

Übungen zu Information Mining, Sommersemester 2008

Ingo Frommholz (LF 138)
Sprechstunde nach Vereinbarung
ingo.frommholz@is.inf.uni-due.de

Übungsblatt 6

Abgabe bis
keine Abgabe

Aufgabe 12: Clustering im IR

Beschreibe 3 Anwendungen von Clustering im Information Retrieval. Begründe ggf., warum eine Kategorisierung in diesen Fällen nicht in Frage kommt.

Aufgabe 13: k-Means-Clustering im IR

Im Information Retrieval spielt das Vektorraummodell eine wichtige Rolle. Dabei werden alle Dokumente als Vektoren von Termen aufgefasst, deren einzelne Elemente das Gewicht eines jeweiligen Terms im Dokument wiedergibt. Dieses Modell kann auch zum Clustering verwendet werden.

Angenommen, wir haben die Termmenge $T = \{\text{haus, auto}\}$; der dazugehörige Vektorraum ist also 2-dimensional und wird von den beiden Termen "haus" und "auto" aufgespannt. Wir haben nun folgende 5 Dokumente, die sich als Vektoren beschreiben lassen (dabei sei \vec{d} der zu einem Dokument d zugehörige Vektor):

$$\begin{aligned} \vec{d}_1 &= \begin{pmatrix} 0,8 \\ 0,9 \end{pmatrix} & \vec{d}_2 &= \begin{pmatrix} 0,9 \\ 0,65 \end{pmatrix} \\ \vec{d}_3 &= \begin{pmatrix} 0,5 \\ 0,5 \end{pmatrix} & \vec{d}_4 &= \begin{pmatrix} 0,2 \\ 0,25 \end{pmatrix} \\ \vec{d}_5 &= \begin{pmatrix} 0,25 \\ 0,1 \end{pmatrix} \end{aligned}$$

"haus" hat also im Dokument d_1 das Gewicht 0,8, während "auto" dort das Gewicht 0,9 hat, usw.

- Fasse die Dokumente mittels k-Means-Clustering zusammen. Es sollen dabei 2 Cluster gebildet werden. Als initiale Seeds sollen d_4 und d_5 genommen werden. Breche ab, wenn sich die Cluster nicht mehr verändern.
- Skizziere graphisch den Dokumentenraum und die Bewegung der Zentroiden. Was fällt auf?
- Berechne für jeden Iterationsschritt die Purity und den Rand-Index. Gehe dabei von folgender manuellen Klassifikation aus: $Cl_1 = \{d_1, d_2, d_3\}$ und $Cl_2 = \{d_4, d_5\}$.