Übung 02: XML

Programmiertechniken in der Computerlinguistik II, FS 16

Abgabedatum: 7. April 2016, 18:00

Hinweise zur Abgabe

- Bitte gib jedes Python-Programm in einer eigenen Datei ab, die die Dateiendung `.py` hat und *ausführbaren* Python-Code enthält.
- Gib all deine Antworten, welche keine Skripts sind, in einem PDF-Dokument *vorname_nachname_uebung02.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).
- Um das Hochladen der Abgabe auf OLAT zu erleichtern, kannst du die Dateien mit `zip` oder `tar` (oder einem anderen verbreiteten Format) archivieren / komprimieren.

1 Wohlgeformtheit

Betrachte für diese Aufgabe das beiliegende XML Dokument *books.xml*. Zähle alles auf, was dieses Dokument bei einem Test für Wohlgeformtheit durchfallen lässt und notiere jeweils die dazugehörige XML Syntax Regel, welche gebrochen wird. Zusätzlich sollst du das XML Dokument duplizieren und korrigieren, so dass es den Test für Wohlgeformtheit besteht.

Abzugeben: Ein Dokument mit deinen Antworten und dein korrigiertes XML-Dokument mit dem Namen `books_corrected.xml`.

2 XPath

Folgend sind Anfragen auf das beiliegende XML Dokument *plants.xml* aufgelistet. Wandle diese Anfragen zu XPath-Ausdrücken um.

a) Der botanische Name der zweiten Pflanze

b) Anzahl Pflanzen mit Zone 'Annual'

c) Die Zone der Pflanzen mit 'Shade' als Lichtbedingung

d) Alle Namen der Pflanzen mit Zone '2'

e) Pflanzen, welche einen Preis grösser als 5 haben

Abzugeben: Das mit den XPath-Ausdrücken ergänzte Dokument aus Aufgabe 1.

3 XML mit Python: Informationen extrahieren

Schreibe ein Python Programm, welches aus dem XML Dokument *plants.xml* verschiedene Informationen extrahiert. Dein Programm soll folgende Fragen über die Pflanzen im Dokument beantworten und die Antworten im Terminal herausgeben.

- Welche Pflanzen haben bei der Eigenschaft 'Zone' einen Wert von 5?
- Wie viel würde es kosten, wenn man eines von jeder Pflanze kaufen würde?
- Welche Pflanzen brauchen Schatten, 'Shade'? Nenne ihre botanischen Namen.
- Wie viele Pflanzen gibt es insgesamt? Benutze hierfür XPath.

Benutze für diese Aufgabe die *lxml* Bibliothek.

Abzugeben: Dein Python-Programm mit dem Namen `ex03.py`.

4 XML mit Python: Datei erstellen

Schreibe nun ein Python Programm, welches aus dem Brown Korpus folgende Informationen sammelt und in Form eines XML Dokuments speichert. Das resultierende XML Dokument soll die Informationen in der hier angegebenen Form speichern.
Falls dir 'Brown Korpus' nichts sagt, befolge die Anweisungen am Ende dieser Übung.

- Anzahl Kategorie und Texte im Korpus als Attribute (des Wurzelelementes browncorpus)
- Pro Text:
  - Text-ID als Attribut
  - Kategorie als Attribut
  - Anzahl Wörter als Subelement
  - Der letzte Satz als Subelement

Abzugeben: Dein Python-Programm mit dem Namen `ex04.py`.

Zusätzliche Anweisungen zu Aufgabe 4

1) Für die Aufgabe 4 brauchst du das Natural Language Toolkit für Python. Überprüfe, ob du das NLTK Paket schon installiert hast. Öffne hierfür den Python Interpreter im Terminal und gib folgendes ein:

   ```
   >> import nltk
   ```

   Falls keine Warnungen auftauchen, gehe zu Punkt 2. Ansonsten musst du das NLTK Paket wie hier beschrieben installieren.

2) Überprüfe nun, ob du das NLTK Brown Korpus schon heruntergeladen hast, indem du im Terminal folgendes initialisierst:

   ```
   >> from nltk.corpus import brown
   ```

   Falls keine Fehlermeldungen erscheinen, kannst du mit der Aufgabe 4 fortfahren. Ansonsten musst du zusätzlich die NLTK-Daten des Pakets 'book' mittels Python-Interpreter im Terminal installieren:

```
>> nltk.download()
```
Nun sollte ein Fenster erscheinen, in dem du die Daten auswählen kannst.
Achtung: Wenn du alle NLTK-Daten (`all` anstelle von nur `book` und `all-corpora`) auswählst, benötigt das sehr viel Platz und kann eine Weile dauern.

Um zu verstehen, wie du an die Informationen im Brown Korpus gelangst, lies folgendes Kapitel 1.3 Brown Corpus im NLTK Buch.

Reflexion/Feedback

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

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