Projektgruppe

Jennifer Post

Clustering und Fingerprinting zur Erkennung von Ähnlichkeiten

2. Juni 2010

Motivation

- Immer mehr Internet-Seiten

  → Immer mehr digitale Texte

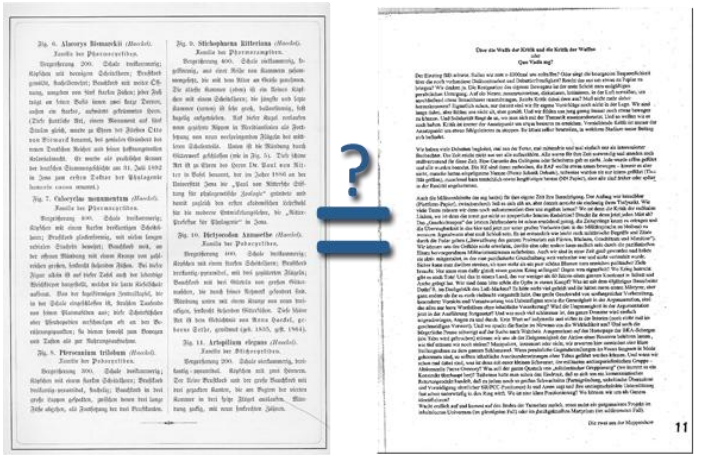
- Viele Inhalte ähnlich oder gleich

- Effektive und effiziente Methoden zur Organisation sind notwendig

Wann sind Texte ähnlich?

Ähnliche…

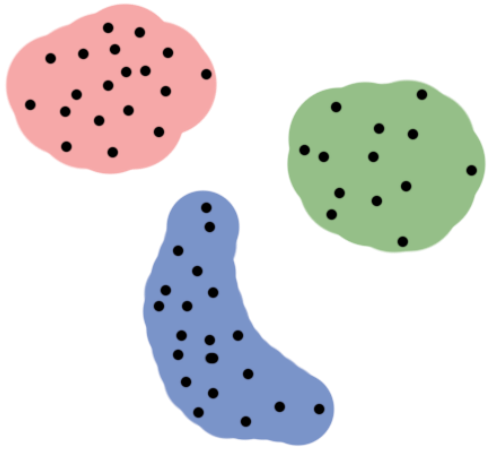
- Thematik
- Satzbau
- Dokument-Struktur



Clustering – Was ist das?

- Aufteilung einer Menge von Objekten

  in disjunkte Teilmengen (=Cluster)



- Möglichst große Ähnlichkeiten von

  Objekten innerhalb eines Clusters

- Möglichst geringe Ähnlichkeit zwischen Objekten verschiedener Clustern

Wie werden Texte repräsentiert?

Bag-of-Words-Modell:

- Nur Worte entscheidend
- Andere Informationen (Satzbau, Interpunktion) werden ignoriert

John likes to watch movies . He also likes to watch football games.

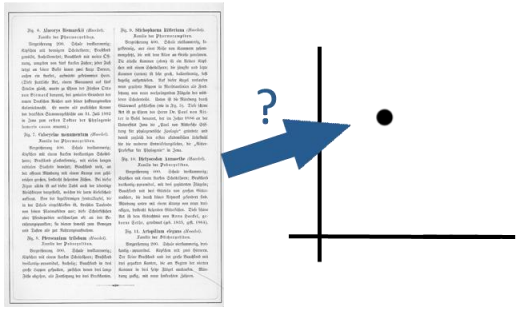
Sandra likes to watch movies too.

{„John“, „likes“, „to“, „watch“, „movies“, „he“, „also“, „football“, „games“, „Sandra“, „too“}

Wie werden Texte repräsentiert?

Dokumentvektor

- Komponenten entsprechen Dimensionen
- Häufigkeit der Komponenten entspricht dem Wert



John likes to watch movies .
He also likes to watch football games.

Sandra likes to watch movies too.

{„John“, „likes“, „to“, „watch“, „movies“, „he“, „also“, „football“, „games“, „Sandra“, „too“}

(1, 2, 2, 2, 1, 1, 1, 1, 1, 0, 0)

(0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1)

TF/IDF

- Termfrequenz allein ist nicht aussagekräftig

- Terme, die in vielen Dokumenten vorkommen, sind nicht so relevant wie Terme, die nur in wenigen Dokumenten vorkommen

- Termfrequenz / Inverse Dokumentfrequenz als Gewichtung für Termfrequenzen

TF/IDF (2)

- Termfrequenz gibt Hinweis auf Bedeutung des Terms innerhalb eines Dokuments:

$$tf_{i,j} = \frac{freq_{i,j}}{max_l(freq_{l,j})}$$

- Inverse Dokumenthäufigkeit misst allgemeine Bedeutung des Terms für alle betrachteten Dokumente:

$$idf_i = \log \frac{N}{n_i}$$

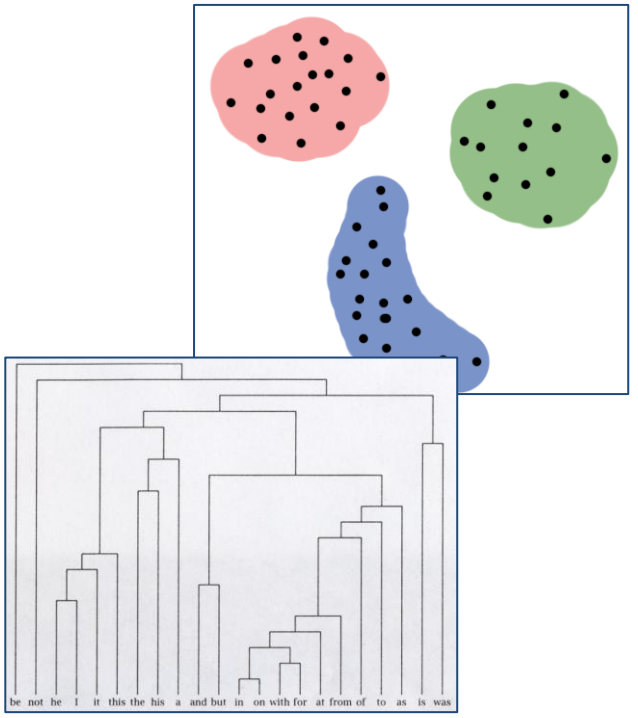
- Gewicht des Terms i in Dokument j nach TD/IDF:

$$w_{i,j} = tf_{i,j} \cdot idf_i = \frac{freq_{i,j}}{max_l(freq_{l,j})} \cdot \log \frac{N}{n_i}$$

