Übung zur Vorlesung “Computerlinguistik I”

Wintersemester 2017/2018, Prof. Dr. Udo Hahn, Sven Büchel
Übungsblatt 6 vom 7.12.2017
Abgabe bis 11.12.2017, 23.59 Uhr; per Email (PDF-Format) an sven.buechel@uni-jena.de

---

Aufgabe 1 Algorithmus zur Satzsegmentierung 5

Eine der grundlegendsten Aufgabe in der Computerlinguistik ist es, in einem Fließtext Sätze zu erkennen. Ein Programm zur Satzerkennung bekommt als Eingabe einen Text, beispielsweise “Peter lief nach draußen. Seine Jacke wurde vom Regen durchweicht.”. Die Ausgabe des Programms wären dann die Sätze “Peter lief nach draußen.” und “Seine Jacke wurde vom Regen durchweicht.”

a) 1

Eine naive Herangehensweise zur Satzerkennung ergibt sich, wenn jeder Punkt im Text als Satzgrenze aufgefasst wird. Welche Probleme sehen Sie bei diesem Vorgehen? Nennen Sie Beispiele.

b) 2

Geben Sie einen einfachen Algorithmus an, der einen Text entgegen nimmt und diesen in einzelne Sätze zerlegt. Die entsprechende Funktion soll eine Liste von Sätzen zurückgeben. Hinweis:

- Als Hilfestellung sei die Funktion `getWords(text)` gegeben. Diese Funktion bekommt einen Text, trennt ihn an Leerzeichen und gibt die so erhaltenen Wörter (ggf. mit Interpunktion!) in einer Liste zurück. Die Leerzeichen selbst gehen dabei verloren. Beispiel: `words` ← `getWords("Peter lief nach draußen.")` zerlegt den eingegeben Satz in die Elemente `"Peter"`, `"lief"`, `"nach"` und `"draußen."` (man beachte, dass das letzte Element auch den Punkt enthält, da er nicht durch ein Leerzeichen abgetrennt ist). Die Variable `words` enthält nun alle diese Wörter. Wie gehabt kann mit `words[0]` auf das `1`te Wort, mit `words[1]` auf das `2`te Wort etc. zugegriffen werden.

c) 2

Verfeinern Sie ihren Algorithmus, indem sie versuchen die Probleme, die Sie in Teilaufgabe a) identifiziert haben, zu lösen. Nehmen Sie dafür die benötigten Listen mit bestimmten Wörtern oder Funktionen zur Identifikation bestimmter problematischer Strukturen als gegeben an. Falls Sie solche Listen oder Funktionen verwenden, geben Sie bitte deren Inhalt bzw. Zweck an.

Aufgabe 2 Mengen 5

Welche der folgenden Aussagen sind richtig?

$| \{a, b, \{c, d, e\}, f\} | = 4$
$\{a, b, c\} \subset \{a, b, c\}$
$(a, b, x) \in \{a, b, c\} \times \{x, y, z\}$
$(a, 2) \in \{a, b, c\} \times \{0, 2, 4\}$
$| \{a, b, c, d\} | = 4$
$\{(b, 0), (b, 4)\} \subseteq \{a, b, c\} \times \{0, 2, 4\}$
$x \in \{a, d, m\} \cup \{r, t, x\}$
$\bigcup_{i=1}^{3} \{a, b\}^i = \{aa, b, ab, ba, bb, aaa, aab, a, aba, baa, abb, bab, bba, bbb\}$
$\{d, e, f\} \subseteq \{d, e, f\}$
$x \in \{a, d, m\} \cap \{r, t, x\}$