9. Übungsblatt - Abgabe: 15.01.2013

Aufgabe 9.1 - Statistische Modellierung

Ein statistisches Modell zur Klassifikation von Wörtern w in Adjektive ($ADJA$) und Nichtadjektive ($NADJA$) wurde auf einigen Hundert Sätzen trainiert. Man erhält dabei die folgende Frequenztabelle:

Wort *w* Artikel?	**Wort *w+1* großgeschrieben?**	***w* endet auf -er/-es/-e/-en/-em?**	**ADJA**	**NADJA**
Falsch	Falsch	Falsch	0	1758
Falsch	Falsch	Wahr	44	932
Falsch	Wahr	Falsch	3	473
Falsch	Wahr	Wahr	255	127
Wahr	Falsch	Falsch	0	0
Wahr	Falsch	Wahr	0	128
Wahr	Wahr	Falsch	0	32
Wahr	Wahr	Wahr	0	233

(a) Wie viele Klassen gibt es, wie heißen sie?

(b) Wie viele Features gibt es? Wie viele Werte haben die Features jeweils?

(c) Welches sind die möglichen Ereignisse, die sich aus unterschiedlichen Featurekombinationen ergeben? Wie groß ist also der Ereignisraum? Geben Sie zu jeder Kombination aus einem Ereignis und einer Klasse falls möglich einen Beispielsatz/eine Beispielphrase an, der/die zu dieser Kombination passt. Markieren Sie jeweils das Wort, auf das sich die Kombination bezieht. Wenn Sie Probleme haben, ein geeignetes Beispiel zu finden, beschreiben Sie, woran das liegt.

(d) Für welchen Anteil des Ereignisraumes hat das Modell Trainingsinstanzen gesehen? Was bedeutet das für die Abdeckung des Modells auf neuen Daten?

(e) Formulieren Sie Regeln, die jedem Ereignis eine Klasse zuordnen.

Aufgabe 9.2 - Evaluation

Die Folien zur vorletzten Vorlesung zeigen eine Beispielevaluation für eine binäre Klassifikationsaufgabe in Adjektive ($ADJA$) und Nichtadjektive ($NADJA$) und gibt Precision, Recall und F-Score für die Klasse $ADJA$ an.

(a) Berechnen Sie Precision und Recall für die Klasse $NADJA$.

(b) Vergleichen Sie die Ergebnisse für *NADJA* mit den Ergebnissen für *ADJA*. Was fällt Ihnen auf? Beschreiben Sie umgangssprachlich, warum dieses Ergebnis zustande kommt.

(c) Welche der beiden Evaluationen ist für die tatsächliche Brauchbarkeit des Modells in der Praxis aussagekräftiger? Begründen Sie kurz.

Aufgabe 9.3 - Evaluation

Bei der Evaluation eines Klassifiers wird das Klassifikationsergebnis für einen Text (Testkorpus) mit einer manuellen Annotation (Goldstandard) verglichen.

Gegeben ist der folgende Testkorpus (als Hilfe für Sie sind die STTS-Tags gegeben):

> Das (ART) unbeständige (ADJA) Wetter (NN) der (ART) beiden (PIDAT) letzten (ADJA) Tage (NN) wird (VAFIN) sich (PRF) auch (ADV) am (APPR) Montag (NN) fortsetzen (VVINF). (\$.) Bei (APPR) mäßigem (ADJA) Westwind (NN) treten (VVFIN) vermehrt (ADV) starke (ADJA) Regenschauer (NN) und (KON) gelegentlich (ADV) auch (ADV) Unwetter (NN) auf (PTKVZ). (\$.) Am (APPR) kommenden (ADJA) Wochenende (NN) erwarten (VVFIN) wir (PPER) den (ART) kältesten (ADJA) Tag (NN) der (ART) Woche (NN).(\$.) An (APPR) den (ART) drei (CARD) Tagen (NN) zu (APPR) Wochenanfang (NN) tritt (VVFIN) wieder (ADV) eine (ART) leichte (ADJA) Erwärmung (NN) auf (PTKVZ). (\$.)

(a) Annotieren Sie für jedes Wort (einschließlich der Satzzeichen) den Goldstandard (also die tatsächliche Zugehörigkeit zur Klasse ADJA, als + oder -), die Merkmale (entsprechend den Vorlesungsfolien) und das sich daraus ergebende Klassifikationsergebnis (+/-), das das regelbasierte Modell von den Vorlesungsfolien (Statistik I) liefert.

(b) Geben Sie für diese Evaluation die Konfusionsmatrix an.

(c) Berechnen Sie aus der Konfusionsmatrix Akkuratheit, Präzision und Recall für die Klasse ADJA.

Aufgabe 9.4 - WSD mit Bayes-Klassifier

Mit Hilfe von Kontextwörtern soll ein Klassifikator zur Word-Sense-Disambiguierung für die beiden Lesarten von *Schloss* aus einem Trainingskorpus mit 100 Dokumenten je Lesart gelernt werden.

Dabei ermittelt man die folgenden Kontextwortfrequenzen:

	Schloss_1	Schloss_2
Tür	5	23
Graf	22	3
Fahrrad	11	15
Neuschwanstein	17	1
Schlüssel	2	33
Ausflug	35	5

Außerdem gelten a-priori-Wahrscheinlichkeiten von $P(\text{Schloss}_1) = 0,4$ und $P(\text{Schloss}_2) = 0,6$.

Beobachtet wird nun ein Auftreten von Schloss mit Merkmalsmuster v_i:

	v_i
Tür	0
Graf	1
Fahrrad	1
Neuschwanstein	0
Schlüssel	0
Ausflug	1

Bestimmen Sie die wahrscheinliche Lesart!

Abgabe in Gruppen von bis zu drei Studierenden bis **15.01.2013** 10 Uhr entweder als Email im pdf-Format an **i2cl@coli.uni-sb.de** oder auf Papier im Briefkasten an der Tür von Raum 1.04 in C7.2.