

Web-Informationssysteme, WS 2009/10
Übungsblatt 3

Besprechung am Di 10.11.2009

Aufgabe 3-1 Vector Space Model, TF_IDF

Gegeben eine Dokumentensammlung bestehend aus den folgenden 5 Dokumenten:

D_1 : "If it walks like a duck and quacks like a duck, it must be a duck."

D_2 : "Beijing Duck is mostly prized for the thin, crispy duck skin with authentic versions of the dish serving mostly the skin."

D_3 : "Bugs' ascension to stardom also prompted the Warner animators to recast Daffy Duck as the rabbit's rival, intensely jealous and determined to steal back the spotlight while Bugs remained indifferent to the duck's jealousy, or used it to his advantage. This turned out to be the recipe for the success of the duo."

D_4 : "6:25 PM 1/7/2007 blog entry: I found this great recipe for Rabbit Braised in Wine on cookingforengineers.com."

D_5 : "Last week Li has shown you how to make the Sechuan duck. Today we'll be making Chinese dumplings (Jiaozi), a popular dish that I had a chance to try last summer in Beijing. There are many recipes for Jiaozi."

Zur Vereinfachung seien nur die folgende Terme relevant für unsere Anwendungsklasse:

$$T \in \{\text{beijing, dish, duck, rabbit, recipe, roast}\}.$$

Sei N = Anzahl der Dokumente in der Dokumentensammlung (hier $N = 5$).

$DF(T)$ = Anzahl der Dokumente, in denen der Term T vorkommt.

$DL(D)$ = Länge des Dokuments D gemessen in Termvorkommen = $\sum_T RTF(D, T)$.

$RTF(D, T)$ = Anzahl der Vorkommen von Term T in Dokument D .

a) Geben Sie jeweils die Dokument-Term-Matrix für die obige Dokumentensammlung an unter den folgenden Maßen für die Term-Frequenz:

1. **Binäre Term-Frequenz:** Eine Zelle m_{ij} der binären Dokument-Term Matrix ist 1 falls in Dokument D_i der Term T_j vorkommt, 0 andernfalls.
2. **Einfache Term-Frequenz:** Eine Zelle m_{ij} der einfachen Dokument-Term Matrix ist die einfache (raw) Term-Frequenz $RTF(D_i, T_j)$, die Anzahl der Vorkommen von T_j in D_i .
3. **Normalisierte Term-Frequenz:** Eine Zelle m_{ij} der normalisierten Dokument-Term Matrix ist $TF(D_i, T_j) = \frac{RTF(D_i, T_j)}{DL(D_i)}$.
4. **TF_IDF:** Eine Zelle m_{ij} der TF_IDF Dokument-Term Matrix ist $TF(D_i, T_j) \cdot IDF(T_j)$, wobei $IDF(T_j) = \log \frac{N}{DF(T_j)}$ die logarithmisch gedämpfte *inverse document frequency* sei.

- b) Für die Anfrage $Q = \text{"Beijing duck recipe"}$, bestimmen Sie zu jedem der Maße RTF, TF, TF_IDF die beiden höchst-gerankten Dokumente und kommentieren Sie, wie relevant diese für die Anfrage sind. Verwenden Sie dabei die Cosinus-Ähnlichkeit mit $Q = (q_1, \dots, q_n)^T$:

$$\text{sim}(Q, D_i) = \frac{Q \cdot D_i}{\|Q\|_2 \cdot \|D_i\|_2} = \frac{\sum_j q_j m_{ij}}{\sqrt{\sum_j q_j^2} \sqrt{\sum_j m_{ij}^2}}$$

Aufgabe 3-2 Vector Space Model, TF_IDF

Gegeben eine Dokumentensammlung mit 20.000 Dokumenten, so dass:

- „Maximilian“ kommt in 300 Dokumenten vor.
- „Ludwig“ kommt in 40 Dokumenten vor.
- „University“ kommt in 10.500 Dokumenten vor.
- „in“ kommt in 19.500 Dokumenten vor.
- „Munich“ kommt in 6.000 Dokumenten vor.
- „Germany“ kommt in 10.000 Dokumenten vor.

