

Dokumentdistanz (Vektormodell)

Motivation

Die Dokumentdistanz ist eine nichtnegative Zahl, mit der sich die Ähnlichkeit zweier Dokumente beschreiben lässt. Wozu?

- automatische Klassifikation von Dokumenten
- Document Retrieval (Auffinden von Dokumenten)
- Erkennung von Plagiaten

Das Modell

In einer Sammlung von Dokumenten wird jedes Dokument als Menge von Wörtern aufgefasst. Wird die Häufigkeit der Wörter berücksichtigt, spricht man von einer *Multimenge* (*Bag of Words*).

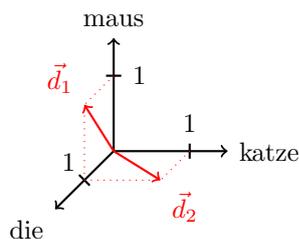
Wählt man eine willkürliche aber feste Reihenfolge in der Menge aller Wörter in der Dokumentsammlung ($\vec{w}_1, \vec{w}_2, \dots, \vec{w}_n$), so lässt sich jedes Dokument d der Kollektion als Vektor

$$\vec{d} = h_1 \cdot \vec{w}_1 + h_2 \cdot \vec{w}_2 + \dots + h_n \cdot \vec{w}_n$$

im Vektorraum der n fest gewählten (Basis)Wörter darstellen. Die skalaren Komponenten h_1, h_2, \dots, h_n stellen die Häufigkeiten der entsprechenden Wörter dar.

Beispiel 1

$\vec{d}_1 =$ „die maus“	Basisvektor	\vec{d}_1	\vec{d}_2
$\vec{d}_2 =$ „die katze“	die	1	1
	maus	1	0
	katze	0	1



Nun können wir den Winkel zwischen den vektorisierten Dokumenten \vec{d}_1 und \vec{d}_2 als Maß für ihre Distanz auffassen. Dafür verwenden wir die aus der Vektorgeometrie bekannte Zwischenwinkelformel:

$$\varphi = \arccos \frac{\vec{d}_1 \cdot \vec{d}_2}{|\vec{d}_1| \cdot |\vec{d}_2|}$$

Beispiel 1 (Fortsetzung)

$$\vec{d}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \text{ („die katze“)} \quad \vec{d}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \text{ („die maus“)}$$

$$\begin{aligned} \varphi &= \arccos \frac{1 \cdot 1 + 1 \cdot 0 + 0 \cdot 1}{\sqrt{1^2 + 1^2 + 0^2} \cdot \sqrt{1^2 + 0^2 + 1^2}} \\ &= \arccos \frac{1}{\sqrt{2} \cdot \sqrt{2}} = \arccos \frac{1}{2} = 60^\circ \end{aligned}$$

Beispiel 2

\vec{d}_1 = „der hund jagt die katze“

\vec{d}_2 = „die katze jagt den hund“

\vec{d}_3 = „die katze jagt die maus“

Basisvektor	\vec{d}_1	\vec{d}_2	\vec{d}_3
der	1	0	0
hund	1	1	0
jagt	1	1	1
die	1	1	2
katze	1	1	1
den	0	1	0
maus	0	0	1

Welche der drei Dokumente haben die kleinste „Distanz“?

$$\text{dist}(\vec{d}_1, \vec{d}_2) = \arccos \frac{4}{\sqrt{5}\sqrt{5}} = 36.87^\circ$$

$$\text{dist}(\vec{d}_2, \vec{d}_3) = \arccos \frac{4}{\sqrt{5}\sqrt{7}} = 47.46^\circ$$

$$\text{dist}(\vec{d}_3, \vec{d}_1) = \arccos \frac{4}{\sqrt{5}\sqrt{7}} = 47.46^\circ$$

Bemerkungen

- Es treten nur Winkel zwischen 0° (wortmässige Übereinstimmung) und 90° (disjunkte Wortmengen) auf.
- In bestimmten Situationen kann es sinnvoll sein, Wörter höchstens einfach zu zählen oder Wörter aus den Dokumenten zu entfernen, die keinen Beitrag zu ihrer Charakterisierung leisten (*Stop words*).

Geschichte

Das Vektorraummodell geht auf Gerhard Salton zurück, der es in den 60er Jahren bei der Arbeit am SMART-Projekt (System for the Mechanical Analysis and Retrieval of Text) entwickelt hat (Fuhr, 2006).

Quellen

Demaine, E. (2011). Lecture 2: Models of Computation, Document Distance. 33'–43'.
<https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-006-introduction-to-algorithms-fall-2011/lecture-videos/lecture-2-models-of-computation-document-distance/>
(7.9.2018)

Fuhr, N. (2006). *Information Retrieval*. Skriptum zur Vorlesung im SS 06.
http://www.is.informatik.uni-duisburg.de/courses/ir_ss06/fohlen/irskall.pdf (7.9.2018)