

DIS12: Information Retrieval – Online-Sitzung 11.12.2020

Übung 03: Ranking, Vektorraum und Co.

Bitte bearbeiten Sie die Aufgaben vor den Übungsterminen. In den Online-Sitzungen arbeiten wir dann mit den Ergebnissen Ihrer Vorarbeiten. Z.B. stellt jemand aus Ihren Reihen die vorbereitete Lösung vor. Die Lösung kann aber auch im Plenum diskutiert, erweitert und in Kontext gesetzt werden. Auch Peer-Reviews sind möglich.

Die Präsentation der Lösung kann beispielsweise über "Teilen des Bildschirms", Kurzreferate oder interaktive Kollaborationswerkzeuge wie Google Jamboards erfolgen.

I Retrieval mit dem Vektorraummodell

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

Dokument 1: „traditional french recipe“

Dokument 2: „french dinner“

Dokument 3: „traditional christmas dinner“

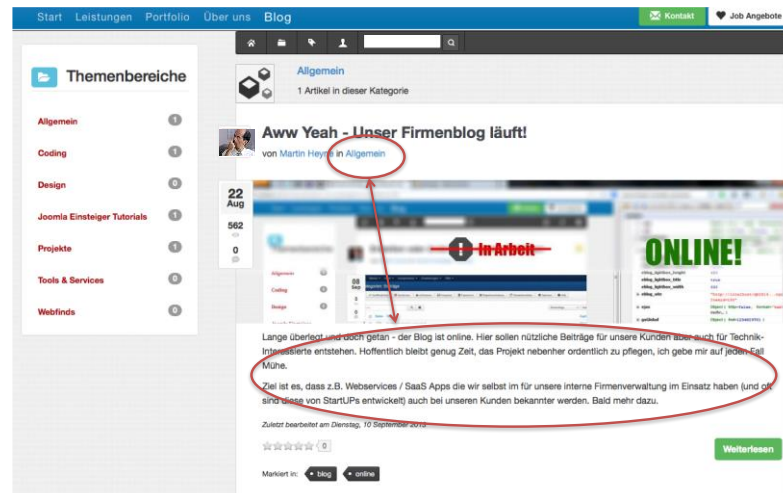
Die Suchanfrage q an das Retrievalsystem lautet „traditional french christmas recipe“.

- Erstellen Sie die Term-Dokument-Matrix mit einer **tf-idf-Gewichtung** von $(1 + \log_{10}(tf_{d,f})) * \log_{10}(N/df_i)$.
- 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 (Aufgabe für 2 Personen)

Sie arbeiten in einen Betrieb in dessen Firmenblog die neusten Posts Kategorien zugeordnet werden. Das geht aber leider immer schief, weil ihr Chef, der die Beiträge schreibt, nie die passende Kategorie vergibt. Sie müssen jedes Mal nachkorrigieren. Ihr Chef erinnert sich an die gute Arbeit, die Sie bei der neuen Firmensuche gemacht haben. „Da hatten Sie doch auch dieses Vektor-Dinges eingesetzt... Wäre das nicht was für unser Problem?“. Aber sicher!

Sie können diese Aufgabe gerne zu zweit vorbereiten und präsentieren. Ansonsten halt zwei Einzelpersonen, die dann entweder den ersten oder zweiten Teil übernehmen.



Machen Sie Ihren Chef glücklich! 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.

| | 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 (?) | ? | ? | ? | ? |

- 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?

III Zipfs Gesetz in Shakespeares Gesamtwerk

Stück für Stück wächst unsere Suchmaschine. Aber vorher steigen wir noch etwas tiefer in das Thema Teststatistik ein.

- Erstellen Sie eine Teststatistik über alle Terme in Shakespeares Gesamtwerk, unterscheiden Sie also nicht nach einzelnen Dateien. Sortieren Sie alle Terme nach Häufigkeit.
- Geben Sie die Terme und Termfrequenzen als CSV-Datei aus.
- Visualisieren Sie die Term-Häufigkeiten als Graph (logarithmieren Sie beide Achsen). Nutzen Sie z.B. Excel oder eine andere Software Ihrer Wahl. Sehen Sie eine Gerade?
- Gilt für dieses Korpus Zipfs Gesetz (für $C=0,1$)? Vergleichen Sie die gemessenen Werte mit den theoretischen.