Clustering-Algorithmen

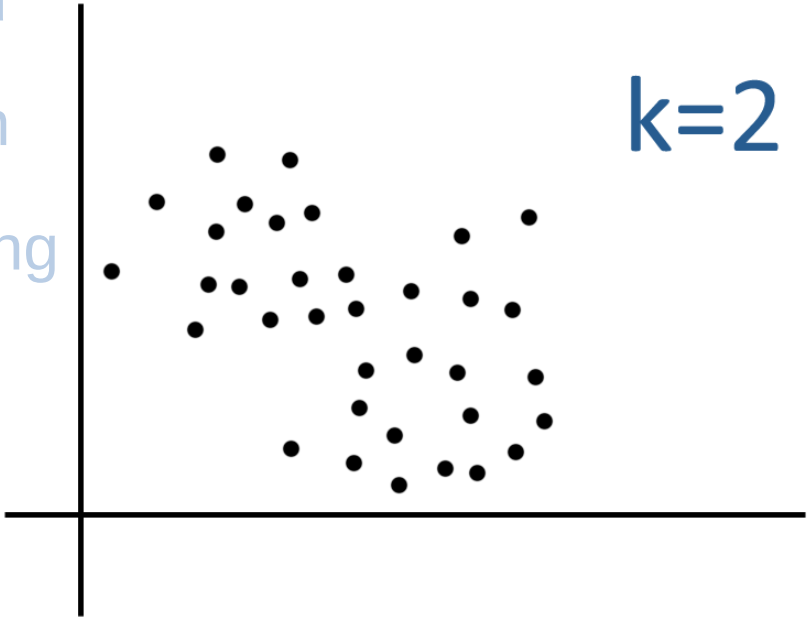
- Ziel ist eine gute Gruppierung der Objekte
- Gruppen sind anfangs unbekannt

  → Unüberwachtes Lernen
- Verschiedene Ansätze
  - Iterativ
  - Hierarchisch
  - Dichtebasiert



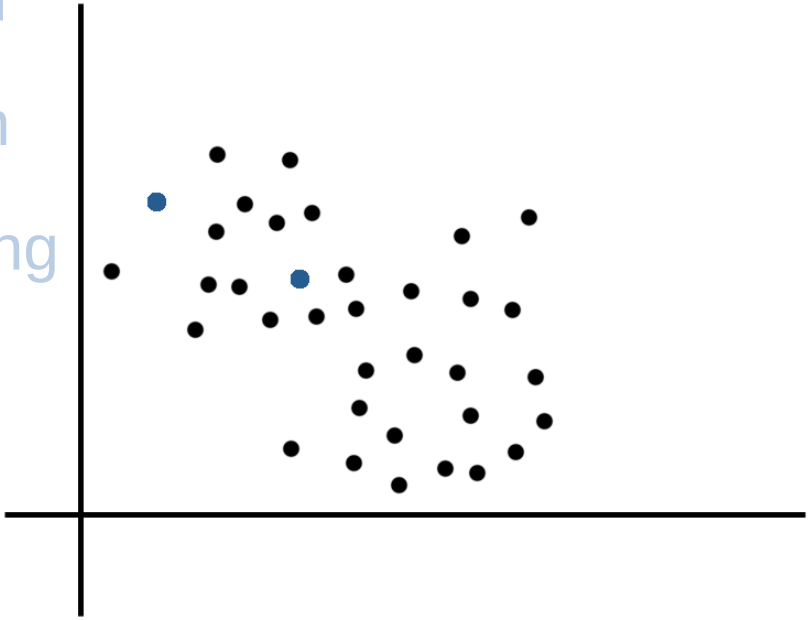
K-Means

- wähle Anzahl der Cluster: k
- wähle k zufällige Centroiden der Cluster
- weise jeden Punkt dem nächsten Centroiden zu
- berechne die Centroiden neu
- wiederhole die letzten beiden Schritte bis sich die Zuweisung nicht mehr ändert



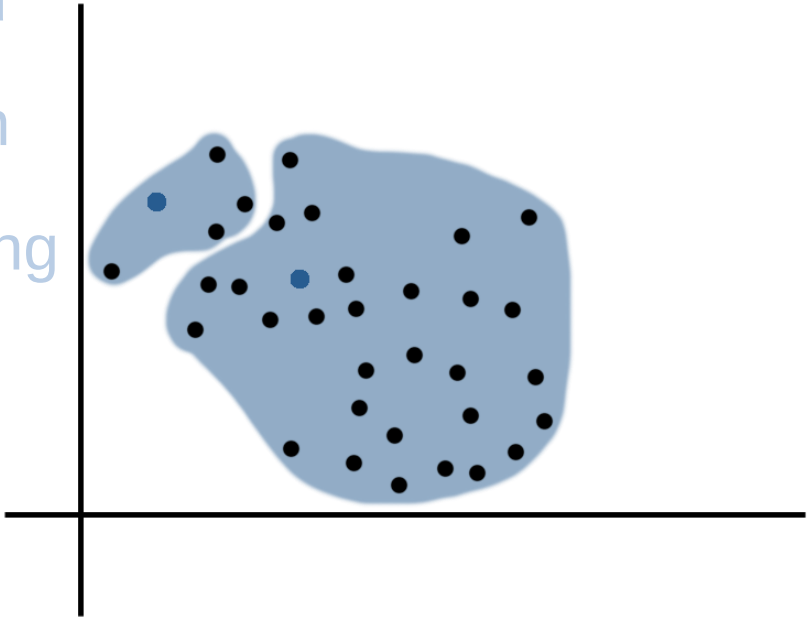
K-Means

- wähle Anzahl der Cluster: k
- wähle k zufällige Centroiden der Cluster
- weise jeden Punkt dem nächsten Centroiden zu
- berechne die Centroiden neu
- wiederhole die letzten beiden Schritte bis sich die Zuweisung nicht mehr ändert



