9. Übungsblatt - Abgabe: 17.01.2011

Aufgabe 9.1 - Mutual Information

Die *pointwise mutual information* (PMI) ist ein Maß, das angibt, wie stark zwei Begriffe statistisch voneinander abhängig sind:

$$I(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)}$$

Dabei ist $P(x)$ die Wahrscheinlichkeit, dass der Begriff x in einem Dokument vorkommt, und $P(x,y)$ die Wahrscheinlichkeit, dass x und y gemeinsam in einem Dokument auftreten.

(a) Berechnen Sie die PMI paarweise für die Begriffe *Geige, Orchester, Sonate, Brot* und *Käse*. Schätzen Sie dabei die benötigten Wahrscheinlichkeiten, indem Sie die Begriffe googlen und aus den Trefferzahlen die Wahrscheinlichkeiten schätzen. Nehmen Sie dazu an, dass google auf 20 Milliarden Webseiten arbeitet.

(b) Geben Sie ein Beispiel, warum solche gegoogleten bzw. errechneten PMI-Werte nicht immer eine Aussage über die semantische Ähnlichkeit von zwei Begriffen liefern.

(c) Eine weitere Möglichkeit, die Ähnlichkeit von Begriffen zu bestimmen, besteht darin, ihren Kontext als Merkmale zur Bestimmung der Ähnlichkeit zu benutzen. Gegeben sind die Wörter *Geige, Orchester* und *Brot*. Berechnen Sie für jedes der folgenden Wörter einen Feature-Vector, indem Sie die Häufigkeit für das gemeinsame Auftauchen den Begriffs mit jedem der folgenden Kontext-Wörter als Merkmal bestimmen: *Musik, Sonate, Mehl* und *Butter*. Berechnen Sie dann die Cosinus-Ähnlichkeit für jedes der drei Wortpaare.

Hinweis: Konfigurieren Sie google so, dass nur deutschsprachige Seiten gefunden werden.

Aufgabe 9.2 - 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:

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

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

Bestimmen Sie die wahrscheinliche Lesart!

Abgabe in Gruppen von bis zu drei Studierenden bis **17.01.2011** 18 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.