12.03.2012

$D = [\ ]_{m \times n}$ → $C = [\ ]_{m \times m}$ for doc.

→ $C' = [\ ]_{n \times n}$ for term

m

$$n_c = \sum_{i=1} c_{ii} = \sum \underbrace{\delta_i}_{\text{decoupling coef.}}$$

$$n'_c = n_i = \sum_{i=1}^{r} c_{ii}$$

D contains unique documents

documents containing no common terms

$$C = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad n_c = m$$

D contains identical documents

$$C = \begin{bmatrix} 1/m \end{bmatrix} \quad n_c = 1$$

C^3M

1. Find n_c
2. Select cluster seeds (initiators)
3. Assign no seeds to cluster seeds.

terms

$$D = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$ documents

$$C = \begin{bmatrix} 0.361 & 0.250 & 0.194 & 0.111 & 0.083 \\ 0.188 & 0.563 & 0.063 & 0.000 & 0.188 \\ 0.194 & 0.083 & 0.361 & 0.278 & 0.083 \\ 0.117 & 0.000 & 0.417 & 0.417 & 0.050 \\ 0.125 & 0.375 & 0.125 & 0.000 & 0.175 \end{bmatrix}$$ $m \times m$

$m + (m - n_c) \times n_c$

calculate this many elements of C-matrix

# of non-seeds

# of seeds

$n_c \ll m$

C^3M (revised)

1. Find $n_c = \sum \delta_i = \sum c_i \approx 2$

2. Select cluster seeds

$P_i = \delta_i \times \Psi_i \times X_{d_i}$

$= c_{ii} \times (1 - c_{ii})$

depth of indexing for d_i. # of elem. in d_i.

P_i = seed power of d_i

Calculate cluster seed power of each document

Choose n_c of them as cluster seeds.

3. Assign non-seeds to cluster seeds

Assign a non-seed to the seed documents which provides the max. coverage.

$\max(c_{ij})$ where $d_j \in$ cluster seed

How to identify identical seeds

$$P_i \approx P_{ji}$$

$$P_1 = \delta_i \times \Psi_1 \times X_{di}$$

$$= C_{11} \times (1 - C_{11}) \times (\text{row sum})$$

$$= 0.361 \times (0.361) \times 3 = 0.692$$

$P_2 = 0.984$

$P_3 = 0.692$

$P_4 = 0.484$

$P_5 = 0.469$

sorted

P_2, P_3, P_1, P_4, P_5

choose

$d_1 \neq d_3$

$C_{03} \mid C_{01} \mid C_{11} \mid C_{13}$

d_2 is a cluster seed

↳ contains t_1, t_2, t_3, t_4

$d_3 \rightarrow t_1, t_3, t_6$

$d_1 \rightarrow t_1, t_4, t_6$

it contains more # of terms which are not used in previously selected seeds. So select d_3 as the 2nd seed.

$$C_{ij} = \alpha_i \sum_{k=1}^{\wedge} d_{ik} \times p_k \times d_{jk}$$

$d_1 \rightarrow \langle t_1, 1\rangle\langle t_2, 2\rangle\langle t_6, 1\rangle$

$d_2 \rightarrow \langle t_1, 1\rangle\langle t_2, 1\rangle\langle t_3, 1\rangle\langle t_4, 1\rangle$

Inverted Index for Seed Documents

$t_1 \rightarrow \langle d_2, 1\rangle\langle d_3, 1\rangle$

$t_2 \rightarrow \langle d_2, 1\rangle$

$t_3 \rightarrow \langle d_2, 1\rangle$

$t_4 \rightarrow \langle d_2, 1\rangle$

$t_5 \rightarrow \langle d_3, 1\rangle$

$t_6 \rightarrow \langle d_3, 1\rangle$

Calculate $C_{12} = 0$ $C_{13} = 0$

t_1

$$C_{12} = C_{12} + \alpha_1 \times (d_{11} \times \beta_1 \times d_{21}) = 1/12$$

$$C_{13} = C_{13} + \alpha_1 \times (d_{11} \times \beta_1 \times d_{31}) = 1/12$$

