Textmining

Dokumentähnlichkeit

Folien von Christian Wartena

Merkmale I

- Objekte können oft über Merkmale beschrieben werden.
- Wenn Merkmale numerische Werte haben, können Objekte mit einer Reihe von Zahlen beschrieben werden.
- Eine solche Reihe von Zahlen nennen wir einen *Merkmalsvektor*.
- Die Merkmale haben eine feste Reihenfolge.
- Die erste Zahl steht für den Wert des ersten Merkmals, die zweite für den wert des zweiten Merkmals, usw.

Merkmale II

- Beispiele:
  - Koordinaten eines Objektes
  - Die Cepstrum-Koeffizienten in der Spracherkennung
  - Farbwerte in einem Bild
  - Erkannte Teilmuster in einem Bild
  - Wörter in einem Text
  - "Gelikte" Berichte für eine Person
  - usw.

Merkmalsraum I

- Wenn wir zwei Merkmale haben, z.B. Längengrad und Breitengrad, können wir ein Objekt als Vektor oder Punkt auf einer Fläche darstellen.
- Bei drei Merkmalen brauchen wir einen 3-D Raum.
- Wir können jetzt auch Entfernungen zwischen Objekten berechnen!
- Es gibt verschiedene Möglichkeiten eine Entfernung zu berechnen.
- Einfach einen Abstand mit dem Satz von Pythagoras zu berechnen, funktioniert beispielsweise nicht auf einer gekrümmten Oberfläche, wie die der Erde.
- Wir schauen und drei Abstandsmaße etwas genauer an:
  - Jaccard Koeffizient
  - Manhattan-Distanz

Merkmalsraum II

- Euklidischer Distanz
- Kosinusähnlichkeit

Jaccard Koeffizient

- Jaccard Koeffizient ist ein Maß für die Ähnlichkeit zwischen Mengen
  - Also nicht Vektoren!
- Einfach und schnell zu berechnen.
- ZB geeignet für Mengen von Schlagwörtern

$$jac(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- Wenn A und B die gleichen Elemente enthalten, ist der Wert 1
- Wenn A und B keine gemeinsame Elemente enthalten, ist der Wert 0

Manhattan-distanz I

Figure: Mannheim. Quelle:Deutsches Architektur Forum

Manhattan-distanz II

- Auch genannt: Taxi-Distanz, Cityblock-Distanz oder Mannheimer Distanz.
- Angewandt auf ein 2D Problem: die Entfernung die man in Mannheimer Stadtzentrum laufen müsste. Man darf sich also ausschließlich entlang den Koordinaten bewegen und keine diagonale Bewegungen ausführen.
- Für zwei Vektoren a und b gilt nun:

