Information Retrieval

Aufgabe 1

Gegeben sei ein Korpus von 1000 Dokumenten d1 bis d1000. Zu einer Anfrage liefert ein IR-System die folgende Rangliste zurück:

1. d877
2. d432
3. d558
4. d121
5. d47
6. d932
7. d111
8. d865
9. d99
10. d56

Ein menschlicher Experte würde für dieselbe Anfrage die folgende Rangliste relevanter Dokumente (aus dem gesamten Corpus) aufstellen:

1. d558
2. d633
3. d47
4. d955
5. d877
6. d111

$$
\begin{aligned}
Prec@5 &= \frac{\text{Anzahl relevanter Dokumente in den Top5}}{5} \\
&= \frac{3}{5}
\end{aligned}
$$

$$
\begin{aligned}
Prec@10 &= \frac{\text{Anzahl relevanter Dokumente in den Top10}}{10} \\
&= \frac{4}{10} = \frac{2}{5}
\end{aligned}
$$

$$
\begin{aligned}
Rec@10 &= \frac{\text{Anzahl der gefundenen und relevanten Dokumente}}{\text{Anzahl der relevanten Dokumente}} \\
&= \frac{4}{6} = \frac{2}{3}
\end{aligned}
$$

$$
\begin{aligned}
F &= \frac{1}{\alpha\frac{1}{Prec@10} + (1-\alpha)\frac{1}{Rec@10}} \\
&= \frac{1}{\alpha\frac{1}{\frac{4}{10}} + (1-\alpha)\frac{1}{\frac{4}{6}}} \\
&= \frac{4}{10\alpha + 6 - 6\alpha} \\
&= \frac{4}{4\alpha + 6} \\
&= \frac{2}{2\alpha + 3}
\end{aligned}
$$

Mit R_k bezeichnen wir die Ergebnisse der Suche bis zum relevanten Treffer k. Damit ergibt sich für die uninterpolierte Durchschnittspräzision UMP folgendes.

$$
\begin{aligned}
UMP@10 &= \frac{1}{6}\sum_{k=1}^{6} Pre(R_k) \\
&= \frac{1}{6}(2/3 + 0 + 3/5 + 0 + 1 + 4/7) \\
&= 298/630 \approx 0,473
\end{aligned}
$$

Aufgabe 2

Um die Aufgabe zu lösen, konstruieren wir ein Gegenbeispiel, bei dem das Hinzufügen eienr Anfrage mit identischem Resultat für beide Systeme die Reihenfolge der beiden Gütewerte vertauscht.

Für eine Anfragemenge $\{q_1, \ldots, q_n\}$ ist die

$$
\text{Mikrobewertung} = \frac{\sum_{i=1}^{n} \text{Anzahl der für } q_i \text{ relevanten und gefundenen Dokumente}}{\sum_{i=1}^{n} \text{Anzahl aller für } q_i \text{ gefundenen Dokumente}}.
$$

Seien IR_1 und IR_2 zwei IR-Systeme und $\{q_1, \ldots, q_n\}$ die Anfragemenge. Die Summe der relevanten und gefundenen Dokumente von IR_1 sei gleich 10 und die von IR_2 gleich 1, sowie die Summe der gefundenen Dokumente von IR_1 gleich 15 und die von IR_2 gleich 2. Dann

$$
\text{Mikrobewertung}_{IR_1} = 10/15 = 2/3
$$

$$
\text{Mikrobewertung}_{IR_2} = 1/2
$$

Nun nehmen wir eine Anfrage q_{n+1} hinzu, bei der für beide IR-Systeme die Anzahl der relevanten und gefundenen Dokumente 10 und die Anzahl der gefundenen Dokumente ebenfalls 10 ist, so gilt für die Mikrobewertungen der Anfragemenge $\{q_1, \ldots, q_{n+1}\}$

$$
\text{Mikrobewertung}_{IR_1} = 20/25 = 4/5 = 0,8
$$

$$
\text{Mikrobewertung}_{IR_2} = 11/12 \approx 0,917 > 0,8
$$

Aufgabe 3

Zeigen Sie, dass für L2-normierte Vektoren (d.h. Vektoren der Länge 1) die Cosinus-Ähnlichkeit und die Euklidische Distanz im Vektorraummodell dasselbe Ranking der Treffer ergeben.

