Aufgabenblatt 10

Einführung in die Computerlinguistik (WS19/20) Abgabe bis

Aufgabe 1) Term-Term-Ähnlichkeiten **Punkte:** –

Gegeben sind sind die folgenden Kookurrenzen aus dem BNC-Corpus:

	species	motion	area	environment	star	idea	heat
animal	616	15	230	266	53	138	41
atmosphere	8	7	84	39	19	46	93
carnivore	21	0	3	0	0	4	0

a) Kosinusähnlichkeit *Punkte:* –

Berechnen Sie die Kosinusähnlichkeit zwischen den Begriffen *animal*, *atmosphere* und *carnivore* indem Sie die Spalteneinträge als Featuredimensionen behandeln.

b) PPMI-Umwandlung *Punkte: x*

Beschränken wir uns nun zu Veranschaulichungszwecken auf einen kleineren Ausschnitt der Matrix:

	species	motion	area
animal	616	15	230
atmosphere	8	7	84

Wandeln Sie die Kookurrenzen in PPMI-Einträge um.

Aufgabe 2) Euklidische und Cosinus-Distanz für Einheitsvektoren **Punkte:** –

Zeigen Sie, dass für beliebige Einheitsvektoren $\overrightarrow{p}$, $\overrightarrow{v}$, $\overrightarrow{w}$ gilt:

$$d_{euklid}(\overrightarrow{p}, \overrightarrow{v}) \leq d_{euklid}(\overrightarrow{p}, \overrightarrow{w}) \Leftrightarrow d_{cos}(\overrightarrow{p}, \overrightarrow{v}) \leq d_{cos}(\overrightarrow{p}, \overrightarrow{w}).$$

Hierbei ist $d_{cos}(\overrightarrow{x}, \overrightarrow{y}) := 1 - sim_{cos}(\overrightarrow{x}, \overrightarrow{y})$ für alle Vektoren $\overrightarrow{x}, \overrightarrow{y}$ definiert.[1] Einheitsvektoren sind Vektoren der Länge 1.

[1] Dies ist im allgemeinen keine Metrik, da u.a. die Dreiecksungleichung nicht gilt. Oft wird dies stattdessen eine *dissimilarity* genannt.

Aufgabe 3) Pointwise Mutual Information **Punkte:** –

Aus Suchanfragen von Unigrammen und Bigrammen einer bekannten Online-Suchmaschine lässt sich folgende Matrix erstellen:

	Emily	Charlotte	...	
Brontë	2.620	3.120	...	25.100
Dickinson	12.000	18	...	92.100
...	...	...	...	...
	604.000	733.000	...	100.000.000

Emily liefert also 604.000 Ergebnisse, während das Bigram *Emily Dickinson* 12.000 mal gefunden wurde. Insgesamt wurden 100 Millionen Suchergebnisse ausgewertet.

1. Berechnen Sie Pointwise Mutual Information für *Emily Dickinson*, *Emily Brontë* und *Charlotte Dickinson*.

2. Deckt sich das Ergebnis mit ihrer Intuition?

3. Angenommen der Gebirgszug *Ephel Dúath* aus Herr der Ringe taucht in unseren Suchergebnissen einmal auf und insbesondere treten *Ephel* und *Dúath* jeweils nur einmal und nur zusammen auf. Berechnen Sie PMI für *Ephel Dúath*. Wie verändert sich der Wert für ein größeres N, d.h wenn wir mehr Ergebnisse auswerten, aber keine weiteren Vorkommen der beiden Wörter finden? Halten Sie diese Tendenz für gerechtfertigt?

Aufgabe 4) Agglomeratives Clustering **Punkte:** –

Gegeben seien die folgenden Daten[2]:

ID	Stern	scheinbare Helligkeit	absolute Helligkeit
0	Proxima Centauri	11	16
1	α Centauri A	0	4
2	α Centauri B	1	6
3	Barnards Pfeilstern	10	13
4	Wolf 359	14	17
5	Lalande 21185	7	10
6	α Canis Majoris A	-1	1
7	α Canis Majoris B	8	11

Wir nutzen für diese Aufgabe die Euklidische Distanz.