Nehmen Sie an, daß „in“ als Stopwort behandelt wird.

- a) Berechnen Sie den TF_IDF Vektor für das folgende Dokument. Stopwörter werden entfernt. Die Terme seien alphabetisch geordnet.

D_0 : „Ludwig-Maximilian University Munich, in Munich, Germany“

- b) Ist die *inverse document frequency* (IDF) eines Terms immer endlich? Falls ja, begründen Sie. Falls nein, geben Sie ein Gegenbeispiel!
- c) Welchen Wert nimmt der IDF eines Terms an, wenn dieser Term in *allen* Dokumenten vorkommt?
- d) Wie wirkt sich die Basis des Logarithmus auf die Berechnung des IDF aus?
- e) Vergleichen Sie die Verwendung von TF_IDF mit der Verwendung von Stopwörtern. Lohnt es sich, die beiden Techniken zu kombinieren?

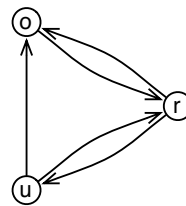
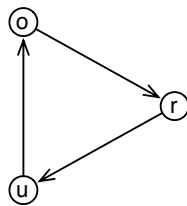
Aufgabe 3-3 Web-Graphen

Zum Zweck der Link-Analyse abstrahieren wir die (Hyper-)Link-Struktur des Web zu einem gerichteten Graphen (bzw. seiner Adjazenzmatrix).

- a) Zeichnen Sie einen gerichteten Graphen mit der folgenden Adjazenzmatrix:

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

- b) Interpretieren Sie die beiden folgenden gerichteten Graphen als Web-Graphen und überlegen Sie, wie Sie die (drei) Knoten ranken würden. Begründen Sie ihre Entscheidung.



- c) Geben Sie Adjazenzmatrizen für die beiden Graphen an. Wie viele unterschiedliche Adjazenzmatrizen für jeden der Graphen gibt es?
- d) Nehmen Sie für den zweiten Graphen je 12 Surfer pro Seite an und überlegen Sie, wieviele Surfer unter einem zufälligen Surfer-Modell¹ sich nach einem Schritt auf jeder der Seiten aufhalten. Nach weiteren Schritten?

Angenommen es gibt eine *stabile Konfiguration*, in der die Anzahl der Surfer pro Seite beim nächsten Schritt gleich bleibt. Wieviele Surfer befinden sich in dieser Konfiguration auf welchen Seiten?

¹ Surfer verfolgen zufällig ausgehende Links. „Sprünge“ zu beliebigen (nicht verlinkten) Seiten finden nicht statt.

Aufgabe 3-4 Lineare Algebra: Wiederholung

Im folgenden betrachten wir der Einfachheit halber nur 2×2 -Matrizen.

- a) Sei $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $B = \begin{pmatrix} 4 & 3 \\ 2 & 1 \end{pmatrix}$. Berechnen Sie AB und BA .
- b) Sei $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, $B = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$. Berechnen Sie AB und BA . Was fällt auf?
- c) Bestimmen Sie die inversen Matrizen zu $A = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}$ und $B = \begin{pmatrix} 1 & b \\ 1 & b \end{pmatrix}$.
- d) Sei $A^2 = AA$, $A^3 = AAA$ etc. Berechnen Sie A^2 , A^3 , A^4 , A^{100} für $A = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}$.
- e) Zeigen Sie: $A = \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}$ kommutiert mit jeder Diagonalmatrix.
Für $a \neq d$ kommutiert A nur mit Diagonalmatrizen.
- f) Zeigen Sie: $(AB)^T = (B^T)(A^T)$.