Sprachtechnologie

6. Übung

Prof. Tobias Scheffer
Christoph Sawade

Sommer 2009

Ausgabe am: 09.06.09
Besprechung am: 15.06.09

Aufgabe 1 *Vektorraum-Modell*

Entwickeln Sie einen Textklassifikator für das Marsianische. Der Klassifikator soll Pläne für die Invasion der Erde (Klasse +1) von Texten anderen Inhalts (Klasse −1) unterscheiden. Die Liste der relevanten Terme umfasst nur *argh* und *zonk*, die in 79 bzw. 90 von 100 Texten vorkommen.

Als Trainingsmenge liegen vier von SETI abgefangene marsianische Texte vor:

a) „*argh bob argh*“, Klasse +1

b) „*zonk zonk bob*“, Klasse −1

c) „*argh zonk bob*“, Klasse +1

d) „*zonk zonk argh*“, Klasse −1

Bestimmen Sie die TF-IDF-Merkmalsvektoren und repräsentieren Sie diese im Vektorraum-Modell.

Aufgabe 2 *Lineare Klassifikatoren*

Simulieren Sie das Training eines Rocchio- und eines Perzeptron-Klassifkators mit den Trainingsdaten aus Aufgabe 1 von Hand und stellen Sie die erhaltenen Modelle graphisch dar.

Aufgabe 3 *ROC-Kurve*

Nachdem Sie die Modelle traininert haben, erhalten sie fünf weitere entschlüsselte Nachrichten. Die entsprechenden TF-IDF-Vektoren und die dazugehörigen Klassen sind in der folgenden Tabelle dargestellt.

ID	1	2	3	4	5
TF-IDF	$\begin{pmatrix}0.02\\0.99\end{pmatrix}$	$\begin{pmatrix}0.35\\0.94\end{pmatrix}$	$\begin{pmatrix}0.60\\0.80\end{pmatrix}$	$\begin{pmatrix}0.86\\0.50\end{pmatrix}$	$\begin{pmatrix}0.99\\0.09\end{pmatrix}$
Klasse	-1	+1	-1	+1	+1

Geben Sie für ihre gelernten Hypothesen jeweils die ROC-Kurve an und bestimmen Sie den AUC-Wert. Welcher Klassifikator ist besser?