$$d(a, b) = \Sigma_i |a_i - b_i|$$

Manhattan-distanz III

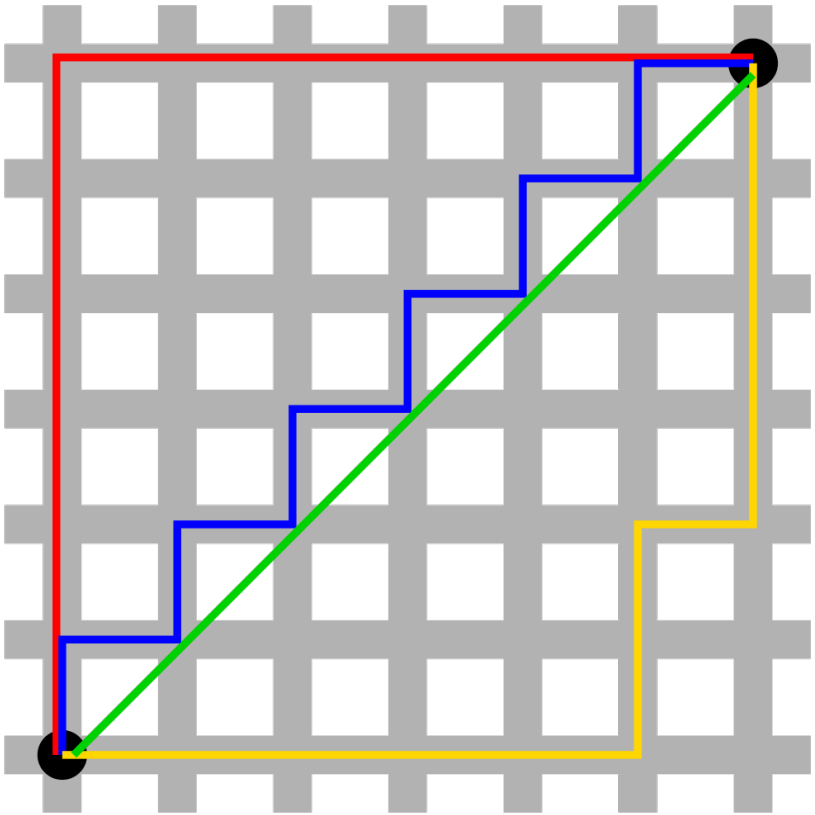


Figure: Manhattan-Distanz zwischen zwei Punkten. Quelle:Wikipedia

Euklidischer Distanz I

- Abstand wird berechnet mit dem Satz von Pythagoras
- Im 2D-Fall:

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

- Im Allgemeinen:

$$d(a,b) = \sqrt{\Sigma_i (a_i - b_i)^2}$$

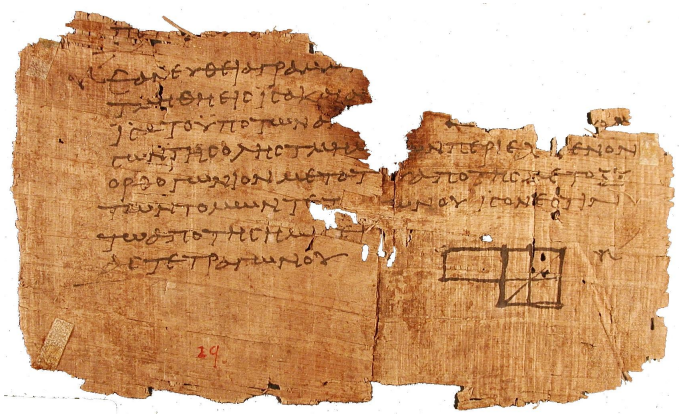
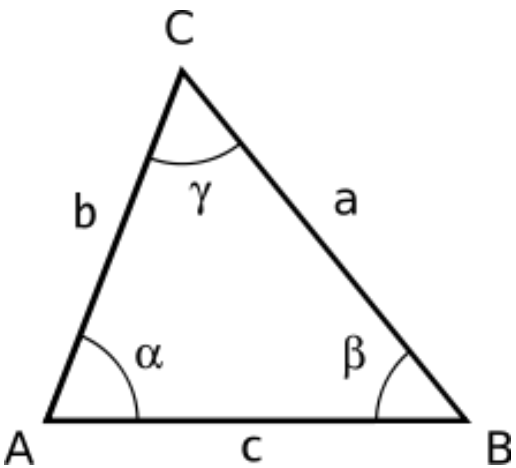


Figure: Kopie eines Werkes von Euklid (3. Jhdt vor Chr.)) aus dem 1. Jhdt. nach Chr. Quelle:Wikipedia

Kosinusähnlichkeit

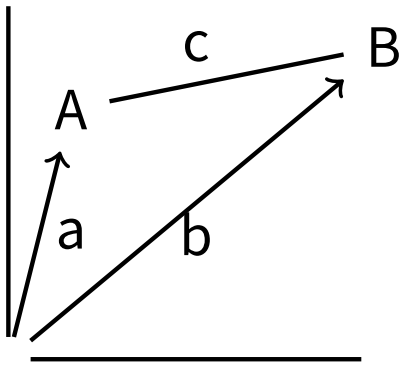
- Manchmal ist die Position der Punkte unwesentlich, sondern geht es um die Richtung der Vektoren.
- Die Richtung kann mit dem *Winkel* zwischen den Vektoren ausgedrückt werden.
- Kosinussatz:

$$\begin{aligned} c^2 &= a^2 + b^2 - 2ab\cos\gamma \\ \Leftrightarrow cos\gamma &= \frac{a^2 + b^2 - c^2}{2ab} \end{aligned}$$



