Mathematische Methoden der Computerlinguistik III: Statistische Methoden

Sommersemester 2015

Matt Crocker, Enrico Lieblang

Übung 4, 11.6.2015

Insgesamt gibt es 38 Punkte für dieses Übungsblatt.

1. Aufgabe(24 Punkte)
   Für die 5 Katagorien Art, Adj, NN, Vfin, Vinf sei folgende Übergangsmatrix gegeben (die Kategorien in der Reihenfolge , wie sie hier genannt sind):

$$\begin{pmatrix} 0 & 0,3 & 0,7 & 0 & 0 \\ 0 & 0,2 & 0,8 & 0 & 0 \\ 0 & 0 & 0,2 & 0,6 & 0,2 \\ 0,4 & 0,1 & 0,5 & 0 & 0 \\ 0 & 0 & 0,2 & 0,8 & 0 \end{pmatrix}$$

   Die Wahrscheinlichkeit, dass ein Satz mit der Kategorie 'Art' beginnt, sei 1. Es werde die Beobachtung ' das gute Essen' untersucht. Für die Ausgabewahrscheinlichkeiten der einzelnen Wörter seien folgende Wahrscheinlichkeiten gegeben:
   $P(das \mid Art) = 0,3$; $P(gute \mid Adj) = 0,6$; $P(gute \mid NN) = 0,4$; $P(essen \mid NN) = 0,2$; $P(essen \mid Vfin) = 0,6$; $P(essen \mid Vinf) = 0,2$

   (a) Was berechnet der Vorwärtsalgorithmus, was der Viterbi-Algorithmus? (4 Punkte)

   (b) Berechnen Sie mit Hilfe des Vorwärtsalgorithmus die Wahrscheinlichkeit für die Beobachtung, also P('das gute Essen'). (8 Punkte)

   (c) Führen Sie die Schritte des Viterbi-Algorithmus zur Berechnung der maximalen Wahrscheinlichkeit der Phrase 'das gute Essen' durch. (8 Punkte)

   (d) Geben Sie die Kategorienfolge an, welche die größte Wahrscheinlichleit der Beobachtung liefert.

   (4 Punkte)

2. Aufgabe (6 Punkte)
   Von wissenschaftlichen Texten sei bekannt, dass die mittlere Satzlänge 12 Wörter betrage und diese normalverteilt sei mit Varianz 4. Eine Stichprobe vom Umfang 25 aus wissenschaftlichen Texten ergibt eine mittlere Satzlänge von 11 Wörtern. Spricht dies auf dem Signifikanzniveau $\alpha = 0,01$ für die Tatsache, dass sich die mittlere Satzlänge von 12 Wörtern auf 11 Wörter verkürzt hat? Führen Sie einen Gauß-Test hierzu durch.

3. Aufgabe (8 Punkte)
   Es soll überprüft werden, ob das Wort 'blanker Hans' eine Kollokation darstellt. Hierbei wurden folgende Häufigkeiten festgestellt und in einer Kontingenztabelle aufgelistet:

	Hans	nicht Hans	Summe
blanker	2	45	47
nicht blanker	40	45000	45040
Summe	42	45045	45087

   Führen Sie einen Chi-Quadrat-Test zum Niveau $\alpha = 0,05$ zur Überprüfung der Hypothese durch, dass 'blanker Hans' eine Kollokation darstellt.