

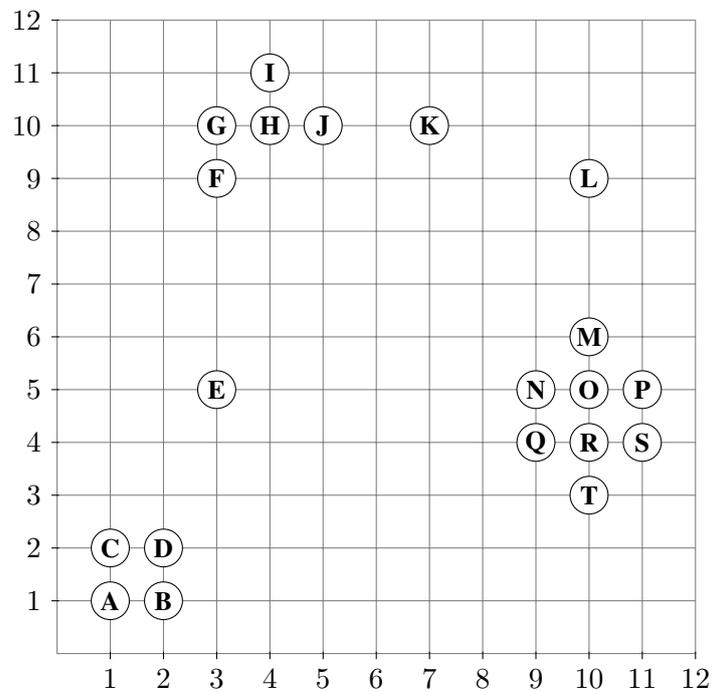
Managing Massive Multiplayer Online Games
SS 2013

Übungsblatt 7: Clustering / Levenshtein / Suffixbäume / Markov-Modelle

Besprechung: 04./05.07.2013

Aufgabe 7-1 DBSCAN

Gegeben sei folgender Datensatz:



Clustern Sie den oben dargestellten Datensatz mit Hilfe des DBSCAN-Algorithmus. Verwenden Sie als Distanzfunktion die L_1 -Norm (Manhattan-Distanz)

$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2|.$$

Verwenden Sie dabei die Parameter $\varepsilon = 1.1$ und $minPts = 3$.

Aufgabe 7-2 *Suffix Bäume*

Gegeben sei das Alphabet $A = \{A, B, E, N, S\}$.

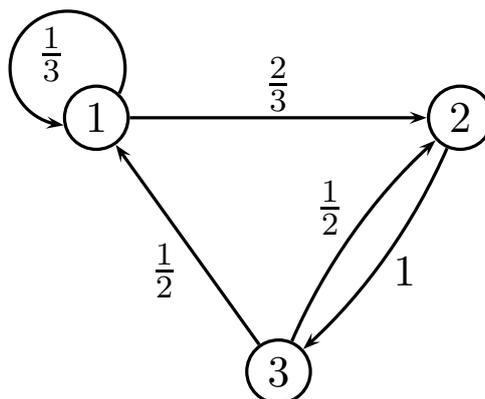
- (a) Fügen Sie die Sequenz $G_1 = \{B, A, N, A, N, E\}$ in einen leeren Suffix Baum ST ein.
- (b) Fügen Sie zusätzlich die Sequenz $G_2 = \{A, N, A, N, A, S\}$ in ST ein.
- (c) Finden Sie die Subsequenz $S_1 = \{N, A, N, A\}$. Welche Sequenz enthält sie?
- (d) Welches ist die längste gemeinsame Subsequenz von G_1 und G_2 ?
- (e) Welche Erweiterung wäre notwendig, um das Finden der häufigsten Subsequenz der Länge n oder länger zu unterstützen?

Aufgabe 7-3 *Levenshtein Distanz*

Bestimmen Sie die Levenshtein Distanz der Sequenzen $BANANE$ und $ANANAS$.

Aufgabe 7-4 *Markov Ketten*

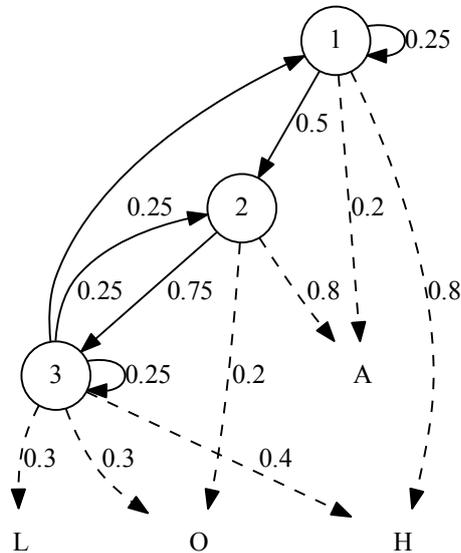
Gegeben sei folgende Markov Kette M in Graph-Darstellung. Dabei stehen Knoten für Zustände, und Kanten für mögliche Übergänge. Kantenlabels entsprechen den Übergangswahrscheinlichkeiten.



- (a) Geben Sie M in Matrix Schreibweise an. Nehmen Sie dabei an, dass die Startzustände gleichverteilt sind, und Sequenzen nur nach Zustand 3 enden, und zwar mit einer Wahrscheinlichkeit von 50%.
- (b) Wie hoch ist die Wahrscheinlichkeit, die Sequenz $3 - 1 - 1 - 2 - 3$ zu beobachten?
- (c) Wie hoch ist die Wahrscheinlichkeit, die Sequenz $2 - 3 - 2 - 1 - 2$ zu beobachten?

Aufgabe 7-5 *Hidden Markow Modelle*

Gegeben sei folgendes Hidden Markow Modell:



- Geben Sie die Zustandsmenge A und die Beobachtungsmenge B an. Leiten Sie die Übergangsmatrix D und die Output-Matrix F aus dem Modell ab. Nehmen Sie an, dass die Startwahrscheinlichkeiten gleichverteilt sind und die Wahrscheinlichkeit, dass Sequenzen in einem Zustand enden, dem zur Summe 1 fehlenden Werten entsprechen.
- Berechnen Sie die Wahrscheinlichkeit, dass die Beobachtung $O_1 = \{H, A, L, L, O\}$ durch das HMM generiert wird.
- Welche Sequenz (s_1, s_2, \dots, s_k) mit $s_i \in A$ erklärt die Beobachtung $O_2 = \{A, L, O, H, A\}$ am besten?