**Task 1: Query Hierarchy**

Given is the following database, with two additional tables: “Cite” (to join Paper and Paper) and “Write” (to join Author and Paper).

<table>
<thead>
<tr>
<th>Author</th>
<th>Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>name</td>
</tr>
<tr>
<td>1</td>
<td>Charlie Carpenter</td>
</tr>
<tr>
<td>2</td>
<td>Michael Richardson</td>
</tr>
<tr>
<td>3</td>
<td>Michelle</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1.1 Create the query hierarchy for the query $K = \{\text{Michelle,XML}\}$, where no query has more than two joins.

1.2 Which query construction option would you definitely not present to the user?

**Task 2: Estimating the Probability of a Query Interpretation**

Given are the following database, keyword query and query:

<table>
<thead>
<tr>
<th>Actor</th>
<th>Movie</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>name</td>
</tr>
<tr>
<td>1</td>
<td>Tom Hanks</td>
</tr>
<tr>
<td>2</td>
<td>Collin Hanks</td>
</tr>
<tr>
<td>3</td>
<td>Tom Cruise</td>
</tr>
</tbody>
</table>

$K = \{\text{hanks,2001,cruise}\}$, 
$Q = \sigma_{\text{hanks} \in \text{name}(\text{Actor})} \bowtie \text{Acts} \bowtie \sigma_{\text{2001} \in \text{year}(\text{Movie})} \bowtie \text{Acts} \bowtie \sigma_{\text{cruise} \in \text{name}(\text{Actor})}$. 

Task 2.1 Explain in words what $P(Q|K)$ describes.

Task 2.2 Estimate $P(Q|K)$. 
Task 3: Selecting a Query Construction Option

Given is the following query hierarchy:

Predict which query construction option (O₁, O₂, or O₃) should be presented to the user.

Task 2.1 Explain in words what the graph shows and which criteria lead to the QCO selection.

Task 2.2 What do you need to compute?

Task 2.3 Compute the needed values and decide for a QCO.
Appendix: Formulas

QCP: Query Construction Plan
QCO: Query Construction Option
IG: Information Gain

Probability of a query interpretation
\[ P(Q|K) = P(I, T|T) \propto \left( \prod_{k_i \in k} P(A_i : k_i|A_i) \right) \cdot P(T) \]
\( I \) – the set of keyword interpretations \( \{A_i : k_i\} \) in \( Q \)
\( T \) – the template of \( Q \)

Probability of a keyword interpretation
\[ P(\sigma_{k_i} \in A_i | \sigma_\sigma \in A_i) \text{ can be estimated using Attribute Term Frequency (ATF):} \]
\[ ATF(k_i, A_i) = \frac{TF(k_i, A_i) + \alpha}{N_{A_i} + \alpha B} \]
\( N_{A_i} \) – the number of keywords in \( A_i \)
\( \alpha \) – a smoothing parameter, typically \( \alpha = 1 \) (Laplace smoothing)
\( B \) – the vocabulary size

Probability of a query template
\[ P(T) = \frac{\#occurrences(T) + \alpha}{N + \alpha B} \]
\( \#occurrences(T) \) – number of queries in the log using \( T \) as a template
\( N \) – the total number of queries in the log
\( \alpha \) – a smoothing parameter, typically set to 1 (Laplace smoothing)
\( B \) – the vocabulary size

When the query log is absent or is not sufficient, we assume that all query templates are equally probable.

A measure of QCO efficiency and probability estimation for QCOs
\[ H(\zeta) = -\sum_{I \in \zeta} P(I) \cdot \log_2 P(I) \]
\( H(\zeta) \) – entropy of the query interpretation space
\[ IG(O) = H(\zeta) - H(\zeta|O) \]
\( IG(O) \) – the expected information gain of a QCO as entropy reduction
\[ H(\zeta|O) = P(O) \cdot H(\zeta|O) + P(\neg O) \cdot H(\zeta|\neg O) \]
\( H(\zeta|O) \) – the entropy of the interpretation space given the QCO
\[ P(O) = \sum_{I \in \zeta/O} P(I) \]
\( P(O) \) – the probability of a CQO using probabilities of the subsumed query interpretations