14.03.2012

$$n_c = \sum c_{ii}$$

$$n_c = \frac{m \times n}{t} \rightarrow$$ # of non-zero elements in the D-matrix

$$D = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$n_c = \frac{2 \times 3}{6} = 1$$

$$n_c = \frac{m \times n}{t} = \frac{n}{x_d} \rightarrow$$ depth of indexing. (avg. number of terms/doc)

$= t/m$

$$n_c = \frac{m \times n}{t} = \frac{m}{t_g} \rightarrow \text{term generality (\# of docs/term)}$$

$$= \frac{t}{n}$$

$$t_g = t/n$$

$$\max(t) = m \times n$$

$$\min(t) = \max(m, n)$$

$$\max(t_g) = \max(t)/n = m \times n/n = m$$

$$\min(t_g) = \min(t)/n = \max(m, n)/n$$

$$= \max(m/n, 1)$$

$$n_c = m/t_g$$

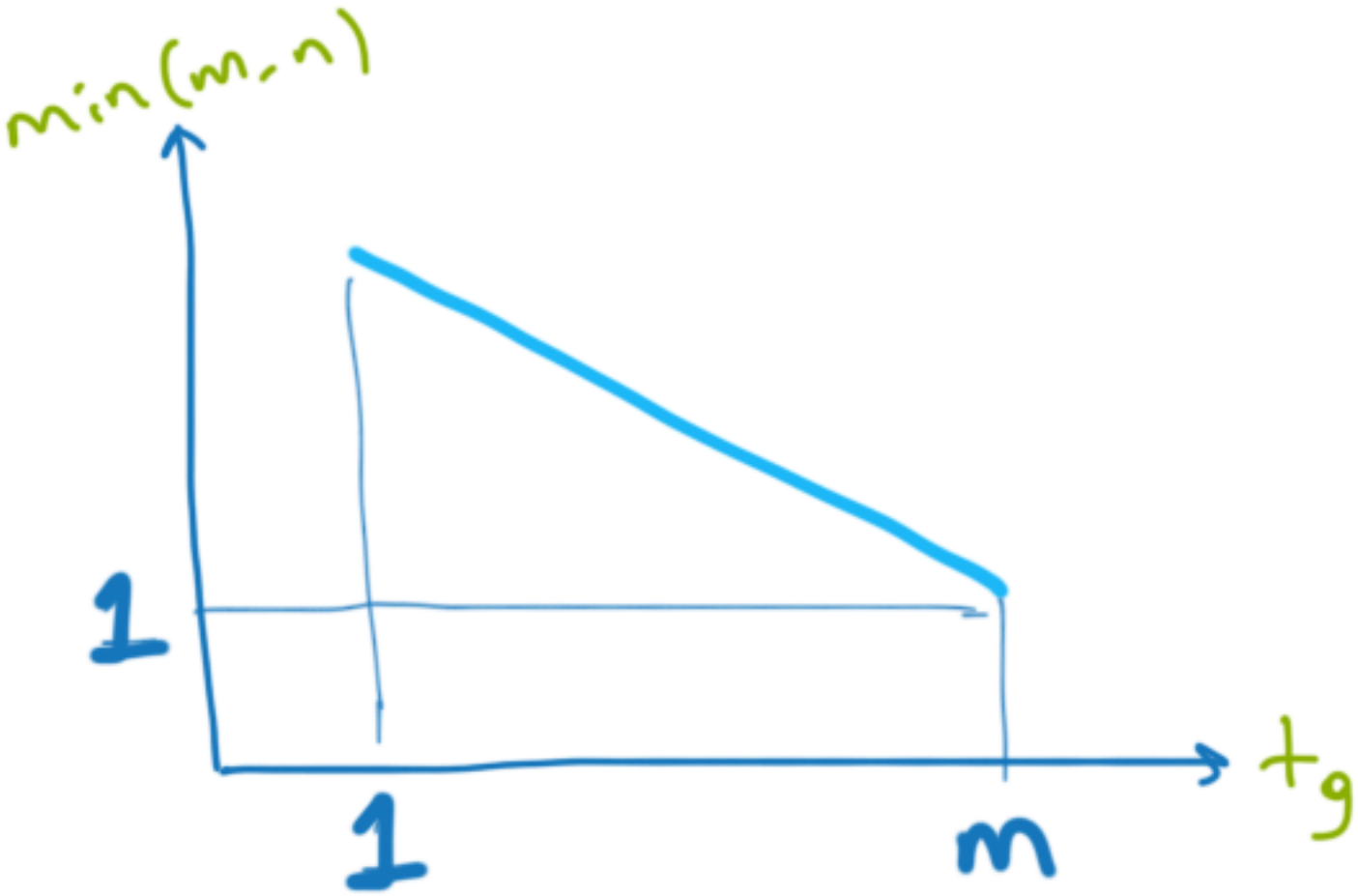
$$\max(n_c) = m/\min(t_g) = m/\max(m/n, 1)$$

