

Big Data – Exercises

Spring 2020 – Week 4 – ETH Zurich

Introduction

This exercise will cover XML and JSON well-formedness.

We recommend you use an online editor to validate and check your solutions for well-formedness, such as [this \(https://www.xmlvalidation.com/\)](https://www.xmlvalidation.com/) for XML and [this \(https://jsonlint.com/\)](https://jsonlint.com/) for JSON. To edit XML files, something like [this \(https://www.webtoolkitonline.com/xml-formatter.html\)](https://www.webtoolkitonline.com/xml-formatter.html) might work better, which offers syntax highlighting and formatting.

If you prefer, you can instead install oXygen, an XML/JSON development IDE. Download links can be found [on the course website \(https://www.systems.ethz.ch/courses/spring2020/bigdataforeng/material\)](https://www.systems.ethz.ch/courses/spring2020/bigdataforeng/material), and the license key is available on the course Moodle.

1 XML

1.1 Well-formedness

For each of the following XML documents, say if it is well-formed or not; if not, then also correct the errors. An online editor may help with these tasks, but first try to solve them without software support.

Document 1

```
<Burger>
  <Bun>
    <Pickles/>
    <Cheese/>
    <Patty/>
  </Bun>
</Burger>
```

Document 2

```
<Burger>
  <Bun>
    <Pickles/>
    <Cheese/>
    <Patty/>
  </Bun>
</Burger>
<Cola>
  <Sugar/>
  <Water/>
</Cola>
<Fries kind="French"/>
```

Document 3

```
<Email>
  Dear Mr. John Doe,

  I hereby kindly request you to revise the attached draft
  of the merger contract and provide me with any feedback you may have.

  Best regards,
  Jane Doe
  Legal Department.
</Email>
P.S. Also, please stop sending me funny cat pictures; I laugh all day and cannot work.
```

Document 4

```
<Band name="Metallica"
  member="James Hetfield"
  member="Lars Ulrich"
  member="Kirk Hammett"
  member="Robert Trujillo"/>
```

1.2 Create your own XML

Copy the text of the introduction above (including the title until the text 'by email') and paste it into an editor as plain text. Create a possible XML document keeping the same context and including formatting (title, sections, style, links, etc.).

Questions

1. Is your XML well-formed? If not, correct any mistakes. You can use an editor to help you with this task.
2. Is this data structured, unstructured, or semi-structured?

1.3 XML Names

Which of the following are valid XML Names?

1. <_bar/>
2. <Xmlelement/>
3. <Foo/>
4. <foo123/>
5. <foo_123/>
6. <foo-123/>
7. <foo#123/>
8. <foo.123/>
9. <-123/>
10. <123foo/>
11. <doctype/>

1.4 Predefined entities

XML has only 5 predefined entities. Associate each escape code with the corresponding character.

- | | |
|-----------|---|
| 1. < | > |
| 2. & | " |
| 3. > | ' |
| 4. " | & |
| 5. ' | < |

2. JSON

2.1 JSON Values

JSON documents are composed of name–value pairs. List the 6 possible JSON value types.

2.2 JSON well-formedness

For each of the following JSON documents, state if you see any syntax mistakes. If yes, correct them.

Document 1

```
{
  "burger" : {
    "bun" : ["pickles", "cheese", "patty"],
    "extraIngredients" :
  }
}
```

Document 2

```
{
  "pizza" : {
    "topping" : "salami",
    "topping" : "cheese",
    "topping" : "oregano"
  }
}
```

Document 3

```
{
  name : "John Doe",
  age : 42,
  occupation : "Penguin Turner",
  motto : "Up is life!"
}
```

3. Conversions from a relational database

Imagine that we have a relational table which stores chat messages. Translate this table into XML and JSON.

conversation_id	people	sender	content	timestamp	is_read	attachment_id
42	charlie,ari,jesse	charlie	hey, here's the doc ><	1510410193	TRUE	NULL
42	charlie,ari,jesse	charlie	NULL	1510410244	TRUE	doc_6492
42	charlie,ari,jesse	ari	thanks!	1510432987	FALSE	NULL
17	rudy,sage	rudy	look at this cute cat!	1500897189	TRUE	img_91847
17	rudy,sage	NULL	aww	1506610190	TRUE	NULL

In []:

4.