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

Technology
Arts Sciences
TH Köln

DIS12: Information Retrieval – Online-Sitzung 04.12.2020
Übung 02: TF-IDF

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 Aufgabe 1: Berechnung von idf

Gegeben ist folgendes Vokabular und dazugehörige Dokumente.

Vokabular: {Aubergine, Karotte, Kartoffel, Paprika, Tomate}

Dokumente (Bag of Words):

1. Aubergine, Kartoffel, Tomate
2. Tomate, Tomate, Karotte, Kartoffel
3. Kartoffel, Kartoffel, Aubergine, Aubergine, Aubergine, Karotte, Karotte, Tomate
4. Kartoffel, Kartoffel, Aubergine
5. Paprika, Paprika, Tomate

Bearbeiten Sie die folgenden Aufgaben:

a. Erstellen Sie eine Term-Dokument-Matrix, die die Termhäufigkeiten enthält.
b. Bestimmen Sie df_t für jeden Term.
c. Berechnen Sie die idf-Gewichte für alle Begriffe des Vokabulars und nutzen Sie hierfür die folgende idf-Formel: $\mathbf{\log_{10}(N/df_t)}$

II Ein Gefühl für idf bekommen

Sie haben eine Dokumentenkollektion in der 1 Millionen Dokumente enthalten sind (N=1.000.000).

a. Vervollständigen Sie die folgende Tabelle und nutzen Sie die o.g. Formel aus Aufgabe 1c zur Berechnung von idf. Welches Muster fällt auf? Welcher Wert ist als besonders herauszustellen?
b. Verständnisfrage: Wie viele idf-Werte gibt es für jeden Term t in der Kollektion?

Term	**df_t**	**idf_t**
Müller-Lüdenscheidt	1	
Tier	100	
Sonntag	1.000	
Mops	10.000	
unter	100.000	
der	1.000.000	

III Berechnung von tf-idf

Stellen Sie sich folgende Termfrequenzen für die Dokumente 1, 2 und 3 vor:

Term	**Dokument 1**	**Dokument 2**	**Dokument 3**
Müller-Lüdenscheidt	18	0	0
Tier	8	24	0
Sonntag	0	23	21
Mops	9	0	13
unter	0	1	0
der	0	189	0

Berechnen Sie die tf-idf-Werte indem Sie die errechneten idf-Werte aus der vorherigen Aufgabe verwenden. Tipp: Wenn Sie sich das Leben einfach machen wollen, verwenden Sie Excel für die Berechnung.

ACHTUNG: Es gibt verschiedene Arten, wie man tf-idf berechnen kann. Benutzen Sie in dieser Aufgabe zur Berechnung die folgende Formel: $\mathbf{(1 + \log_{10}(tf_{d,f}))} * \mathbf{\log_{10}(N/df_t)}$

IV Verständnisfragen zu idf und Zipfs Gesetz

a. Ab welcher Anzahl von Suchtermen hat idf einen Effekt auf das Ranking? Können Sie Ihre Antwort begründen? Haben Sie ein Beispiel, dass Ihre Antwort illustriert?
b. Rufen Sie sich die Formel zu Zipfs Gesetz aus der Vorlesung wieder in Erinnerung:

$$P_t = \frac{c}{r_t}$$

Welchen Anteil der aufgetretenen Terme würde man aus der Textkollektion entfernen, wenn wir jedes Auftreten der fünf häufigsten Terme aus der Textkollektion entfernen würden?

V tf-idf in Shakespeares Gesamtwerk

Für diese Aufgabe greifen wir wieder auf den Shakespeare-Korpus zurück, den Sie im Moodle finden. In der Vorlesung habe ich einige Term-Dokument-Matrix-Beispiele mit Zahlen aus dem Korpus verwendet. Allerdings sind die Zahlen wohl offensichtlich falsch (da nur erfunden!).

a. Erstellen Sie ein kleines Skript (z.B. in Python), das Ihnen die tf, df, und tf-idf-Werte für einen beliebigen Term ermittelt
b. Komplettieren Sie hiermit die Tabellen, sowohl für die binäre, Häufigkeits- und tf-idf-Variante.

Nutzen Sie zur Berechnung von tf-idf die Formel aus Aufgabe III. Ermitteln Sie erst einmal nur die Werte für die Terme, die auch in der Vorlesung genannt wurden.