Informationsintegration Allgemeines/Übung 1

SS 2016

Yvonne Lichtblau

Allgemeines

Ablauf der Übung

- Insgesamt 5 Übungszettel
- Abgabe in Gruppen von 2-3 Personen
- Pro Übung ~zwei Wochen Bearbeitungszeit

- 6 Pflichttermine (alle zwei Wochen)
  - Ausgabe neuer Übungszettel
  - Vorstellung der Lösungen letzter Übungszettel

- Termine dazwischen
  - Klärung von Fragen
  - Übungen nach Wunsch (nach Möglichkeit vorher Email)

- Webseite:
  https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/teaching/archive/ss16/ue_infoint/

Termine im Einzelnen

- **27.04.2016**, Ausgabe 1. Übung: Web Scraping
- **11.05.2016**, Ausgabe 2. Übung, Korrektur 1. Übung
- **25.05.2016**, Ausgabe 3. Übung, Korrektur 2. Übung
- **08.06.2016**, Ausgabe 4. Übung, Korrektur 3. Übung
- **29.06.2016**, Ausgabe 5. Übung, Korrektur 4. Übung
- **13.07.2016**, Korrektur 5. Übung

Ansonsten können jeden Mittwoch Fragen geklärt werden.

Übungsschein

- Schein: Voraussetzung für die Prüfung

- Abgabe der Übungszettel in Gruppen von 2-3 Personen
  - Jeder Zettel muss bestanden werden!
  - Gruppen bestehen/scheitern nur als Ganzes
  - Vorstellung der Lösungen der letzten Übung durch 2-3 Gruppen
  - Ein Student der Gruppe muss Lösung vortragen
    → immer einen Vortrag parat haben
  - Wir behalten uns vor den Student zu bestimmen
  - Ziel: Jeder Student trägt einmal vor

Aufgaben: Abgabe

- **Implementationen in Java, C++, Python, ...**
  aber muss auf gruenau2 ausführbar sein!

- Abgaben
  - Für die Abgabe Programm und evt. Metadaten als Archiv (.zip, .tar.gz, .jar) inklusive Sourcecode und Programmaufruf per Email
  - Muss auf gruenau2 laufen!
  - I.d.R. sind Eingabe, Ausgabe und Aufrufform vorgegeben
  - PDF-Abgaben ebenfalls per Email
  - Quellcode muss einsehbar sein!

- Nichteinhaltung der Regeln: Punkteabzug

Wettbewerb

- Einige Aufgaben sind als Wettbewerb konzipiert.
- Punkte gibt es für die schnellste/korrekteste Lösung (unabhängig von den Punkten zum Bestehen der Übung)
- Gemessen wird mehrmals mittels Linux „time“ (user+sys, oder real)
- Parallelisierung lohnt sich i.d.R. nicht!
- Beste Gruppe bekommt am Ende der Veranstaltung eine kleine Überraschung.
- Wettbewerbspunkte:
  1. Platz: 5 Punkte
  2. Platz: 3 Punkte
  3. Platz: 1 Punkt

Gruppeneinteilung

Gruppe1:

Gruppe2:

Gruppe3:

Gruppe4:

Gruppe5:

Gruppe6:

Gruppe7:

Gruppen bitte als „GruppeX“ in Goya eintragen!

Übung 1
Web Scraping

(Größtenteils übernommen von Sebastian Wandelt, danke!)

Aufgabe 1

Erstellen Sie ein Programm, dass für beliebige Namen die ethnische Herkunft (Geburtsland) „errät“.

- Auf der Homepage ist eine Datei bereitgestellt, die 90 Namen von Sportlern und 10 weitere Namen beinhaltet: **01_INPUT.txt**
- Namen können Typos enthalten
  (Editabstand zum echten Namen ist maximal 3)

- Eingabe (in UTF-8):
  - ein Name pro Zeile
  - newline: \n
  - letzter Name hat ein abschließendes \n
- Ausgabe:
  - Textdatei (.tsv) mit Name und ISO-3166 Ländercodes
  - newline: \n
  - letzter Ländercode hat ein abschließendes \n

Details Eingabe/Ausgabe

- Eingabe:
  ```
  Gerd Müller
  Lin Dan
  Zinedine Zidane
  ```

- Ausgabe:
  ```
  Gerd Müller<tab>DE
  Lin Dan<tab>CN
  Zinedine Zidane<tab>IT
  ```

- Fehler der Ausgabe:
  - IT sollte FR sein
  - Korrektheit hier: 66%

Kriterien zum Bestehen der Übung

- Mindestens 30% aller Namen müssen korrekt erkannt werden
- Eingabedatei zum Testen ist natürlich vorher unbekannt!
- Eingabedatei enthält ebenfalls 90 Sportlernamen und 10 weitere Namen. Der Editabstand zu den echten Namen ist maximal 3.
- Pro Name sollte das Programm nicht länger als 10 Sekunden brauchen.
- Kurze Beschreibung Eures Ansatzes in einer PDF-Datei (Stichpunkte reichen)

Kriterien für den Wettbewerb

Zwei Wettbewerbe:

- Schnellstes Ergebnis (mit mindestens 30% Korrektheit)
- Bestes Ergebnis (gemessen an der Korrektheit)

Messung auf gruenau2 mittels Linux „time“ (real)

Es werden also zweimal 5, 3 und 1 Wettbewerbspunkte vergeben!

Hilfsmittel

- Alles ist erlaubt! Euer Programm kann (soll!) auch gerne externes Wissen verwenden, z.B. in Form von Dateien, Anfragen von Web-Services etc.

- https://en.wikipedia.org/wiki/Lists_of_sportspeople
  - Achtung: die Namen in der Eingabe sind nicht notwendigerweise in Wikipedia erfasst (Beispiel Kreisklassespieler in Berlin)!
  - Dieser Link ist nur ein Anhaltspunkt zum Start.

- ISO 3166 Ländercode: http://laendercode.net
  (Es gibt sicherlich noch andere Quellen)

Zur Orientierung

Zum Testen Eures Programms ist die Musterausgabe zu der bereitgestellten Eingabedatei auf der Homepage verfügbar: **01_Muster.txt**

Bei den 100 Personen handelt es sich (in der Reihenfolge) um:

- 10 beliebteste Sportler Deutschlands
- 10 bestbezahlten Sportler
- 10 bestbezahlten Sportlerinnen
- 10 Teilnehmer Olympics und Paralympics
- 10 aus dem UEFA Team 2013
- 10 Schwimmrekordhalter
- 10 Frauen Squash World Ranking
- 10 100m Weltrekordhalter Frauen
- 10 Kaderathleten Badmintonverband Hamburg
- 10 Max Mustermann International

Abgabe

- Abgabe bis Montag den 09.05.2016 um 12:00 Uhr

- Für die Abgabe Programm und evt. Metadaten als Archiv (.zip, .tar.gz, .jar) inklusive Sourcecode und Programmaufruf per Email an yvonne.lichtblau@informatik.hu-berlin.de senden

- Gerne auch Fragen zur Übung per Emai!

- Lauffähigkeit:
  - Programm muss auf gruenau2 ausführbar sein
  - Maximal einen Thread verwenden!
  - Der erste Parameter enthält beim Programmaufruf den Pfad auf die Eingabedatei, als zweiten Parameter den Pfad der Ausgabedatei
    Beispiel Programmaufruf:
    ```
    ./program data/eingabe.txt ausgabe.txt
    ```