INF 551 – Fall 2016 (Afternoon)

Quiz 4: File formats (10 points)

Solution

a = 10

C = 12 = 8+4

D = 13 = 8+4+0+1

1\.

110x xxxx 10xx xxxx

11 + 7

Hex UTF-8: C2 A2 42

Binary UTF-8: 11000010 10100010 01000010 (1.5)

Binary Code Point: 00010100010 1000010 (1.5)

Hex Code Point: U+00A2 U+0042 (1)

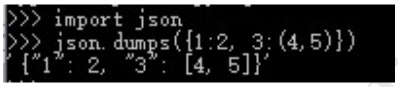
C2A242(hex) = 1100 0010 1010 0010 0100 0010 (binary)

→ binary code point = 000 1010 0010(bi) & 100 0010 (binary)

⇒ U+ A2 U+ 42

2\.

'{"1":2, "3":[4,5]}'



```
>>> import json
>>> json.dumps({1:2, 3:(4,5)})
'{"1": 2, "3": [4, 5]}'
```

json.dumps() is a JSON encoder which converts Python object to JSON documents:

Python list => JSON array

Python tuple => JSON array (1)

Python dictionary => JSON object (1), keys as strings (1)

3\.

The value of attribute "price"(1) in "book" elements under "bib"(0.5) that contain an "author" subelment(0.5) whose content contains the word "Ullman"(1).

Name: ______________________________ USC ID: ________________________

E 14

INF 551 – Fall 2017 (Morning section)

Quiz 5: File Formats (10 points), **15 minutes**

1. [5 points] Unicode code point for the Chinese character 中 (means “middle”) is U+4E2D. Give its **UTF-8** encoding in both **binary** and **hexadecimal** formats.

U+ 4E2D

Hexadecimal: **E4 B8 AD**

0100 1110 0010 1101

Binary: 1110 0100 1011 1000 1010 1101

1110 0100 1011 1000 1010 1101

E4 B8 AD

2. [5 points] For each JSON value in the table below, indicate if it is valid. If it is not valid, provide a reason in the last column.

JSON Value	Valid? (Y/N)	Reason (if it is not valid)
{[25]} ✗	N	You have the key but no value in JSON object
[25, {}, Null] ✗	N	Null should be null
“name” : ”john” ✗	N	Should be inside {}
[“name” : 25] ✗	N	Should be either {“name” : 25} or [“name” , 25]
{“name” : []} ✓	Y	
☆ ["foo", {"bar": ["baz", null, 1.0, 2]}]	Y	
{25: “mary”} ✗	N	Key should be string

Name: ______________________________ USC ID: ________________________

INF 551 – Fall 2017 (Afternoon section)

Quiz 5: File Formats (10 points), **15 minutes**

c = 12 = 8 + 4 + 0 0
d = 13 = 8+4+0+1

1. [6 points] Unicode code point for the Chinese character 你 (means “you”) is U+4F60. Give its **UTF-8** encoding in both **binary** and **hexadecimal** formats.

U+4F60: 0100 111101 100000
Binary UTF-8: 11100100 10111101 10100000

Hexadecimal UTF-8: E4 BD A0

U+4F60
= ~~0010~~ 0100 1111 0110 0000

encode into UTF-8 format:
1110 ~~0010~~ 0100 1011 1101 1010 0000 (binary)
= e~~2~~4 bd a0 (hex)

呜呜… 太不小心了

2. [4 points] Consider an JSON document “person.json” as show below. Consider loading it into Python as an object p: p = json.load(open(“person.json”)). Write a Python script for each question below.

   a. What is the last name of the person?

   p[“lastName”]

   b. Which city does the person live?

   p[“address”][“city”]

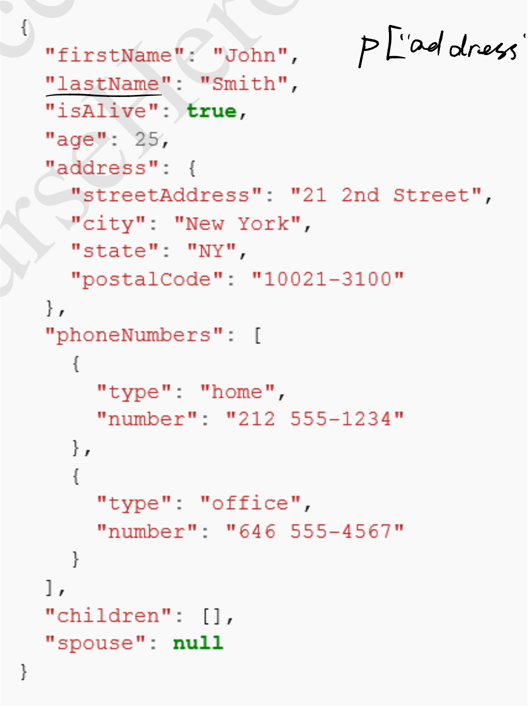
   c. What is his **second** phone number?

   p[“phoneNumbers”][1][“number”]

只有XML是从1开始排序的.