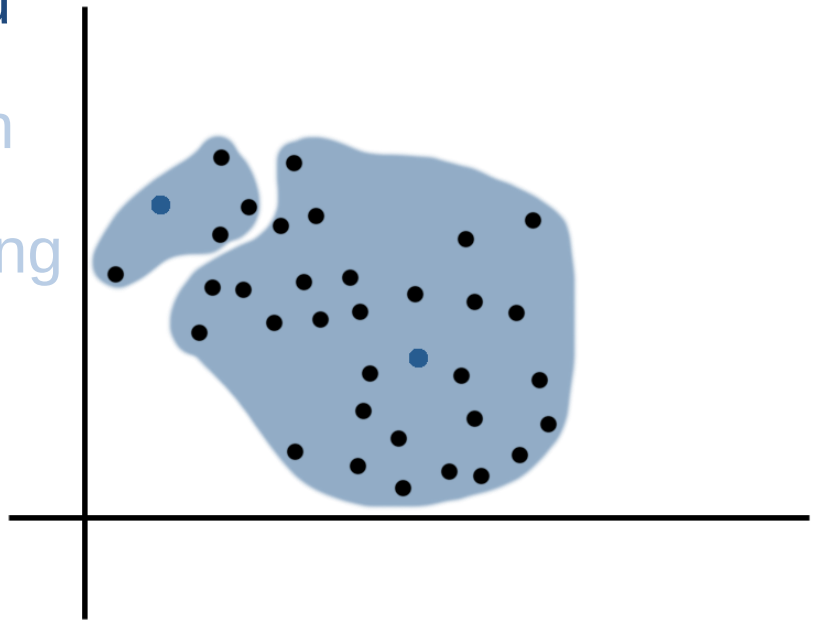
K-Means

- wähle Anzahl der Cluster: k
- wähle k zufällige Centroiden der Cluster
- weise jeden Punkt dem nächsten Centroiden zu
- berechne die Centroiden neu
- wiederhole die letzten beiden Schritte bis sich die Zuweisung nicht mehr ändert



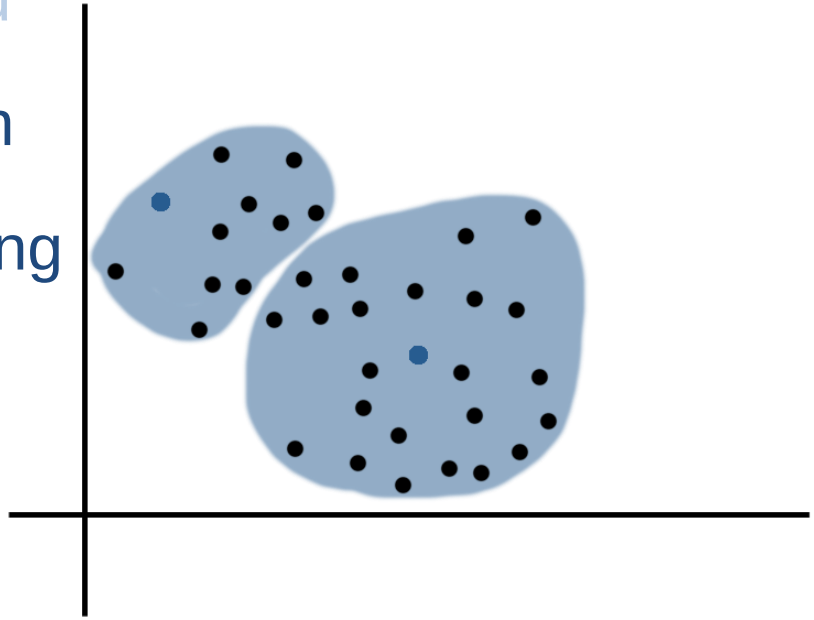
K-Means

- wähle Anzahl der Cluster: k
- wähle k zufällige Centroiden der Cluster
- weise jeden Punkt dem nächsten Centroiden zu
- berechne die Centroiden neu
- wiederhole die letzten beiden Schritte bis sich die Zuweisung nicht mehr ändert



