Vector Space Model

Roman Behrends und Sebastian Zaffiro

Repräsentation von Dokumenten als Vektor

- Dokumente werden als Vektor eingezeichnet
- Die Dimensionen sind verschiedene Eigenschaften
  - z. B. Schlüsselwörter
- Eigenschaften müssen nicht gleich gewichtet sein
- Wird vor allem zum Vergleichen von Dokumenten genutzt
  - z. B. wie Relevant sind Google-Ergebnisse

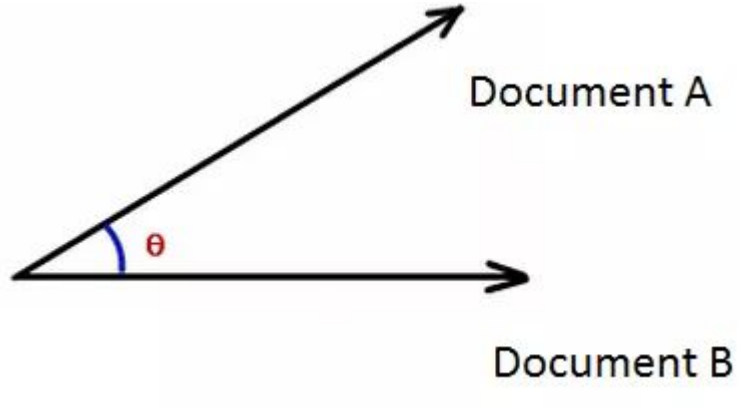
Skalarprodukt, Kosinus Ähnlichkeit

- Je geringer der Winkel zwischen dem Vektor der Suche und dem des Dokumentes desto relevanter ist es.
- Der Winkel ist häufig nicht so einfach zu bestimmen
  ⇒ Der Kosinus wird statt dem Winkel verwendet
- Je höher der Kosinuswert, desto besser passt es.

$$\cos\theta = \frac{\mathbf{d_2} \cdot \mathbf{q}}{\|\mathbf{d_2}\| \, \|\mathbf{q}\|}$$

Queries beantworten mit Vector Space Model

- Queries, also Useranfragen werden erst auch in so einen Vektor umgewandelt
- Hiernach wird dieser Vektor meist mithilfe des Kosinus mit den Dateien verglichen.
- Je höher der Kosinuswert, desto besser ist das Ergebnis.



Ranking

- Zweck: Auflistung der Dokumente nach Relevanz für den Nutzer
- Je größer der Cosinus-Wert, desto relevanter ist das Dokument und steht somit weiter oben
- Begriffe, die oft auftreten, aber in wenigen verschiedenen Dokumenten auftauchen, werden als wichtiger bewertet als welche bei denen das Gegenteil der Fall ist

Rechenbeispiel

d_1 = „Nachweis von Chloridionen“

d_2 = „Nachweis von Acetationen“

d_3 = „Salze der Chloridionen“

q = „Nachweis der Acetationen“

⟹ Das zweite Dokument ist als oberstes Ergebnis zu erwarten

Vor- und Nachteile

+ Durch den Kosinuswert gibt es eine Auflistung
+ Auch Dokumente in denen nicht alle Wörter der Suchanfrage sind, tauchen in der Auflistung auf
+ Basiert auf relativ einfacher Mathematik
- Die Position und Reihenfolge der Wörter in den Dokumenten wird nicht beachtet
- Synonyme von Suchwörtern werden auch außer Acht gelassen
- Lange Dokumente können unberechtigterweise eine geringere Relevanz zugeschrieben bekommen

Quellen

http://www.site.uottawa.ca/~diana/csi4107/cosine_tf_idf_example.pdf

https://en.wikipedia.org/wiki/Vector_space_model

http://www.iai.uni-bonn.de/III/lehre/vorlesungen/InformationRetrieval/WS04/Vorlesung-041104neu.pdf

https://en.wikipedia.org/wiki/Bag-of-words_model