$$= 1/\max(1/n, 1/m)$$

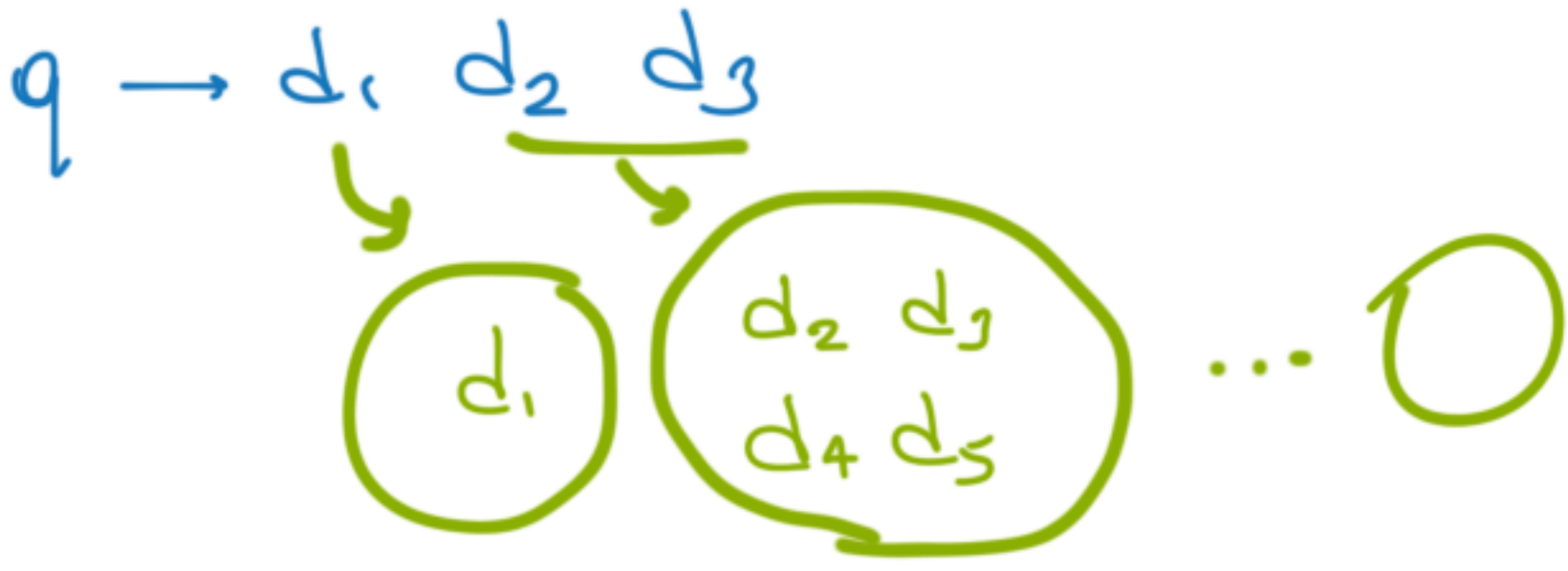
$$= \frac{1}{[\min(m, n)]^{-1}} = \min(m, n)$$

$$\min(n_c) = 1$$

$$\max(n_c) = \min(m, n)$$



Yao's Formula



for q we have 2 tar. clus.