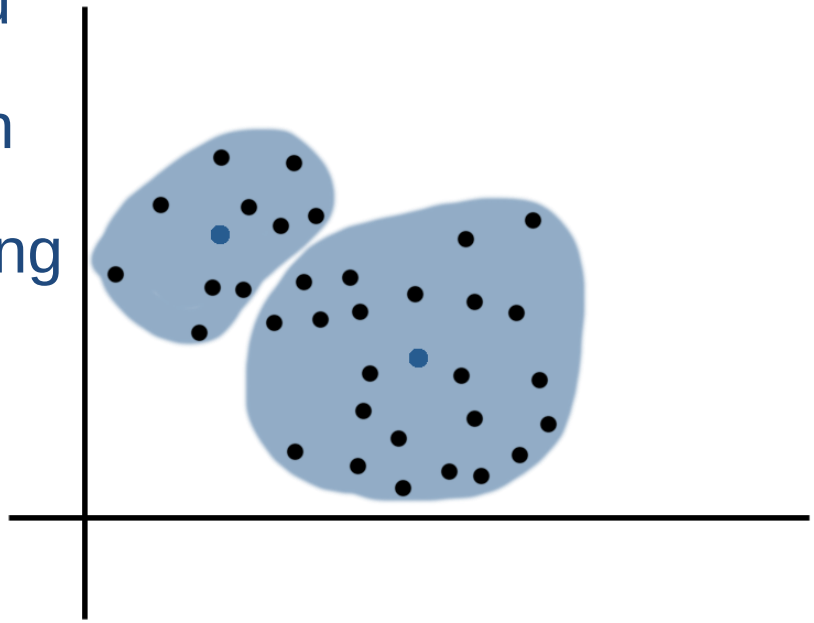
K-Means

- wähle Anzahl der Cluster: k
- wähle k zufällige Centroiden der Cluster
- weise jeden Punkt dem nächsten Centroiden zu
- berechne die Centroiden neu
- wiederhole die letzten beiden Schritte bis sich die Zuweisung nicht mehr ändert



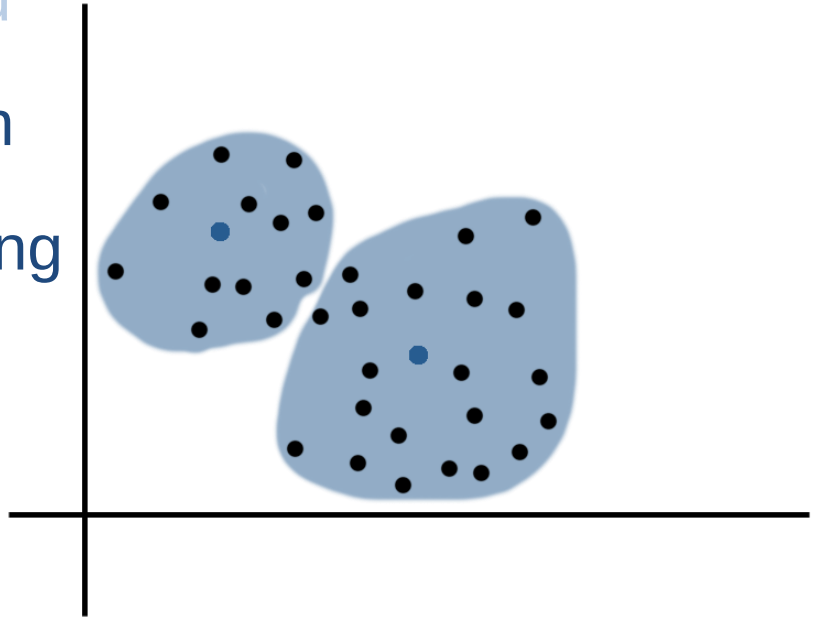
K-Means

- wähle Anzahl der Cluster: k
- wähle k zufällige Centroiden der Cluster
- weise jeden Punkt dem nächsten Centroiden zu
- berechne die Centroiden neu
- wiederhole die letzten beiden Schritte bis sich die Zuweisung nicht mehr ändert



K-Means

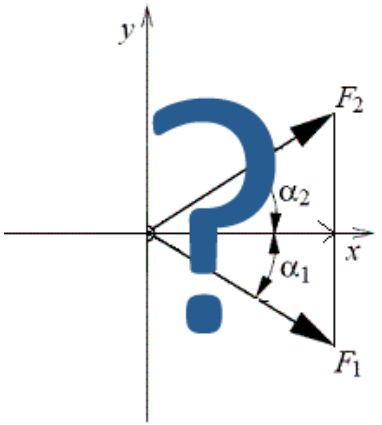
- wähle Anzahl der Cluster: k
- wähle k zufällige Centroiden der Cluster
- weise jeden Punkt dem nächsten Centroiden zu
- berechne die Centroiden neu
- wiederhole die letzten beiden Schritte bis sich die Zuweisung nicht mehr ändert



Ähnlichkeitsmaße

- Spiegeln die Nähe bzw. Distanz zwischen zwei Dokumenten wieder

- Für unterschiedliche Clustering-Probleme sind unterschiedliche Ähnlichkeitsmaße sinnvoll



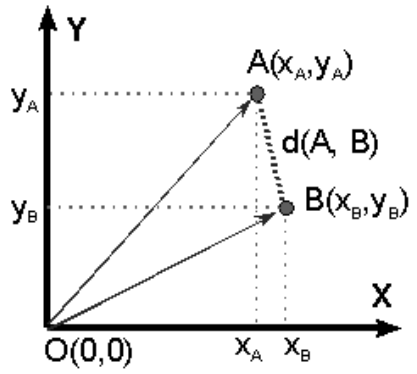
- Tragen entscheidend zur Clusterqualität bei

Euklidische Distanz

- Standardmaß für geometrische Abstände

$$d(x,y) = \|x-y\|_2 = \sqrt{(x_1-y_1)^2+\cdots+(x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

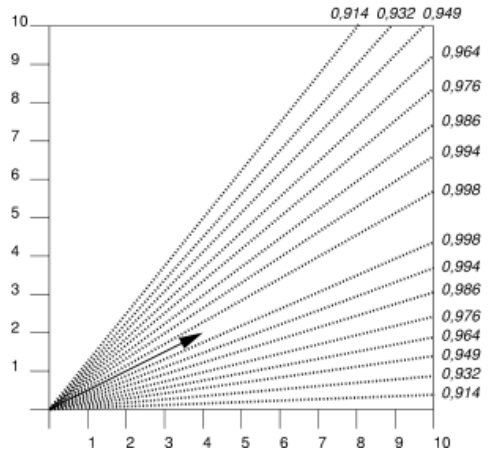
- Länge der Strecke zwischen zwei Punkten