Kosinusähnlichkeit (im 1. Quadrant)

- Die Längen von a, b und c können über den Satz von Pythagoras berechnet werden.



$$
\begin{aligned}
a &= \sqrt{A_x^2 + A_y^2} \\
b &= \sqrt{B_x^2 + B_y^2} \\
c &= \sqrt{(B_x - A_x)^2 + (B_y - A_y)^2}
\end{aligned}
$$

Kosinusähnlichkeit

- Wir kombinieren das jetzt mit dem Kosinussatz:

$$\begin{aligned}\cos(a,b) &= \frac{a^2+b^2-c^2}{2ab} \\ &= \frac{A_xB_x+A_yB_y}{\sqrt{A_x^2+A_y^2}\cdot\sqrt{B_x^2+B_y^2}}\end{aligned}$$

- Im allgemeinen gilt:

$$\cos(a,b) = \frac{\Sigma_i a_i b_i}{\sqrt{\Sigma_i a_i^2}\cdot\sqrt{\Sigma_i b_i^2}}$$

Dreiecksungleichung

- Es gibt viele weitere Möglichkeiten, Ähnlichkeiten zwischen Objekten zu definieren.
- Für die meiste Zwecke brauchen wir keine willkürliche Ähnlichkeiten, sondern einen metrischen Raum.
- Folgende Axiome müssen für beliebige Elemente x, y und z gelten:
  1. $d(x, y) \geq 0$
  2. $d(x, y) = d(y, x)$ (Symmetrie)
  3. $d(x, y) \leq d(x, z) + d(z, y)$ (Dreiecksungleichung)

Visualisierung / Multidimensional Scaling

- Wenn wir erst mal Ähnlichkeiten zwischen Elementen haben, können wir diese dann auch abbilden?
- Ja, aber nicht unbedingt in einem 2-dimensionalen Raum!
- Wir können aber versuchen, die Elemente in einer Fläche abzubilden, und die Abstände dabei möglichst wenig zu verändern.
- Dieses Verfahren heißt: *Multidimensional Scaling (MDS)*

Dokumentähnlichkeit

Maß

- Für die Ähnlichkeit zwischen zwei Dokumenten können alle genannte Ähnlichkeitsmaße benutzt werden.
- Kosinus ist am üblichsten

Darstellung

- Festgelegte Menge von Wörter:
  - Position 1: Wert für Wort 1
  - Position 2: Wert für Wort 2
  - usw.

Wortgewichte

Maß

- 0 oder 1: Wort kommt vor oder nicht
- Anzahl der Vorkommen des Wortes
- Anzahl der Vorkommen dividiert durch die Gesamtanzahl der Wörter

Problem

- Viele Wörter kommen in allen Texten vor
- Alle Texte der gleichen Sprache sind sich sehr ähnlich
- Lösung: wenn Wörter, die das Thema bestimmen ein höheres Gewicht bekommen, messen wir eher eine thematische Ähnlichkeit

Inverse Dokument Frequenz (IDF)

Inverse Dokument Frequenz (Idee)

- Ein Wort, das in jedem Dokument vorkommt, sagt wenig über das Thema des Dokumentes aus
- Ein Wort, das nur in wenigen Dokumenten vorkommt, ist charakteristisch für das Dokument, in dem es vorkommt.

Inverse Dokument Frequenz (Formel)

- $|D|$ die Zahl der Dokumente in D
- $N_{D,w}$ die Zahl der Dokumente aus D in denen w vorkommt.
- $df_{D,w} = \frac{N_{D,w}}{|D|}$ die Dokumentfrequenz vom w in D
- $idf_{D,w} = \frac{1}{df_{D,w}} = \frac{|D|}{N_{D,w}}$ die **inverse Dokumentfrequenz**
- Der Effekt ist oft zu stark. Deswegen nimmt man oft $idf_{D,w} = 1 + \log \frac{1}{df_{D,w}}$: