10. Übungsblatt - Abgabe: 24.01.2012

Aufgabe 10.1 - Information Retrieval

Lesen Sie Jurfasky & Martin Kapitel 23 bis einschließlich 23.1.3. Kapitel 23.1.2 ist optional. Eine Kopiervorlage des Texts befindet sich im Vorlesungsordner. Dieser Text beschreibt, wie man beim Information Retrieval Dokumente und Suchanfragen als Vektoren repräsentiert und die Ähnlichkeit dieser Vektoren bestimmt.

Gegeben ist die Suchanfrage Q = *beijing duck recipe* und die beiden folgenden Dokumente:

> *D1 = „If it walks like a duck and quacks like a duck, it must be a duck.*
> *D2 = „Beijing Duck is mostly prized for the thin, crispy duck skin with some versions of the dish serving mostly the skin.*“

(a) Geben Sie die Term-Dokument-Matrix (term-by-document-matrix) an. Diese soll für Q, D1 und D2 je eine Spalte mit den absoluten Häufigkeiten der Kontextwörter *beijing, dish, duck, recipe, roast* enthalten, die jeweils dem Vektor für die Anfrage/das Dokument entspricht. (Abweichend vom Buch soll auch die Suchanfrage in der Matrix als eigene Spalte repräsentiert werden.)

(b) Berechnen Sie daraus die Kosinusähnlichkeit der Suchanfrage Q zu jedem der beiden Dokumente. Welches Dokument passt demnach besser zu der Anfrage?
(Formel wie im Buch: $sim_{cos}(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$)

(c) Warum ist es hilfreich einen Stemmer bei der Auszählung der Kontextwörter zu benutzen? Ist das Stemming für das Deutsche oder für das Englische wichtiger?

(d) Welche Wörter aus D1 und D2 würden Sie auf eine Stoppwortliste (stop list) setzen? Warum?

Aufgabe 10.2 - 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 **24.01.2012** 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.