Cosinus-Maß

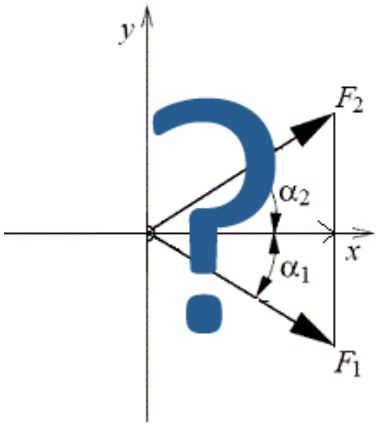
- Basiert auf dem Winkel zwischen den Dokumentvektoren

- Unabhängig von Dokumentlänge



Weitere Ähnlichkeitsmaße

- Es existieren viele weitere Ähnlichkeitsmaße
  - Jaccard-Koeffizient
  - Averaged Kullback-Leibler Divergenz
  - Pearson-Korrelation



- Je nach Anwendungsgebiet muss passendes Ähnlichkeitsmaß gewählt werden

K-Means Vor- und Nachteile

- Schnelle Laufzeit: O(n)
- Allerdings immer Vorgabe von Start-Centroiden & Clusteranzahl notwendig
- Schlechte Wahl von Start-Centroiden kann die Clusterqualität negativ beeinflussen
- Wahl des Ähnlichkeitsmaßes beeinflusst Clusterqualität entscheidend

Hierarchisch agglomeratives Clustering

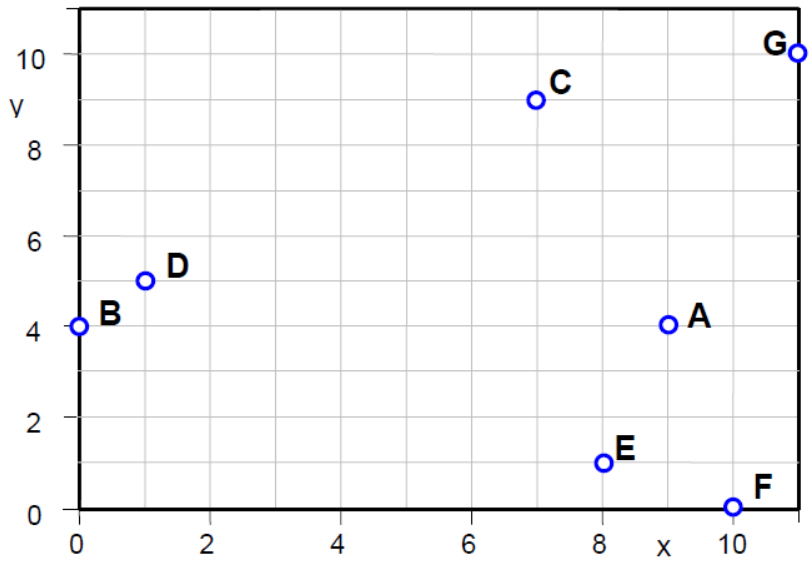
- Hierarchisch strukturiertes Ergebnis

  → informativer als unstrukturierte Cluster von K-Means

- Bottom-up Clustering: angefangen mit ein-elementigen Clustern werden diese nach und nach zusammengefasst

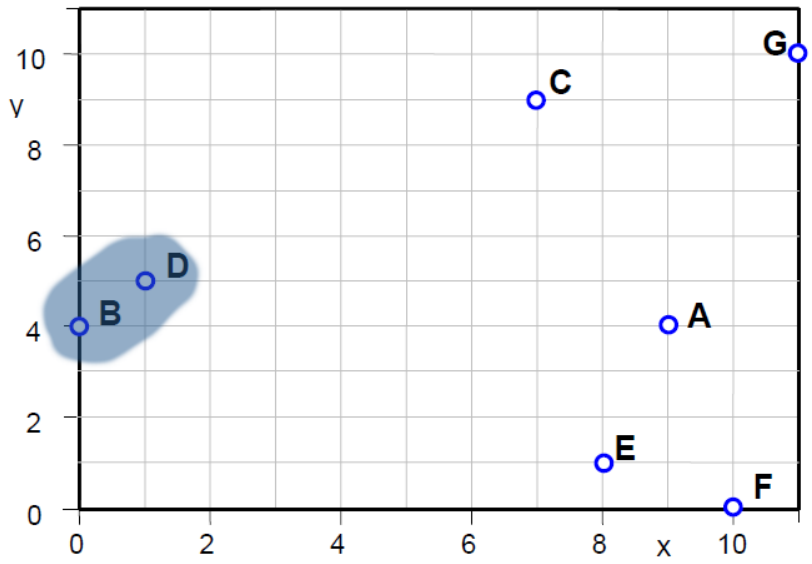
- Allgemeiner Algorithmus:
  - wiederhole, bis nur noch ein Cluster übrig:
    - finde zwei Cluster mit größter Ähnlichkeit
    - Vereine beide Cluster
    - Berechne Ähnlichkeiten neu