$$
\begin{aligned}
\|x - y\|^2 &= \sum_i (x_i - y_i)^2 \\
&= \sum_i x_i^2 - 2\sum_i x_i y_i + \sum_i y_i^2 \\
&= 2 - 2\sum_i x_i y_i \\
&= 2 - 2\cos(x, y)
\end{aligned}
$$

Damit erhalten wir

$$2 = \|x - y\|^2 + 2\cos(x, y)$$

Je ähnlicher zwei Vektoren sind, desto kleiner ist $\|x - y\|$ und desto größer ist $\cos(x, y)$. Nimmt $\cos(x, y)$ in obiger Gleichung größere Werte an, wird die Norm kleiner und umgekehrt.

Aufgabe 4

Nach Stoppworteliminierung und Stammformreduktion erhalten wir folgende Sätze

d1 Marcus try assasin Caesar.

d2 Marcus Rome.

d3 Caesar rule. Rome loyal Caesar hate.

d4 Loyal. People try assasin rule loyal.

und bestimmen diese Werte:

- $idf_i = \frac{N}{df_i}$
- $tf_{ij,normalisiert} = \frac{tf_{ij}}{\max_k tf_{kj}}$
- $idf_{i,gedämpft} = \log_2 \frac{N}{df_i}$
- $w_{ij} = tf_{ij,normalisiert} \cdot idf_{i,gedämpft}$

Terme	df_i	idf_i		absolute tf_{ij}				normalisierte tf_{ij}									
		absolut	gedämpft	tf_{i1}	tf_{i2}	tf_{i3}	tf_{i4}	tf_{i1}	tf_{i2}	tf_{i3}	tf_{i4}	w_{i1}	w_{i2}	w_{i3}	w_{i4}	q_1	q_2
Marcus	2	2	1	1	1	0	0	1	1	0	0	1	1	0	0	0	0
try	2	2	1	1	0	0	1	1	0	0	0,5	1	0	0	0,5	0	0
assasin	2	2	1	1	0	0	1	1	0	0	0,5	1	0	0	0,5	1	1
Caesar	2	2	1	1	0	2	0	1	0	1	0	1	0	1	0	1	0
Rome	2	2	1	0	1	1	0	0	1	0,5	0	0	1	0,5	0	0	0
rule	2	2	1	0	0	1	1	0	0	0,5	0,5	0	0	0,5	0,5	0	0
loyal	2	2	1	0	0	1	2	0	0	0,5	1	0	0	0,5	1	0	1
hate	1	4	2	0	0	1	0	0	0	0,5	0	0	0	1	0	0	0
people	1	4	2	0	0	0	1	0	0	0	0,5	0	0	0	1	0	0

Die Cosinus-Ähnlichkeit errechnet sich über

$$sim(d, q) = \frac{\sum_i d_i q_i}{\|d\| \cdot \|q\|}$$

$$sim(d_1, q_1) = \frac{1}{\sqrt{2}} \approx 71\% \qquad sim(d_1, q_2) = \frac{1}{2\sqrt{2}} \approx 35\%$$
$$sim(d_2, q_1) = 0 \qquad sim(d_2, q_2) = 0$$
$$sim(d_3, q_1) = \frac{2}{\sqrt{22}} \approx 43\% \qquad sim(d_3, q_2) = \frac{1}{\sqrt{22}} \approx 21\%$$
$$sim(d_4, q_1) = \frac{1}{\sqrt{22}} \approx 21\% \qquad sim(d_4, q_2) = \frac{3}{\sqrt{22}} \approx 64\%$$

Somit erhalten wir als Trefferlisten für q_1 und q_2:

q_1**:** d1, d3, d4

q_2**:** d4, d1, d3