17.09.4

a) $N(t,c) = N_{11}$

Term kommt in Klasse c vor

N_{00} ... ted, dec

N_{10} ... ted, dec

N_{01} ... ted, dec

N_{11} ... ted, dec

$N_{et,ec}$ ↑ Term in Doku ↑ Doku in Klasse

$N_{\bullet\bullet}$ ... absolute Werte

$N(\text{brazil}, c) = 51$ *

$N(\text{roasted}, c) = 10$

$N(\text{producer}, c) = 34$ *

Für Feature selection beide (*) auswählen

$$I(U;C) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} P(U=e_t, C=e_c) \cdot \log_2 \frac{P(U=e_t, C=e_c)}{P(U=e_t) \cdot P(C=e_c)}$$

↑ in Bit

$$I(U;C) = \frac{N_{11}}{N} \log_2 \frac{N_{11}}{N} \Big/ \left(\frac{N_{1\bullet}}{N} \cdot \frac{N_{\bullet 1}}{N}\right)$$

$N_{11}+N_{10}$ $N_{01}+N_{11}$

Wtk Term in Doku und Doku in Klasse

$$+ \frac{N_{01}}{N} \log_2 \frac{N_{01}}{N} \Big/ \left(\frac{N_{0\bullet}}{N} \cdot \frac{N_{\bullet 1}}{N}\right)$$

$$+ \frac{N_{10}}{N} \log_2 \frac{N_{10}}{N} \Big/ \left(\frac{N_{1\bullet}}{N} \cdot \frac{N_{\bullet 0}}{N}\right)$$

$$+ \frac{N_{00}}{N} \log_2 \frac{N_{00}}{N} \Big/ \left(\frac{N_{0\bullet}}{N} \cdot \frac{N_{\bullet 0}}{N}\right)$$

$\bullet \Rightarrow 1+0$

$N_{0\bullet} = N_{00}+N_{01}$

$N_{\bullet 0} = N_{10}+N_{00}$

$N_{1\bullet} = N_{11}+N_{10}$

$N_{\bullet 1} = N_{01}+N_{11}$

$$= \frac{N_{11}}{N} \log_2 \frac{N_{11} \cdot N}{N_{1\bullet} \cdot N_{\bullet 1}} + \frac{N_{10}}{N} \log_2 \frac{N_{10} \cdot N}{N_{1\bullet} \cdot N_{\bullet 0}}$$

$$+ \frac{N_{01}}{N} \log_2 \frac{N_{01} \cdot N}{N_{0\bullet} \cdot N_{\bullet 1}} + \frac{N_{00}}{N} \log_2 \frac{N_{00} \cdot N}{N_{0\bullet} \cdot N_{\bullet 0}}$$

$= 0,00074 + 0,000459 + 0,000612 + 0,000406$

$= 0,00104$

$\Rightarrow I(U;C) = 0,00104$

$\Rightarrow$ producer, brazil

$N_{1\bullet} = 153$

$N_{\bullet 1} = 1184 = 782$

$N_{0\bullet} = 98847$

$N_{\bullet 0} = 98848$

15.2

$\hat{P}(t='the'|c) = 0{,}05$ Termfrequenz tf 'the' = Dokumentfrequenz df 'the' =

$\hat{P}(t='the'|c) = 7{,}00$

15.3

a) China $\overline{\text{China}}$

NB: $\gamma_{NB}(d) = \underset{c_j \in C}{\text{argmax}}\ P(c_j|d)$

MNB: $P(c_j|d) \propto P(c_j) \prod_{t \in d} P(t|c_j)^{tf_d(t)}$

(∝ = proportional; disjunkte Terme = nicht Token)

$P(c_j) = \frac{g_j}{|D|}$ (absolute # Doku in Klasse drin)

$P(\text{China}) = \frac{2}{4} = \frac{1}{2}$

$P(\overline{\text{China}}) = \frac{2}{4} = \frac{1}{2}$

$\hat{P}(t|c_j) = \frac{tf_{c_j}(t) + 1}{\sum_{t_i \in V} tf_{c_j}(t_i) + |V|}$ → kleinste WSK — Smoothing

$|V| = 7$

Anteil der tf in gesamte Token mit $P(c_j)$

$\hat{P}(\text{Taiwan}|\text{China}) = \frac{2+1}{5+7} = \frac{1}{4}$ (Taiwan zwei Mal in Doku c = China)

$\hat{P}(\text{Sapporo}|\text{China}) = \frac{0+1}{5+7} = \frac{1}{12}$ (Gesamte Token für c = China)

