Übung 03: Datentyp bytes, Encoding, XML und JSON Formate und Exceptions

Programmiertechniken in der Computerlinguistik II, FS 17

Abgabedatum: 4. März 2017, 23:59

Hinweise zur Abgabe

- Bitte gib jedes Python-Programm in einer eigenen Datei ab, die die Dateiendung `.py` hat und *ausführbaren* Python-Code enthält. Die Programme sollten importierbar sein.
- Gib all deine Antworten, welche keine Skripts sind, in einem PDF-Dokument oder einer Plain-Text-Datei (benannt nach dem Schema *vorname_nachname_uebung0X.pdf*) ab.
- Geize nicht mit Kommentaren direkt im Programm-Code, wo Erläuterungen angebracht sind. Umfangreiche Erklärungen werden hingegen besser in einer separaten README-Datei mitgeliefert (vorzugsweise Plain-Text oder PDF).
- Halte dich an die Vorgaben des Python Style Guide! Grobe Verstösse werden mit Punkteabzug geahndet.
- Um das Hochladen der Abgabe auf OLAT zu erleichtern, kannst du die Dateien mit `zip` (oder einem anderen verbreiteten Format) archivieren / komprimieren.

1 Datentyp bytes

Ziel dieser Aufgabe ist es, eine Binär-Repräsentation für Byte-Ketten zu erstellen. Die in Python eingebaute Funktion *bin()* funktioniert nur für Zahlen. Zu schreiben ist eine Funktion, die bytes-Folgen im Binär-Format in 8er-Blöcken darstellt, ähnlich wie das Unix-Tool hexdump. Schaue dir dazu die Wikipedia Seite zu Dump an: *Link*
Dein Program sollte in der Lage sein, eine Textdatei von STDIN als bytes-Folgen einzulesen und sie anschliessend ins Binär-Format zu konventieren und als 8er-Blöcke darzustellen. In jeder Zeile sollen fünf dieser 8er-Blöcke stehen. Füge hinter jeder Zeile den ursprünglichen Text an. Damit du bei nicht-ASCII-Zeichen in der Textspalte rechts keine Encoding- oder Zeilenumbruch-Probleme bekommst, stelle diese Zeichen jeweils wie hexdump als Punkt dar. Im Aufgaben-Ordner findest du Beispiel Text-Dateien für die Evaluierung deines Programms. Der Output sollte etwa so aussehen:

```
01010100 01101000 01101001 01110011 00100000 |This |
01101001 01110011 00100000 01100001 01101110 |is an|
00100000 01100101 01111000 01100001 01101101 | exam|
01110000 01101100 01100101 00100000 01110100 |ple t|
01100101 01111000 01110100 00101110 |ext.|
```

Abzugeben: Ein importierbares Python-Skript.

2 Encoding

2.1 Debugging

Im Aufgaben-Ordner findest du ein Programme namens *koenig.py*, bei welchen sich Encoding Fehler eingeschichen haben. Korrigiere das Programm, sodass der Output wieder lesbar ausgegeben wird.
Abzugeben: Das korrigierten Python-Skript.

2.2 Datenrettung

Im Aufgaben-Ordner findest du die Textdatei *ueberraschung.txt*. Diese Datei enthält kaputte Umlaute und nicht ASCII-Zeichen. Schreibe ein Python-Skript welches die kaputte Textdatei wieder ins richtige Encoding überführt. Dein Skript soll den richtig codierten Text in eine neue Textdatei mit dem Namen *ueberraschung_encoded.txt* schreiben.
Abzugeben: Ein importierbares Python-Skript.

2.3 Encoding Guesser

Schreibe ein Python-Skript welches das Encoding einer eingelsenen Textdatei errät. Verwende dafür Exception Handling. Dein Skript soll mindestens in der Lage sein fünf verschiedene Encodings auf diese Weise zu erraten. Zum Testen findest du fünf Beispiel-Dateien im Aufgaben-Ordner.
Abzugeben: Ein importierbares Python-Skript.

3 XML zu JSON

Du erhälst eine XML-Datei namens *book.xml*. Diese enthält Informationen zu einzelnen Büchern. Deine Aufgabe besteht darin, die XML-Datei einzulesen und sie in eine JSON-Datei umzuwandeln. Extrahiere die entsprechenden Knoten einzeln, d.h. verwendet keine automatischen Umwandlungstools von Python, wie z.B. *xmltodict*! Gehe dabei wie folgt vor:

- Lies die XML-Datei ein.
- Extrahiere und speichere die einzelnen Elemente der XML-Datei sinnvoll ab. Verwendet dazu Listen, Dictionaries, ect.
- Schreibe alle Elemente in eine JSON-Datei. Die Hierarchie der XML-Datei soll in der JSON-Datei gleich bleiben.

Deine erstellte JSON-Datei soll am Ende wie die Datei *example.json* (siehe Aufgaben-Ordner) aussehen.
Abzugeben: Ein importierbares Python-Skript.

Reflexion/Feedback

a) Fasse deine Erkenntnisse und Lernfortschritte in zwei Sätzen zusammen.

b) Wie viel Zeit hast du in diese Übungen investiert?