HAC Beispiel

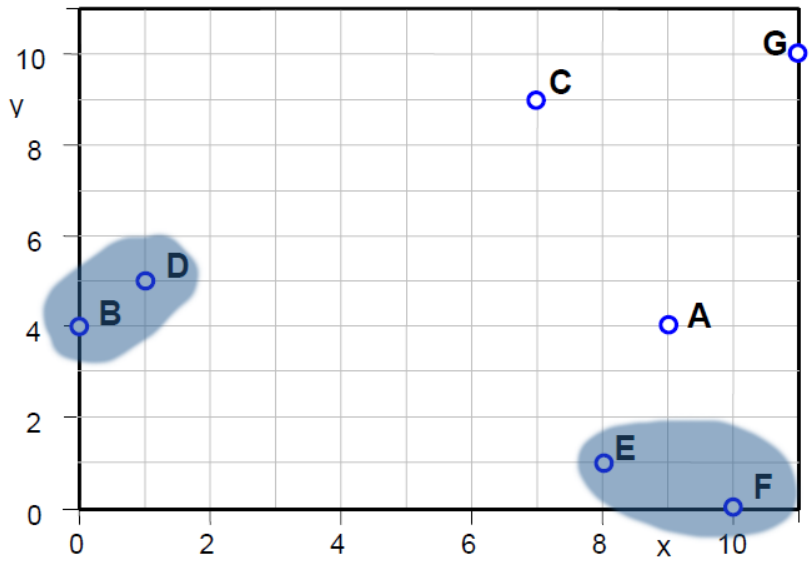


A B C D E F G

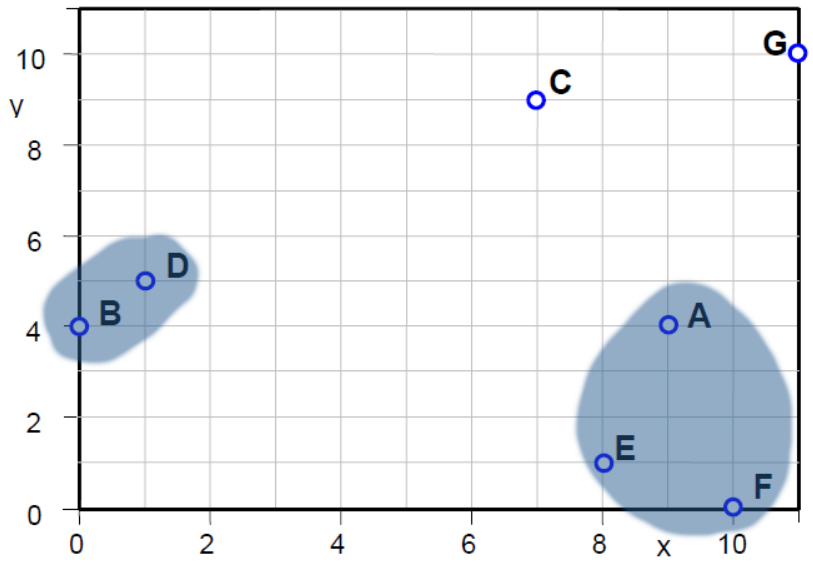
HAC Beispiel

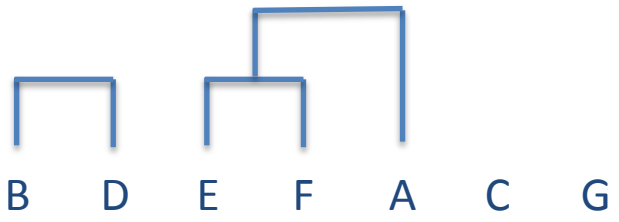


HAC Beispiel

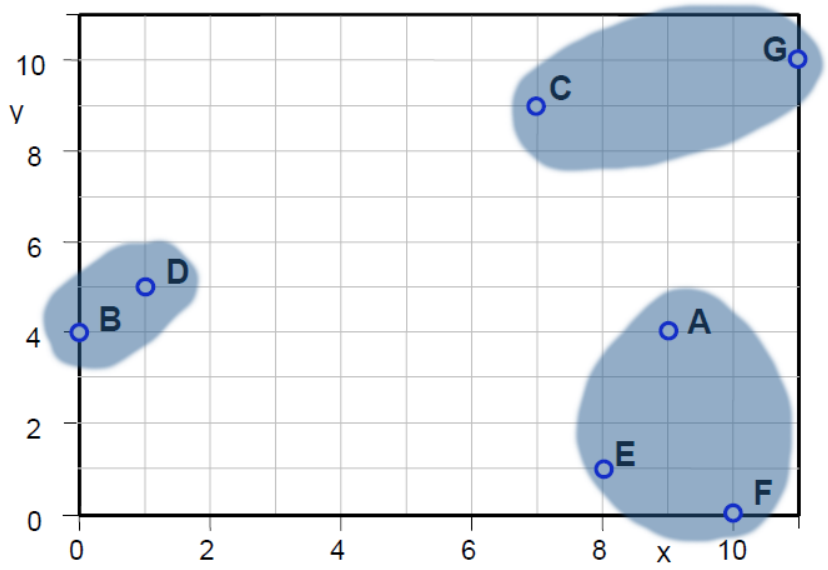


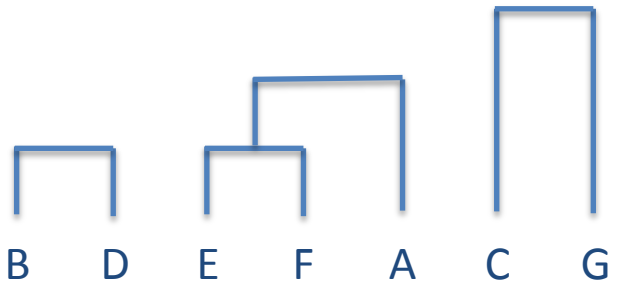
HAC Beispiel



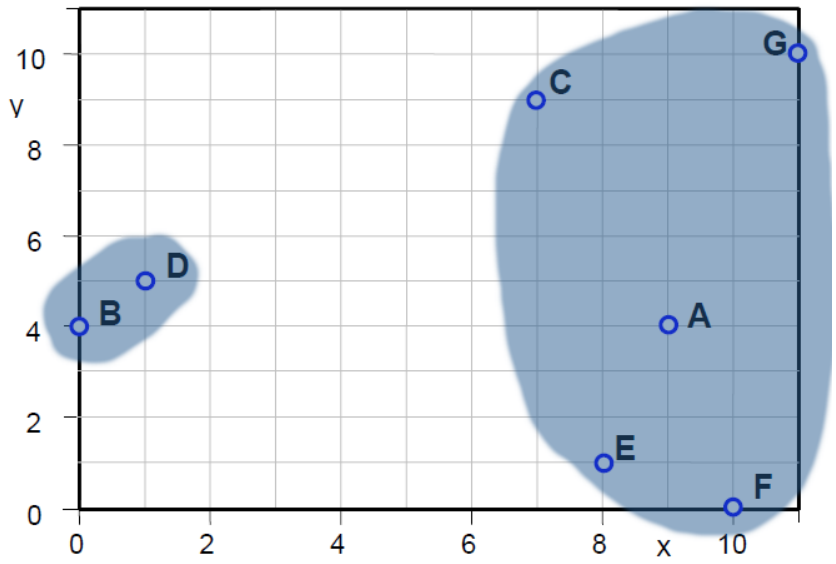


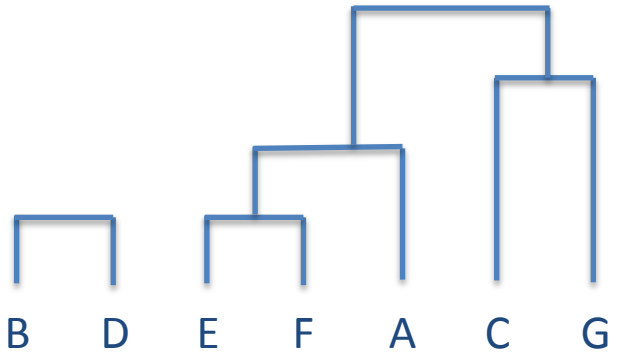
HAC Beispiel



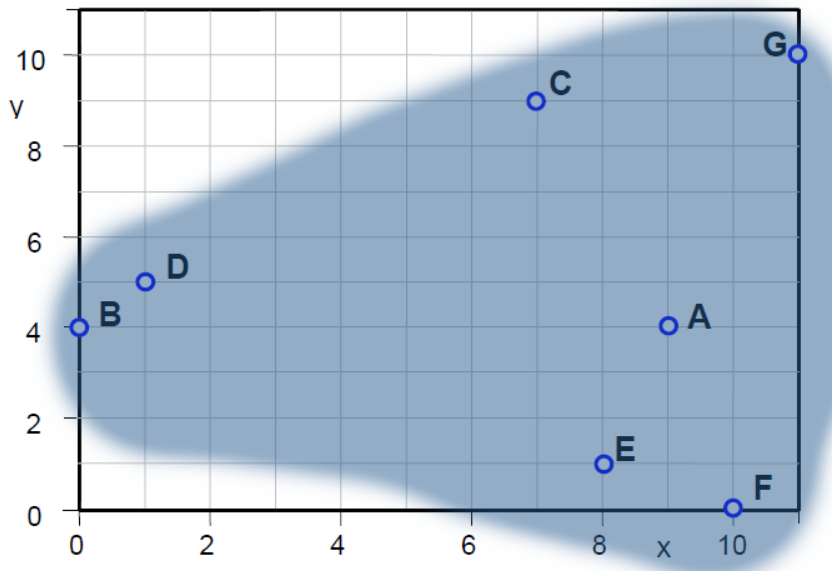


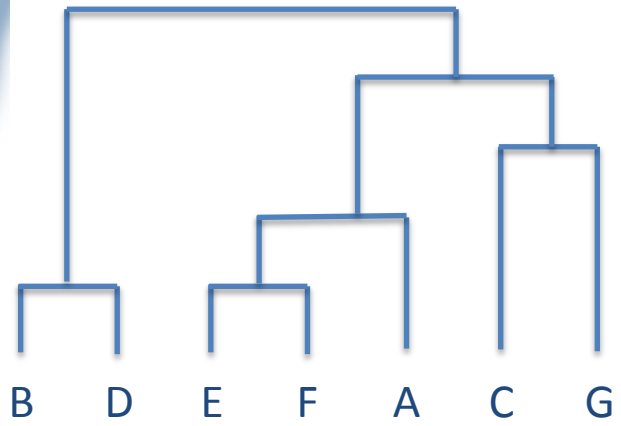
HAC Beispiel





HAC Beispiel





Distanzfunktionen

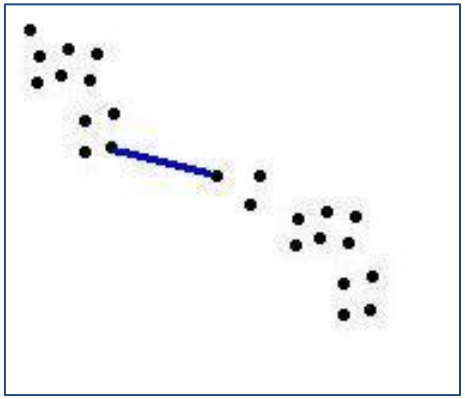
- Es existieren verschiedene Möglichkeiten, die Distanz zwischen zwei Clustern zu berechnen

- Unterschiedliche Distanzfunktionen liefern unterschiedliche Clusterergebnisse

- Distanzfunktion muss je nach gewünschtem Ergebnis gewählt werden

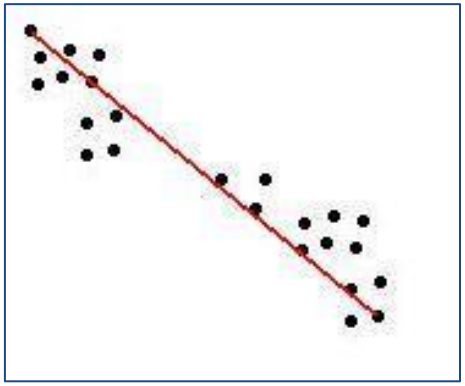
Single Linkage

- Ähnlichkeit ist maximale Ähnlichkeit zwischen zwei beliebigen Objekten aus den Clustern
- Liefert:
  - Große Cluster
  - Schwach ähnliche Elemente
  - Nicht alle Elemente sind sich ähnlich
  - Bildet auch „langgestreckte“ Cluster