$\hat{P}(\text{Taiwan}|\overline{\text{China}}) = \frac{1+1}{5+7} = \frac{1}{6}$

$\hat{P}(\text{Sapporo}|\overline{\text{China}}) = \frac{2+1}{5+7} = \frac{1}{4}$ (tf in d_5; tf in d_5)

$P(\text{China}|d_5) = \hat{P}(\text{China}) \cdot P(\text{Taiwan}|\text{China})^2 \cdot P(\text{Sapporo}|\text{China})^1$

$= \frac{1}{2} \cdot \left(\frac{1}{4}\right)^2 \cdot \left(\frac{1}{12}\right) = \frac{1}{384}$

$P(\overline{\text{China}}|d_5) = \hat{P}(\overline{\text{China}}) \cdot P(\text{Taiwan}|\overline{\text{China}})^2 \cdot P(\text{Sapporo}|\overline{\text{China}})^1$

$= \frac{1}{2} \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{4}\right) = \frac{1}{288}$

$P(\overline{\text{China}}|d_5) > P(\text{China}|d_5) \Rightarrow d_5 = \overline{\text{China}}$

b) BNB:

$P(c_j|d) \propto P(c_j) \cdot \prod_{t \in d} P(t|c_j) \cdot \prod_{t_k \in V \setminus d} P(\overline{t_k}|c_j)$ → Gegenwsk. dass tk nicht vorkommen

b) $\vec{x}_{k,i} \geq \frac{\alpha}{n}$ $\quad \sum_{i=1}^{n} \vec{x}_{k,i} \geq \alpha$

$\sum_{i=1}^{n} \vec{x}_{k,i} > 1$

$\max d = 1 - \alpha$

maximal abstand

$\lim_{X \to 1} \max d = 0$

NB: $\frac{\alpha}{n} + 1 - \alpha$ hier: Ausgangspkt = 1

$\alpha = 1$

$$P = \begin{array}{c|ccc} & A & B & C \\ A & 1/3 & 1/3 & 1/3 \\ B & 1/3 & 1/3 & 1/3 \\ C & 1/3 & 1/3 & 1/3 \end{array}$$

Trick: bei 3 Fälle
$\frac{1}{n}$ $\quad \alpha=1: \frac{\alpha}{n}$ $\quad \alpha=1: \frac{\alpha}{n}$

14.3

$\alpha = 0,1 = \frac{1}{10}$ $\quad n = 3$

Page Rank

$$P = \begin{pmatrix} 1/30 & 29/60 & 29/60 \\ 1/30 & 1/30 & 14/15 \\ 1/30 & 14/15 & 1/30 \end{pmatrix}$$

$\frac{\frac{1}{10}}{3}$

$\frac{\frac{1}{10}}{3} + \frac{1-\frac{1}{10}}{2}$

$\frac{\frac{1}{10}}{3} + \frac{1-\frac{1}{10}}{1}$

$\vec{X_k} = P^T \cdot \vec{X_{k-1}}$ $\quad \vec{X_0} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}$ gleich wahrscheinlich

bis Konvergenz

$\vec{X_k} = \begin{pmatrix} \frac{1}{30} \\ \frac{29}{60} \\ \frac{29}{60} \end{pmatrix}$ (A, B, C) — wie kommen random Server zu A?
durch Transportation $1. (i,j) \in E$ anwenden
$\Rightarrow \frac{\frac{1}{10}}{3} = \frac{1}{30}$

Wsk bei B, C bleibt gleich $\Rightarrow (1 - \frac{1}{30})/2 = \frac{29}{60}$

HITs: $A(i) = \sum_{j \in V / \{i\}} H(j)$ ohne $\{i\}$

$H(i) = \sum_{j \in V / \{i\}} A(j)$ ohne $\{i\}$ ohne selbst

Adjazent-matrix $M = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$ $\quad \vec{A_k} = \vec{X_k} = M^T M \cdot \vec{X_{k-1}}$ $\quad \vec{X_0} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}$ gleichwahrscheinlich

$$M^T M = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

$\vec{X_k} \Rightarrow \frac{1}{|\vec{X_k}|} \vec{X_k}$ theoretisch: normalisierung!

aber hier braucht man die Normalisierung nicht
Aufgabestellung beachten!