A cluster that contains a relevant document is called target cluster.

n_t = # of target clusters using the clustering structure obtained by our clustering algorithm.

n_{tr} = # of target clusters when documents are randomly distributed among the clusters.

If we have a meaningful clustering structure $n_t < n_{tr}$

↑

significant

n: # of records

m: # of blocks (n_c)

block size: n/m

k: # of records to be accessed

p = n/m # of records in the jth block

n-p: # of records in the other blocks

C_k^n: # of combinations that we can have if we try to select k documents out of n documents.

$$C_k^n = \frac{n!}{k!\,(n-k)!} \qquad C_2^3 = \frac{3!}{2!\,(3-2)!} = 3$$

a, b, c
ab, ac, bc

C_k^{n-p}: different ways of seleeting k documents from (n-p) documents

The probobility that no records are selected from the jth block.

$$\frac{C_k^{n-p}}{C_k^n}$$

Let $d = 1 - 1/m$

$$n-p = n - n/m = n\left(1 - \frac{1}{m}\right) = n \cdot d$$

.

$E(I_j)$: probability of selecting at least a record from the jth block

$$1 - C_k^{nd} / C_k^{n}$$

Expected # of blocks to be accessed

$$\sum_{j=1}^{m} E(I_j) = m \times \left(1 - C_k^{nd} / C_k^{n}\right)$$

proof

$$= m \times \left[1 - \frac{\frac{nd!}{k!\,(nd-k)!}}{\frac{n!}{k!\,(n-k)!}}\right] = m \times \left[1 - \frac{nd!}{k!\,(nd-k)!} \cdot \frac{k!\,(n-k)!}{n!}\right]$$

$$= m \times \left[1 - \frac{1 \cdot 2 \cdot \ldots \cdot nd}{1 \cdot 2 \cdot \ldots \cdot (nd-k)} \cdot \frac{1 \cdot 2 \cdot \ldots \cdot (n-k)}{1 \cdot 2 \cdot \ldots \cdot n}\right]$$

$$\Rightarrow m \cdot \left[1 - \prod_{i=1}^{k} \frac{nd-i+1}{n-i+1}\right]$$

$\frac{(nd-k+1)(nd-k+2)\ldots nd}{1}$

$\frac{1}{(n-k+1)(n-k+2)\ldots n}$

Assume that we have clusters with different sizes

cluster size $|c_j|$ $1 \le j \le n_c$

$m_j = m \cdot |c_j|$

$|c_j|, m_j = 1 - |c_j|$

$$P_j = \left[1 - \prod_{i=1}^{k} \frac{m_j - i + 1}{m - i + 1} \right]$$

↑ # of documents

Example

$k = 3$

$m = 100$ (# of docs)

$|C_1| = 5$

$m_1 = 100 - 5 = 95$

$P_1 = 0.14$

19.03.2012

Cluster hypothesis

should be significant

$n_t < n_{tr}$

↳ random clustering

↳ with clustering algorithm

m: # of docs.

