Prof. Dr. Philipp Schaer, Timo Breuer, Björn Engelmann
Institut für Informationswissenschaft

Technology
Arts Sciences
TH Köln

Information Retrieval – Übung Vector Space Model und Ranking

I Retrieval mit dem Vektorraummodell

Stellen Sie sich vor, Sie haben einen Dokumentenkorpus, der wie folgt aussieht:

Dokument 1: „traditional french recipe“

Dokument 2: „french dinner“

Dokument 3: „traditional christmas dinner“

Die Suchanfrage q an das Retrievalsystem lautet „traditional french christmas recipe“.

- Erstellen Sie die Term-Dokument-Matrix mit einer **tf-idf-Gewichtung** von **$(1 + \log_{10}(tf_{d,f})) * \log_{10}(N/df_t)$**.
- Berechnen Sie den **Score** nach für jedes Dokument in Abhängigkeit von der Suchanfrage q nach dem Vektorraummodell bzw. der Kosinus-Ähnlichkeit.
- In welcher Reihenfolge werden die Dokumente in der Ergebnisliste angezeigt?

II Dokumentenklassifikation

Sie arbeiten in einem Betrieb in dessen Firmenblog die neusten Posts Kategorien zugeordnet werden. Wie können Sie Vorschläge für Kategorien für einen neuen Blog-Post generieren, wenn Sie als Grundlage für den Vorschlag das Vektorraummodell verwenden?

- Skizzieren Sie Ihren Lösungsweg und die einzelnen Schritte.
- Sie brauchen die einzelnen Schritte nicht zu rechnen, aber Sie sollten sie mit den korrekten Begriffen aus der Vorlesung beschreiben können.

Nachfolgend finden Sie die Termfrequenzen der bisherigen Posts und des neuen Eintrags.

Kategorie / Terme	Innovation	Preis	Kunde	Produkt
Blog-Post A (Produkte)	0	1	4	5
Blog-Post B (Werbung)	5	6	4	0
Blog-Post C (Allgemein)	2	2	3	2
Neuer Blog-Post (?)	?	?	?	?

Beispielsweise enthält hier der Blog-Post A aus der Kategorie Produkte den Term Kunde vier mal und den Term Innovation gar nicht.

- Wie müsste der neue Blog-Post geschrieben sein, damit er auf jeden Fall der Gruppe „Werbung“ zugeordnet werden würde?
- Wie würden Sie vorgehen, wenn es pro Kategorie mehr als einen Blog-Post gibt? Wie würde sich das Verfahren zur Ähnlichkeitsberechnung ändern?