Exam 2017 – Web Data Management

S. Abiteboul & P. Senellart

March 2017

The exam is 2 hours. All documents are allowed. Internet access and communicating devices are disallowed. The last exercise should by done on a separate sheet, as it will be graded separately.

1 XML typing (7 points)

Consider the following four sets of XML documents:
D_1: ⟨a⟩ ⟨b⟩ ⟨c⟩x⟨/c⟩ ⟨/b⟩ ⟨b⟩ ⟨e/⟩ ⟨/b⟩ ⟨/a⟩
D_2: ⟨a⟩ ⟨b⟩ ⟨c⟩x⟨/c⟩ ⟨/b⟩ ⟨b/⟩ ⟨/a⟩
D_3: ⟨a⟩ ⟨b⟩ ⟨c/⟩ ⟨/b⟩ ⟨b⟩ ⟨c⟩x⟨/c⟩ ⟨c/⟩ ⟨/b⟩ ⟨/a⟩
D_4: ⟨a⟩ ⟨b⟩ ⟨c⟩x⟨/c⟩ ⟨/b⟩ ⟨d⟩ ⟨e⟩y⟨/e⟩ ⟨/d⟩ ⟨/a⟩
where `x` and `y` stand for arbitrary text nodes. Call these sets D_1, D_2, D_3, D_4.

1. For each $D_i \in \{D_1, D_2, D_3, D_4\}$, give a DTD (if one exists) that accepts exactly D_i. Otherwise explain briefly what cannot be captured. The syntax you use for the DTD does not need to be the standard one.

2. For each $D_i \in \{D_1, D_2, D_3, D_4\}$, is there an XML schema that accepts exactly D_i? If yes, you do not need to give it. Otherwise, explain briefly what cannot be captured.

3. Summarize your results of the first two questions in a table of the form:

	D_1	D_2	D_3	D_4
DTD	yes/no	yes/no	yes/no	yes/no
XML Schema	yes/no	yes/no	yes/no	yes/no

4. Each time you answered "no" in the previous question, give the schema (DTD or XML Schema, according to the case) that is as restrictive as you can and validates D_i.

5. Give a DTD that is as restrictive as you can and validates the four sets of documents (i.e., $\bigcup_{i=1}^{4} D_i$).

6. Describe in words (10 lines maximum) an XML Schema as restrictive as you can that validates the four sets of documents (i.e., $\bigcup_{i=1}^{4} D_i$).

2 Distributed Computing and XML (7 points)

A Web site that aggregates reviews of products (say, movies) stores this information as a collection of XML documents, which all have the same structure:

```
<review>
  <item ref="q12344321" />
  <user>helloworld</user>
  <source>MovieReviews</source>
  <text>This movie was so great I did not fall asleep
    watching it.</text>
  <score min="1" max="5">4.5</score>
</review>
```

The `min` and `max` attributes of the `score` element indicates the minimum and maximum score with respect to which the given score is to be interpreted.

1. Assume the entire collection is stored within a native XML DBMS, accessible with the expression `collection('reviews')`. Write a query in XPath 1.0, XPath 2.0, XQuery, or XSLT, as needed or preferred, to compute a list of the following form:

   ```
   <items>
     <item ref="q12344321" avgscore="17.5" />
     <item ref="q12344319" avgscore="12.0" />
     ...
   </items>
   ```

   with the average score of every movie across all reviews for this movie, where scores are normalized to be between 0 and to 20. The precise syntax you use does not have to follow closely the corresponding language standard, but the constructions used need to exist in the language. Do not worry about the number of decimals used to display average scores.

2. Now assume that the entire collection is stored on HDFS, and that one uses MapReduce to run a distributed computation over it. One wishes to produce the same output as in the previous case. Write in pseudo-code `map` and `reduce` functions for this problem.

3. What if one can use Spark instead of MapReduce for the same task? Does it make a difference in efficiency or ease of programming? How so?

3 Data Responsibility (6 points)

In the course, we have discussed issues in responsible data management. Suppose that you are confronted to the following situation: customers are complaining that the prices of staplers sold on a particular web site are unfair, namely, that low income people, women, and senior citizens are charged more. The company claims that this is just the effect of a policy based on the distance to the stores of competitors.

1. You are asked to define a protocol to check who is right. Explain (in half a page max) which data you use, how/where you obtain it, how you proceed. Note that it is reasonable to assume that you dispose of census data, the location of the stores of the competitors, the IP address of customers, some basic information on them (they need to provide a login to the web site); and that you can use testers.

2. Suppose the company is indeed only using the distance to the stores of competitors. How can they make their policy transparent? If they do so, do you think this will make their policy fair?