   d. How many children does the person have?
   len(p[“children”])

p["address"]["city"]



```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```

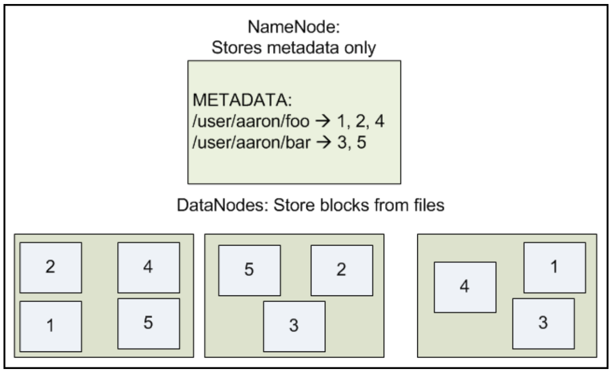
p["phoneNumber"][1]["number"]

2017 Spring

Name: ______________________________ USC ID: ________________________

Quiz 4: HDFS & File Formats (10 points. 15 minutes)

1. [5 points] Refer to the following diagram on an example HDFS. Answer the following questions.



a. [1 point] What is the replication factor in this HDFS?

Each block has two replicas distributed across three DataNodes.

Thus the replication factor in this HDFS is **2**.

b. [1 point] Which node does the client first contact when reading/writing a file?

Client first contacts **NameNode**, which informs the client of the closest DataNodes storing blocks of the file when reading, and selects DataNodes for holding its replicas when writing.

c. [1 point] What is the typical size of a block in HDFS?

64MB, which is much large than disk block size.

2. U+4E2D = 0100 1110 0010 1101

encode into UTF-8:

1110 0100 1011 1000 1010 1101 (binary)

= e4 b8 ad (hex)

d. [2 points] When writing a file in HDFS, how many packets is each block divided into? What is the size of each packet?

One point for each

One block, which is 64MB, is divided into **1024** packets, each of which is **64KB**.

2. [3 points] Unicode code point for the Chinese character 中(means middle) is U+4E2D. Give its **UTF-8** encoding in both **binary** and **hexadecimal** formats.

U+4E2D is within the range from U+0800 to U+FFFF, denoting that the code sequence length being 3. U+4E2D in binary is **0100 1110 0010 1101**. Encode in the following steps:

1. Take 6 bits at a time backwards from end and add leading **10** to form the last two code units;
2. Add leading **111**, which indicates this code point consists of 3 code units, to the rest 4 bits and 0's to any unfitted spaces (one 0 in this case) to form the first code unit;
3. The binary code will be: **11100100 10111000 10101101**.
4. The hexadecimal code will be: **E4 B8 AD**

0.5 point for each transformation error between binary and hexadecimal formats, and each division and completion errors when forming code units. 2 points for wrong number of code units.

3. [2 points] What is the output of `json.dumps(['foo', {25: ('bar', None, 1.0, 2, False)}])`?

```
["foo", {"25": ["bar", null, 1.0, 2, false]}]
```

python tuple → json list

0.5 point deducted for each minor error, such as quotation marks and wrong capitalization. 1 point deducted for each wrong data structure.

Name: ______________________________ USC ID: ________________________

INF 551 – Spring 2018

Quiz 4: File format (10 points), 15 minutes

1. [6 points] The Unicode code point for the math symbol '∈' (meaning "is an element of") is U+2208. Derive its UTF-8 encoding in **both** binary and hexadecimal formats.
   $(2208)_{hex} = (0010\ 0010\ 0000\ 1000)_{binary}$
   Binary: 11100010 10001000 10001000
   Hexadecimal: E2 88 88

U+2208

2208 ⇒ b 0010 0010 ~~0000~~ 1000

The binary form is: 1110 0010 1000 1000 1000 1000

The hex form is: e2 88 88

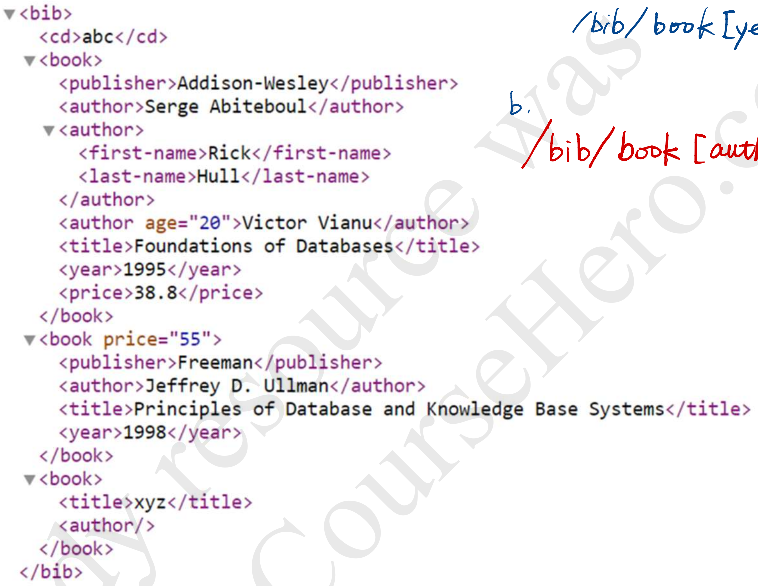
2. [4 points] Consider the following XML document shown in class. Write an XPath for each of the following questions.

a.

/bib/book[year > 1995]/title

b.

/bib/book [author/@age=20] /title



```
▼<bib>
   <cd>abc</cd>
  ▼<book>
     <publisher>Addison-Wesley</publisher>
     <author>Serge Abiteboul</author>
    ▼<author>
       <first-name>Rick</first-name>
       <last-name>Hull</last-name>
     </author>
     <author age="20">Victor Vianu</author>
     <title>Foundations of Databases</title>
     <year>1995</year>
     <price>38.8</price>
   </book>
  ▼<book price="55">
     <publisher>Freeman</publisher>
     <author>Jeffrey D. Ullman</author>
     <title>Principles of Database and Knowledge Base Systems</title>
     <year>1998</year>
   </book>
  ▼<book>
     <title>xyz</title>
     <author/>
   </book>
 </bib>
```

a. [2 points] Find the titles of the books published after 1995.
   /bib/book[year > 1995]/title

b. [2 points] Find the titles of the books written by someone at the age of 20.
   /bib/book[author/@age = 20]/title