Technische Universität Braunschweig
Institut für Informationssysteme
Prof. Dr. Wolf-Tilo Balke
M. Sc. Joachim Selke

Information Retrieval und Web Search Engines
Wintersemester 2008/2009

Übungsblatt 5

3. Dezember 2008

Hinweis: Soweit nicht anders angegeben, gibt es für jede korrekt bearbeitete Teilaufgabe einen Punkt. Die Abgabe der Hausübungen ist bis spätestens zum Beginn der nächsten Vorlesung möglich – entweder persönlich direkt vor der Vorlesung oder per Einwurf in den Briefkasten des Instituts (Informatikzentrum, zweiter Stock, vor Raum 238).

Aufgabe 9 (K-Means-Clustering)

a) In welchen Fällen ist es möglich, daß ein bezüglich RSS optimales Clustering einen leeren Cluster enthält? (2 Punkte)

b) Zu einer beliebigen Dokumentenkollektion sei $\mathrm{RSS}_{\min}(k)$ die minimale RSS über alle Clusterings der Kollektion mit genau k nicht-leeren Clustern. Zeigen Sie, daß $\mathrm{RSS}_{\min}(k)$ monoton in k fällt. Wann ist $\mathrm{RSS}_{\min}(k) = 0$? (2 Punkte)

c) Clustern Sie die folgende Dokumentenkollektion mittels K-Means-Clustering und skizzieren Sie grob den Clustering-Prozeß (geben Sie unbedingt die dabei verwendeten Seeds an). Bestimmen Sie einen geeigneten Wert für k. (Sie müssen die Berechnungen nicht per Hand ausführen; geben Sie aber alle von Ihnen verwendeten Hilfsmittel an.)

Hinweis: Vergessen Sie nicht die Normierung der Dokumentenvektoren auf Einheitslänge.

Dokumentennummer	Inhalt
1	hot chocolate cocoa beans
2	cocoa ghana africa
3	beans harvest ghana
4	cocoa butter
5	butter truffles
6	sweet chocolate
7	sweet sugar
8	sugar cane brazil
9	sweet sugar beet
10	sweet cake icing
11	cake black forest

(2 Punkte)

Aufgabe 10 (Agglomeratives Clustering)

Wir betrachten noch einmal die Dokumentenkollektion aus der vorherigen Aufgabe, inklusive der Normierung der Dokumentenvektoren auf Einheitslänge. Auch hier dürfen Sie geeignete Hilfsmittel verwenden.

a) Geben Sie das zum Single-Link-Clustering gehörige Dendrogramm an. Verwenden Sie als zugrundeliegendes Maß die euklidsche Distanz.

b) Geben Sie das zum Complete-Link-Clustering gehörige Dendrogramm an. Verwenden Sie als zugrundeliegendes Maß die euklidsche Distanz.

c) Geben Sie das zum Group-Average-Clustering gehörige Dendrogramm an. Verwenden Sie als zugrundeliegendes Maß die euklidsche Distanz.