9. Übungsblatt - Abgabe: 12.01.2016

Aufgabe 9.1 - Statistische Modellierung

Ein statistisches Modell zur Klassifikation von Wörtern w in Adjektive ($ADJA$) und Nichtadjektive ($NADJA$) wurde auf einem Trainigscorpus mit 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 aus dem Ereignisraum und einer Klasse falls möglich einen Beispielsatz/eine Beispielphrase an, der/die zu dieser Kombination passt. Markieren Sie darin 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 Teil des Ereignisraumes (also für wieviele von wievielen potenziell möglichen Ereignissen) hat das Modell Trainingsinstanzen gesehen? Was bedeutet das für die Abdeckung des Modells auf neuen Daten?

(e) Formulieren Sie Regeln für einen simplen Klassifikator, die jedem Ereignis genau eine Klasse zuordnen.

Hinweis: Das erste Feature wird Ihnen vielleicht unintuitiv erscheinen. Wenn man den POS-Tag des Wortes so gut kennt, dass man weiß, ob es ein Artikel ist oder nicht, könnte man ja auch gleich überprüfen ob es ein ADJA ist. *Wort w Artikel?* soll hier aber als Kurzschreibweise für *w ist eines der Wörter der, die, das, dem, den, . . . , ein, eines, . . .*

verstanden werden, also ein Wort aus einer fest definierten endlichen Menge. Das geht bei Artikeln ohne großen Aufwand, weil es nur einige wenige gibt. Dagegen ist es nicht so einfach möglich alle Adjektive aufzulisten.

Aufgabe 9.2 - Evaluation

Die Folien zur letzten 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 - Unix-Tools

In dieser Aufgabe sollen Sie die Regeln, die Sie in Aufgabe 9.1 auf einem Trainingskorpus gelernt haben auf einem Testkorpus evaluieren. Gehen Sie dabei folgendermaßen vor:

Benutzen Sie die Datei tiger_bigram.txt. Diese Datei enthält pro Zeile zwei tabseparierte Wörter sowie einen POS-Tag. Das erste Wort ist das relevante Wort w, das zweite Wort das Nachfolgerwort w+1. Nach den beiden Wörtern folgt der manuell annotierte richtige POS-Tag des Wortes w, der sogenannte Goldstandard. Sie sollen nun den Corpus mit den Regeln aus 9.1 annotieren.

(a) Finden Sie mit einem regulären Ausdruck (oder mehreren) mit grep für jede Regel die Zeilen im Corpus, die auf die Regel matchen und geben Sie die jeweiligen Befehle an. (*Hinweis:* Sie brauchen hier reguläre Ausdrücke, bei denen Sie abprüfen, ob eine Zeile auf etwas nicht matcht, also z.B kein Artikel ist. Benutzen Sie dazu *grep -v*. Auf diese Art können Sie dann mehrstufige"Befehle erzeugen: *grep -e allesWasGroßgeschrieben is | grep -v allesWasEinArtikel ist* um z.B. großgeschriebene nicht-Artikel zu bekommen.)

(b) Benutzen Sie dann *sed*, um die Datei mit einer zusätzlichen Spalte zu annotieren, die den durch die Regel gefundenen Tag (ADJA bzw NADJA) enthält. Wenden Sie dazu Ihre Regeln einzeln auf das Korpus an, pipen die Ergebnisse in *sed* und akkumulieren den Output dann in einer neuen Datei. (mit >> leiten Sie die Ergebnisse so in eine Datei um, dass eine bestehende Datei nicht überschrieben wird, sondern der neue Inhalt am Ende angehängt wird.) Geben Sie exemplarisch eine der Befehlsketten an.

(c) Für die Evaluation müssen Sie jetzt auszählen, wie oft ein ADJA als ADJA bzw als NADJA und wie oft ein NADJA als ADJA bzw NADJA erkannt wurde. Dazu

bietet es sich an, dass sie das corpus zunächst auf die beiden relevanten Spalten reduzieren (*cut*) und alle Gold-Tags, die nicht ADJA sind durch NADJA ersetzen. Dann können Sie *wc*, *sort*, *uniq* usw benutzen um die Statistik zu erstellen. Geben Sie Ihre Befehlskette und die daraus resultierende Konfusionmatrix an und berechnen Sie Precision, Recall, F-Score und Accuracy.

Abgabe in Gruppen von bis zu drei Studierenden am **12.01.2016** vor der Vorlesung.