$m_j = m - |C_j|$ $\quad 1 \le j \le n_c$

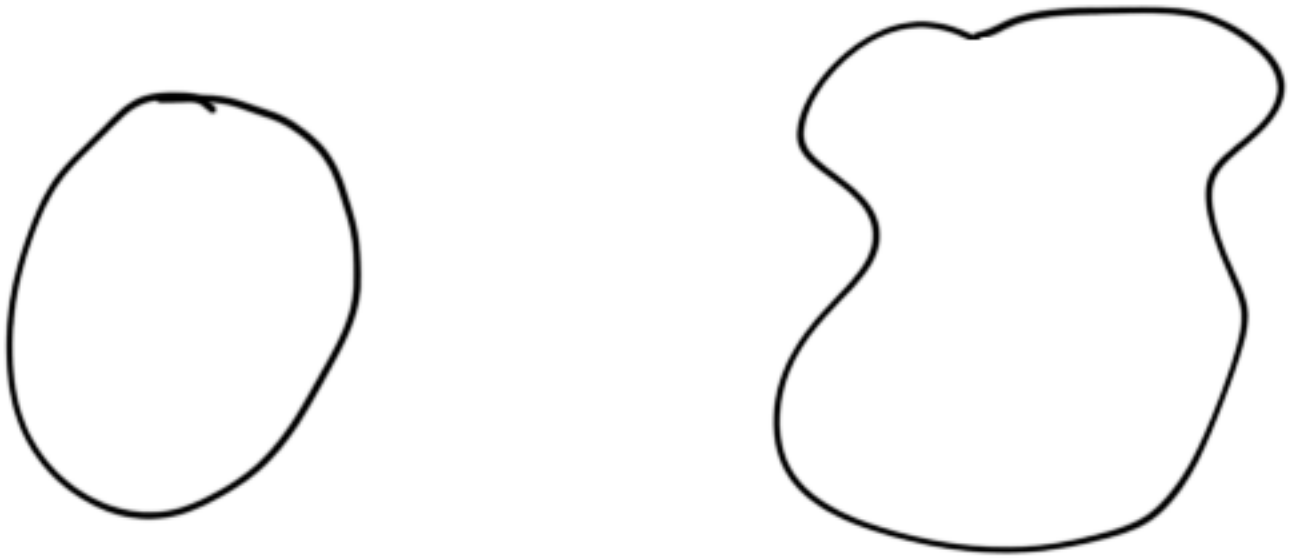
k: # of relevant docs for the query

$$P_j = \left[1 - \prod_{i=1}^{k} \frac{m_j - i + 1}{m - i + 1} \right]$$

$$n_{t_r} = \sum_{j=1}^{n_c} P_j$$

Cluster Validation

Rand Index



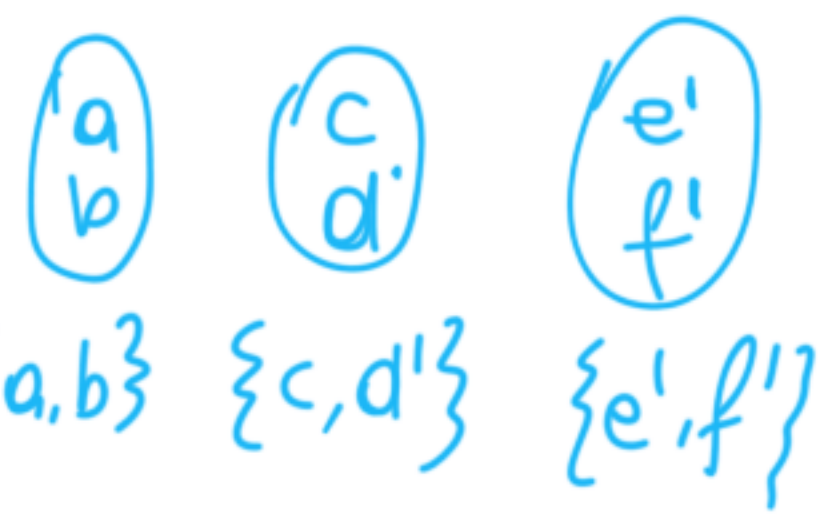
actual clustering structure

$\{a, b, c\}$

$\{d', e', f'\}$

hypothesized clustering structure

generated by our algorithm



$\{a, b\}$ $\{c, d'\}$ $\{e', f'\}$

$$\binom{6}{2} = \frac{6!}{4!\,2!} = 15$$

ab	ac	ad'	ae'	af'	bc	bd'	be'	bf'	cd'	ce'	cf'	de'	d'f'	e'f'
					FN	TN	TN	TN	FP	TN	TN	FN	FN	TP

