Universität Trier
Fachbereich IV – Informatik
Dr. Michael Ley

Trier, den 14.7.2017

Java-Übungsblatt zum kleinen Studienprojekt (Praktikum)

Sommersemester 2017

Abgabe bis 15.8.2017 an `ley@uni-trier.de` Betreff: `Praktikum`

Geben Sie Ihren Namen an!!! Es müssen lauffähige Programme abgegeben werden (.java - Dateien), zusätzlich muss die Ausgabe Ihrer Programme in Text-Dateien dokumentiert (und abgegeben) werden. Das Einpacken in `zip/jar`-Datein ist möglich (aber in der Regel nicht nötig). Die in der Übung besprochenen Programme finden Sie unter `http://dblps.uni-trier.de/~ley/kp17/` . Die `dblp.xml.gz`-Datei und `dblp.dtd` stehen auf dem CIP-Pool im Verzeichnis `/home/ley/`. Außerdem finden Sie die Datei (und die zum Parsen notwendige DTD) unter `http://dblp.uni-trier.de/xml/`.

`dblp.xml` ist die Ihnen vom C-Übungsblatt bekannte eine große ($> 2GBytes$) xml/Text-Datei, die alle bibliographischen Sätze (Records) des dblp-Literaturservers enthält. Diesmal muß die xml-Datei mit einem Java-SAX-Parser einlesen werden und es sollen sinnvolle Hauptspeicherdatenstrukturen mit den Klassen des Java-Collection-Frameworks (Set, List, Map, ...) verwendet werden. Sie können Java-8-Streams benutzen oder sich auf konventionelle Java-7-Techniken (Iteratoren etc.) beschränken. Sie können die unter `http://dblp2.uni-trier.de/src/` zur Verfügung stehenden Java-Klassen verwenden — müssen dies jedoch nicht.

Unter `http://dblp.uni-trier.de/xml/docu/dblpxml.pdf` finden Sie eine ausführliche Beschreibung der dblp-XML-Datei, die weit über das hinausgeht, was Sie für dieses Übungsblatt benötigen.

Ü 1: Chinesische Personennamen — 30 Punkte

Ihr Programm soll die in dblp vorkommenden Personennamen untersuchen. Personennamen stehen in der `dblp.xml`-Datei in `author`- und/oder `editor`-Elementen. Personennamen lassen sich in der Regel in mehrere Teile zerlegen (Vorname(n), Nachname(n), Zwischenname, ...). Außerdem können in dblp am Ende der Namen vierstellige Zahlen auftreten, sie werden verwendet, um bekannte Homonyme zu unterscheiden — sehen Sie sich z.B. die Seite `http://dblp.uni-trier.de/pers/hd/w/Wagner:Markus` an. In Ihrer Software sollten Sie diese Zahlen ignorieren.

Chinesische Vornamen bestehen oft aus zwei Silben (Zeichen) und die Nachnamen (Familiennamen) aus einer Silbe. Häufige chinesische Familiennamen sind `Wang, Chen, Li, Chang, Liu, Yang, Huang, Wu, Lin, Chou, Yeh, Chao, Lu, Hsu, Sun, Chu, Kao, Ma, Liang, Kuo, He, Cheng, Hu, Tsai, Tseng, Wong, She, Teng, Shen, Hsieh, Tang, Hsu` — eine ausführlichere Tabelle und detailiertere Informationen über die Struktur von chineschen Namen finden Sie z.B. in Wikipedia unter `Chinesischer Name`. In dblp wird möglichst die westliche Reihenfolge Vorname-Nachname verwendet.

In der chinesischen Schrift gibt es keine Leerstellen (Blanks) zwischen Zeichen/Worten. Bei der Transskription in das lateinische Alphabeth wird in der Regel zwischen Vor- und Nachname ein Leerzeichen eingefügt, bei den mehrsilbigen Vornamen gibt es keinen einheitlichen Standard, derselbe Name kann in den Formen `KangBin Lin, Kangbin Lin, Kang-Bin Lin, Kang bin Lin, Kang Bin Lin` geschrieben werden. Ihr Programm soll für Personen mit einem in der Tabelle verzeichneten chinesischen Familiennamen in dblp vorkommende Schreibvarianten aufspüren.

Ü 2: Inhomogene Zeitschriftennamen und Konferenzkürzel 20 Punkte

In Aufgabe 3 des C-Übungsblatts haben Sie bereits die in dblp zur Identifikation von bibliographischen Records verwendeten Schlüssel (keys) kennengelernt.

Alle zu einer Zeitschrift gehörenden Keys haben die Form journals/`Z`/..., wobei `Z` ein für die jeweilige Zeitschrift spezifisches Kürzel ist. Alle in der Zeitschrift `Commun. ACM` erschienenen Papiere haben einen Schlüssel journals/cacm/... usw. Den Namen der Zeitschrift — hier also `Commun. ACM` — finden Sie im `journal`-Feld.

Wenn eine Zeitschrift umbenannt wird, bleibt der einmal vergebene Schlüssel-Präfix (journals/`Z`) unverändert, das `journal`-Feld enthält dagegen den jeweils aktuellen Namen. Beispiel: `http://dblp.uni-trier.de/db/journals/ife/`

Ihr Programm soll Zeitschriften (und Konferenzen) mit mehreren Namen auflisten (uneinheitliche Namen können natürlich auch durch Erfassungsfehler entstanden sein).

Konferenzpublikationen haben immer einen Schlüssel der Form conf/`C`/..., `C` ist hier ein für die jeweilige Tagungsserie spezifisches Kürzel. Die Rolle des `journal`-Felds wir hier durch das `booktitle`-Feld übernommen.

Ein Beispiel für eine umbenannte Konferenzserie ist `http://dblp.uni-trier.de/db/conf/iwmm/`.

Bei den Konferenzen werden Sie eine sehr viel größere Menge von inhomogenen Namen feststellen, den Konferenzen untergeordnete Workshops belegen zusätzlich den Namensraum.