$q_m = [0,5;\ 2;\ 2,5;\ 0;\ 1,25]$

10.3

a) $q_m = \alpha q_0 + \beta l_r + \gamma l_{nr}$

$q_0 = \alpha q_0 + \beta l_r + \gamma l_{nr} \mid -q_0$

$0 = (\alpha - 1) q_0 + \beta l_r + \gamma l_{nr}$

Wenn $\alpha = 1$:

$0 = \beta \cdot l_r + \gamma l_{nr}$

$\gamma \cdot l_{nr} = \beta \cdot l_r$

$l_{nr} = \frac{\beta}{\gamma} l_r$

l_{nr} und l_r linear abhängig

b) Gegenbeispiel $q_0 = l_r$

$q_m = \alpha \cdot l_r + \beta l_r - l_{nr}$

10.4

$\alpha = 0,\ \beta = 1;\ \gamma = 0$

10.5

a) → Precision → statistische Centroide → Lernbeschreibung

b) Weniger Information bei several wahrscheinlich

7.2 $|D| = 1$ Gewicht von Term $t_1 \in q$ konstant $\Rightarrow$ unabhängig

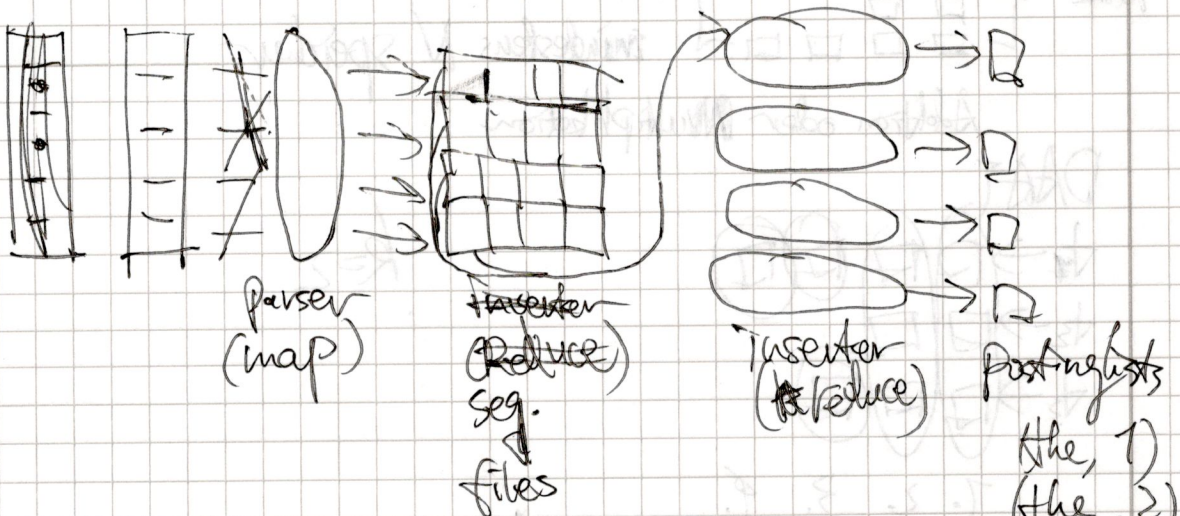
aber: Gewicht von $t_1 \in D$ $D = \{d_1, \dots d_N\}$

$\forall t_i, d_j \Rightarrow$ cos-normalisiert

$t_2 \rightarrow$ $W_{t_2,d_8} > W_{t_1,d_7} > W_{t_4,d_3}$

$\forall B \; k = 2$

7.3



map: $C \rightarrow$ list (term, docID)

reduce: list (<term, list(docID)>)

$\rightarrow$ posting list 1, posting list 2, ...

map: $C \rightarrow$ list (term, 1) ($\rightarrow$ Tupel)

bei mehrfache Wörter kommt auch mehrfache Tupel

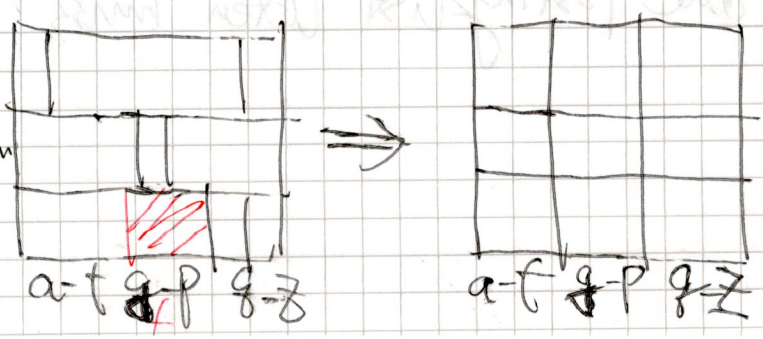
reduce = list (word, (ist(sum)) $\rightarrow$ (word, sum)

7.4 200GB > 100GB

~~Partitionierung~~ Partitionierung nach docIDs

the $\rightarrow$ (100GB) in V_1 (100GB) in V_2

7.5 Im Schlimmsten Fall



Before MapReduce Sampling machen a-priori