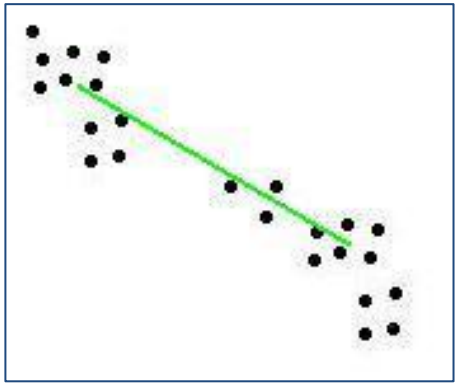
Complete Linkage

- Ähnlichkeit ist minimale Ähnlichkeit zwischen zwei beliebigen Objekten aus den Clustern
- Liefert
  - Kleine Cluster
  - Sehr ähnliche Elemente



Average Linkage

- Durchschnittsähnlichkeit aller Objekte im Cluster
- Liefert
  - Kleine Cluster
  - Sehr ähnliche Elemente



HAC Vor- und Nachteile

- Erzeugt baumartige Hierarchien

- Geringer Rechenaufwand, da bei jedem Schritt die Menge der zu vergleichenden Cluster kleiner wird

- Laufzeit $O(n^2)$

Wofür ist Clustering nicht geeignet?

- Cluster sind nicht gelabelt

- Ähnlichkeitsstrukturen im Textaufbau werden nicht erkannt
  - Plagiate werden nicht erkannt

Was ist ein Fingerprint?

- Kompakte Beschreibung eines Objekts

- Repräsentiert die Eigenschaften bzw. den Inhalt eines Dokuments

Wie werden Fingerprints gebildet?

- Auswahl von Teil-Strings aus dem Text

  Paderborn ist toll

- Anwendung einer Hash-Funktion auf die Teil-Strings

  Paderborn ist toll

  ↓

  9

- Jeder Teil-String ergibt eine Minutia
  - Paderborn → 9, ist → 2, toll → 4
- Fingerprint ist Sammlung von Minutiae
  - FP = (9, 2, 4)

Wie wird Ähnlichkeit gemessen?

- Minutiae repräsentieren Inhalt des Textes
- Ähnlichkeit gemessen durch Vergleich der Minutiae
- Viele gleiche Minutiae → ähnliche Texte

Qualität des Fingerprints

- Funktion, mit der Minutiae gebildet werden
- Granularität – Größe der gewählten Teil-Strings
- Auflösung – Anzahl der verwendeten Minutiae
- Auswahlkriterien für die Teil-Strings

Bildung der Minutiae

- Funktion zur Fingerprint-Berechnung trägt zur Qualität des Fingerprints bei

- Oft Fingerprinting nach Rabin
  - Basiert auf der Modulo-Operation mit zufälligen nicht reduzierbaren Polynomen

- Lineare Laufzeit, kaum Fehler

Bildung der Minutiae (2)

- Fingerprint für „Paderborn“

- $T = (a_1, a_2, \ldots, a_n)$

  $= (80, 97, 100, 101, 114, 98, 111, 114, 110)$

- $FP(T,r) = (a_1 * r^n + a_2 * r^{n-1} + \ldots + a_{n-1} * r^2 + a_n * r) \bmod m$

  $FP(\text{„Paderborn“}, 3) = (80 * 3^9 + 97 * 3^8 + \ldots + 114 * 3^2 + 110 * 3) \bmod 17 = 9$

Granularität und Auflösung

- Granularität

- Auflösung

Teil-String-Auswahl

- Full-Fingerprinting
  - Jeder Teil-String des Dokuments wird ausgewählt
  - Jeder Minutia ist Teil des Fingerprints
  - Hoher Aufwand

- „Paderborn“ = „Pad“, „ade“, „der“, „erb“, „rbo“, „bor“, „orn“

Teil-String-Auswahl

- Randomisierte Auswahl

  Worst-Case

- All-Substring-Selection

  alle nicht-überlappenden Teils-Strings werden gewählt

  „Paderborn“ = „Pad“, „erb“, „orn“

Teil-String-Auswahl

- First-r-Selection

  die ersten r nicht-überlappenden Teil-Strings

  r= 2 → „Paderborn“ = „Pad“, „erb“

- First-r-Sliding-Selection

  wie First-r, allerdings mit überlappenden Teil-Strings

  r= 2 → „Paderborn“ = „Pad“, „ade“

Teil-String-Auswahl

- Rarest-In-Document-Selection

  Teil-Strings wählen, die die seltensten Minutiae innerhalb des Dokuments bilden

- Rarest-In-Collection-Selection

  Teil-Strings auswählen, die selten in der Dokumentensammlung sind

Clustering vs. Fingerprinting

- Clustering erkennt gleiche Thematiken
- Ähnlichkeiten im Textaufbau werden nicht erkannt
- Strukturen werden erkannt

- Fingerprinting mehr für Erkennung gleicher Texte/Plagiate
- Erkennung sprachlicher Gemeinsamkeiten
- Erkennung von Veränderungen im Text
- Gleiche Thematik wird unter Umständen nicht erkannt

Weitere Anwendung

- Zum Teil Kombination von Fingerprinting und Clustering
- Text -> Fingerprint
- Dann Clustern der Fingerprints

Nutzen für die PG

- Anforderungsdokumente können enthalten
  - Abschnitte mit gleicher Thematik
  - Abschnitte mit gleicher Wortwahl

- Beide Methoden interessant für PG

Vielen Dank für die Aufmerksamkeit!

Fragen?