Digital Humanities: Übung 2

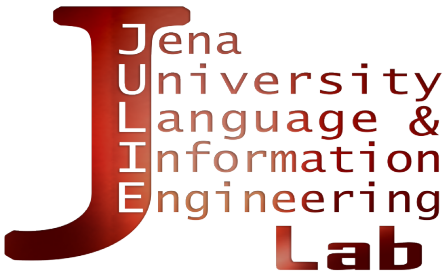
Reguläre Ausdrücke und Suche im DTA

Sven Büchel

Friedrich-Schiller-Universität Jena
Philosophische Fakultät
Institut für Germanistische Sprachwissenschaft
Lehrstuhl für Computerlinguistik
(JULIE Lab)
Sommersemester 2017

http://www.julielab.de

30. Mai 2017



Programm für Heute

Aufgaben von letzter Woche

Reguläre Ausdrücke

Suche im DTA

Aufgabe für nächste Woche

Aufgaben von letzter Woche

Aufgabe 1: Aussagenlogik

Geben Sie den Wahrheitswertverlauf der folgenden aussagenlogischen Formeln an.

1. $\neg a \vee (a \wedge b)$
2. $(a \wedge b) \vee c$

Reguläre Ausdrücke

Cheat-Sheet Reguläre Ausdrücke

Disjunktion:

- `[Bb]`
- `[A-Z], [A-Za-z]`
- `(o|ou)`

Negation:

- `[^a], [^0-9]`

Quantoren:

- `?` 0 oder 1
- `+` 1 bis n
- `*` 0 bis n
- `{n}` genau n
- `{n,m}` n bis m

Zeichenklassen:

- `\d` Ziffern, `\D` keine Ziffern
- `\w` Alphanumerische
- `\s` Whitespace
- `\b` Wortgrenze

Wildcard:

- `.` matcht jedes Zeichen

Non-Greediness:

- `?`, z.B. `a*?`

Übungen Reguläre Ausdrücke

Schreiben Sie jeweils einen Regex der (möglichst genau) folgendes erfasst.

1. Das englische Wort für “Farbe” in amerikanischer und britischer Schreibweise in Klein- und Großschreibung (etwa am Satzanfang).
2. Handynummern mit Ländervorwahl (*+49 157 8557354*, “Vorwahl” des Anbieters abgetrennt)
3. Erwähnungen von einem Herrn Friedrich Mayer, wobei auch die Kurzformen Fritz oder Fritzchen für den Vornamen verwendet werden könnten und Sie sich auch bei der Orthografie des Nachnamens unsicher sind.
4. Email-Adressen der FSU

Suche im DTA

Siehe Vorlesungsskript

Übungen zur Suche (Gruppenarbeit)

Formulieren Sie eine Suchanfrage für

- Werke von Goethe zwischen 1800 und 1830, die das Wort “Gretchen” enthalten
- Werke zwischen 1800 und 1900, die die Wörter “Gevatter” und “Tod” im selben Satz beinhalten, wobei uns vor allem die jüngsten Treffer interessieren.

Übungen zur Suche (Gruppenarbeit)(Lösung)

Formulieren Sie eine Suchanfrage für

- Werke von Goethe zwischen 1800 und 1830, die das Wort “Gretchen” enthalten

```
Gretchen #less_by_date[1800,1830]
#has[author,/Goethe/]
```

- Werke zwischen 1800 und 1900, die die Wörter “Gevatter” und “Tod” im selben Satz beinhalten, wobei uns vor allem die jüngsten Treffer interessieren.

```
Gevatter && Tod #greater_by_date[1800,1900]
```

Aufgabe für nächste Woche

Formalia

- Abgabe bis Montag, den 5.6., 23:59 Uhr
- Per Email an `sven.buechel@uni-jena.de`
- Im PDF-Format
- Unter Angabe von Vorname, Name, Matrikelnummer und Veranstaltung (in der PDF-Datei)
- Gruppenarbeit ist ausdrücklich erlaubt, am Ende muss aber trotzdem jeder eine Lösung abgeben.

Aufgabe 2.1: Reguläre Ausdrücke

Geben Sie einen Regulären Ausdruck zum Auffinden von akademischen Titeln an. Der Ausdruck muss mindestens die folgenden Beispiele erfassen.

- Dr.
- Prof. Dr.
- Dr. h.c.
- Dr. phil.
- Prof. Dr. med.
- Prof. emer.
- PD Dr.
- PD Dr. rer. nat.

Der Ausdruck sollte keine sich wiederholenden Teile enthalten (also nicht einfach alles durch Disjunktion lösen)!

Aufgabe 2.2: Suche im DTA

1. Geben sie die Adjektive an, die in Werken des Zeitraums 1700-1750 vor dem exakten Wort *„Gevatter“* stehen.
2. Welches ist das älteste Werk im DTA, das eine Wortform des Lemmas *Eisenbahn* am Anfang eines Satzes enthält?

Geben Sie jeweils auch die dafür nötige Suchanfrage an!