Um den Rechenaufwand zu minimieren bzw. Programmierung zu vermeiden, ist Ihnen die symmetrische Distanzmatrix vorgegeben.

	0	1	2	3	4	5	6	7
0	0.00	16.28	14.14	3.16	3.16	7.21	19.21	5.83
1	16.28	0.00	2.24	13.45	19.10	9.22	3.16	10.63
2	14.14	2.24	0.00	11.40	17.03	7.21	5.39	8.60
3	3.16	13.45	11.40	0.00	5.66	4.24	16.28	2.83
4	3.16	19.10	17.03	5.66	0.00	9.90	21.93	8.49
5	7.21	9.22	7.21	4.24	9.90	0.00	12.04	1.41
6	19.21	3.16	5.39	16.28	21.93	12.04	0.00	13.45
7	5.83	10.63	8.60	2.83	8.49	1.41	13.45	0.00

Geben Sie die Reihenfolge der Zusammenführung der einzelnen Cluster, sowie die resultierenden Dendrogramme für folgende Gruppenähnlichkeitsmaße an:

1. Single Link
2. Complete Link
3. Average Link

Welche typischen Eigenschaften von complete link vs. single link clustering-Methoden kann man erkennen?

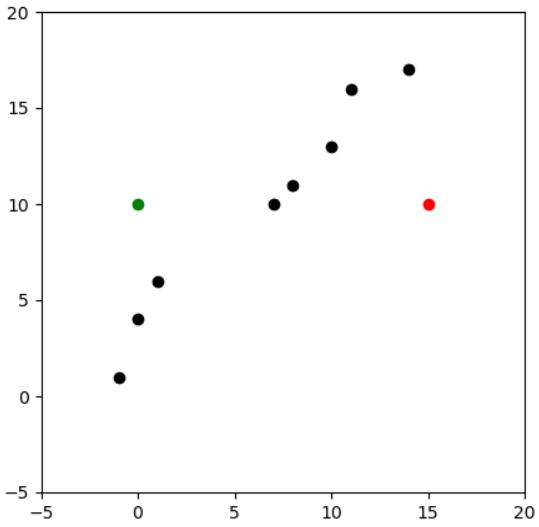
Wenn man nur eine Dimension (zum Beispiel die scheinbare Helligkeit) zum Clustern verwenden würde, würde sich dann in diesem Beispiel viel am Clustering ändern?

[2]Entnommen der Liste der nächsten extrasolaren Sternensysteme

Aufgabe 5) k-means Clustering zum Rechnen **Punkte:** –

Berechnen Sie nun 2-means Clustering bis zur Konvergenz für die Tabelle der Sternensysteme aus der vorherigen Aufgabe 4. Die Zentroiden sind wie folgt initiiert:

Zentroid	Farbe	x	y
c1	rot	15	10
c2	grün	0	10



Geben Sie für jeden Zuweisungsschritt die Koordinaten der zwei Zentroiden, sowie die Menge der Punkte für jeden Cluster an.

Inwieweit hängt die Lösung mit den Lösungen beim agglomerativem Clustering der vorherigen Aufgabe zusammen?

Aufgabe 6) Clustering zum Denken **Punkte:** –

- Gegeben seien Goldcluster $G_1 \dots G_n$ und ein erstelltes Clustering $C_1 \dots C_n$. Das Clustering sei perfekt, d.h. wir haben es geschafft, die Goldcluster genau zu reproduzieren. Wie hoch ist die Reinheit aller Cluster? Eine zweite Methode erstellt $2n$ Cluster, wobei jedes Cluster C_i zufällig in zwei geteilt wurde. Wie hoch ist die Reinheit dieser Cluster? Leiten Sie daraus ein Problem der Benutzung von Reinheit als Evaluationsmaß ab. Wie vermeiden NMI und Rand-Index dieses Problem?

- Überlegen Sie sich ein Beispiel für 3-means clustering (Punkte und Anfangszentroiden), bei dem der Algorithmus einen leeren Cluster generiert.

- Gehen Sie auf die Demo bei `http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html` und versuchen Sie den k-means-Algorithmus zu überlisten, d.h. konstruieren Sie ein Beispiel, so dass keine Konvergenz zum globalen Optimum erfolgen kann.