induction (ID3/C4.5). Answer the following questions:

a. What is the initial entropy of this data set?

b. What is the information gain when using the attribute **AirTemp** for splitting the data set? And when using **Sky**?

c. Compute the complete decision tree for this data set.

Table 2: Training dataset

Instance	**AirTemp**	**Sky**	**Humidity**	**Swimming**
1	> 30	sunny	normal	Y
2	< 20	sunny	high	N
3	> 30	cloudy	high	N
4	> 30	cloudy	normal	Y
5	< 20	cloudy	normal	N
6	$20 - 30$	sunny	normal	Y
7	$20 - 30$	cloudy	high	N
8	$20 - 30$	cloudy	normal	N
9	< 20	sunny	normal	N

Question 4: Latent Semantic Indexing

Consider the following term-document matrix:

	d_1	d_2	d_3	d_4	d_5	d_6
Groovy	1	0	1	0	1	0
Java	1	1	1	0	0	0
Python	0	0	1	1	1	0
Ruby	0	1	0	1	0	1
Scala	0	1	0	0	1	1

The singular value matrix for the above term-document matrix is:

$$S = \begin{pmatrix} 2.80 & 0 & 0 & 0 & 0 \\ 0 & 1.94 & 0 & 0 & 0 \\ 0 & 0 & 1.41 & 0 & 0 \\ 0 & 0 & 0 & 1.12 & 0 \\ 0 & 0 & 0 & 0 & 0.40 \end{pmatrix}$$

The *SVD term matrix K* is the following:

$$K = \begin{pmatrix} 0.48 & -0.53 & 0 & 0.3 & -0.64 \\ 0.46 & -0.21 & -0.71 & -0.31 & 0.38 \\ 0.46 & -0.21 & 0.71 & -0.31 & 0.38 \\ 0.38 & 0.65 & 0 & -0.46 & -0.47 \\ 0.45 & 0.46 & 0 & 0.71 & 0.29 \end{pmatrix}$$

The *SVD document matrix* D^t is the following:

$$D^t = \begin{pmatrix} 0.34 & 0.46 & 0.50 & 0.30 & 0.50 & 0.29 \\ -0.38 & 0.46 & -0.49 & 0.22 & -0.15 & 0.57 \\ -0.5 & -0.5 & 0 & 0.5 & 0.5 & 0 \\ 0 & -0.05 & -0.28 & -0.69 & 0.63 & 0.22 \\ -0.64 & 0.49 & 0.30 & -0.23 & 0.08 & -0.45 \end{pmatrix}$$

Question 4.1 (5pts)

What is (are) the disadvantage(s) of LSI compared to Boolean retrieval? And the one(s) compared to Vector Space retrieval?

Question 4.2 (10pts)

Rank the documents according to their similarity to the query $q =$ (Java, Scala, Python). Use a **2** dimensional concept space.

Hint: Recall that A^{-1} is the inverse of n-by-n matrix A iff $A^{-1}A = AA^{-1} = I_n$, where I_n is the n-by-n identity matrix, i.e. a matrix with ones in the main diagonal and zeros elsewhere.

Question 4.3 (6pts)

We add document $d_7 =$ (Groovy, Ruby) to the collection. Rank documents d_1 to d_7 according to the similarity to the query q in question 4.1. Use a **2** dimensional concept space.

Hint: Any document vector can be transformed into concept space similarly to how a query vector is transformed into concept space.

Question 4.4 (4pts)

Does it affect the results if we don't recompute the SVD every time a new document is added to the document collection? Why?