RI: $\frac{2+8}{15} = 0.66$

Precision $= \frac{2}{3} : 0.66$

Recall $= \frac{2}{6} : 0.33$

$F = \frac{2 \cdot P \cdot R}{P+R} = 0.44$

TP (true positive) two similar documents are assigned to the same h-cluster

TN (true negative) two dissimilar docs. are assigned to two different h-clusters

FP (false positive) two dissimilar docs. are assigned to the same h-cluster

FN (false negative) two similar docs. are assigned to different h-clusters

$$\text{Rand index} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$F = \frac{2PR}{P + R}$$

↓ f-measure

$$\text{recall} = \frac{TP}{TP + FN}$$

Corrected Rand Index: subtract the positive effect of random clusters

Internal Cluster Validation Criterion

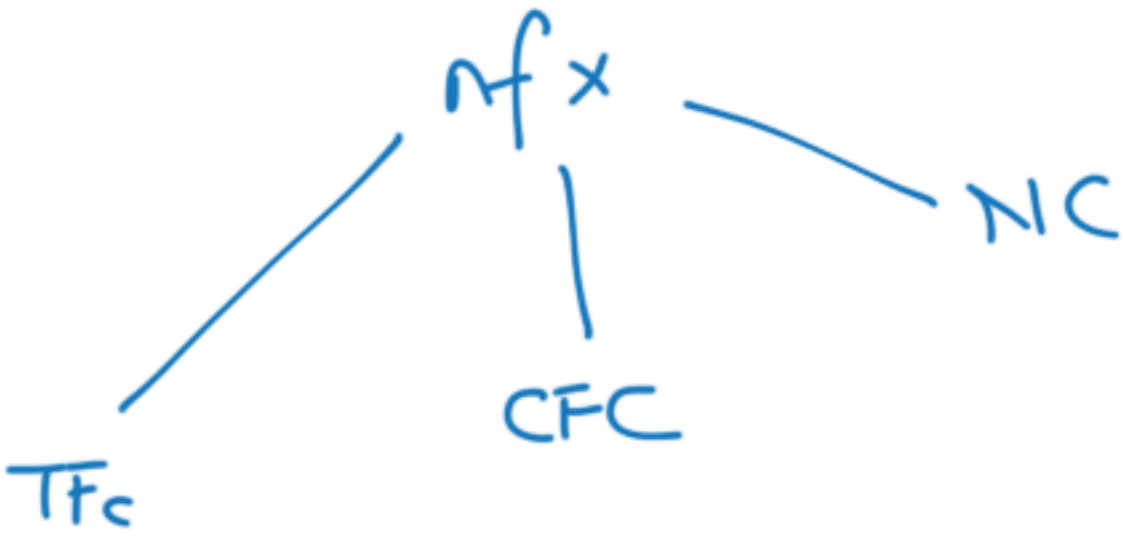
Purity : Each cluster is assigned to the class which is most frequent in the cluster, count the number of correct assignments and divide it by total # of elem.

Example :

26.03.2012

Term weighting (cont.)

Query Weight Normalization



$Q = (1\ 0\ 0\ 2\ 0)$

$t_g\ \ 4\ \ 2\ \ 2\ \ 4\ \ 3$

$TFC = n\ \left(0.5 + 0.5\frac{1}{2}\ \ 0\ \ 0\ \ 0.5\cdot\frac{2}{2}\ \ 0\right)$

$CFC = f\ \ \ln\frac{n}{t_{g_j}} + 1 = (1.22\ \ 1.92\ \ 1.92\ \ 1.22\ \ 1.51)$

nfx: do nothing as normalization

$(0.75 \times 1.22\ \ 0\ \ 0\ \ 1 \times 1.22\ \ 0)$

$(0.92\ \ 0\ \ 0\ \ 1.22\ \ 0)$

$$D = \begin{bmatrix} 0.62 & 0 & 0.49 & 0.62 & 0 \end{bmatrix}$$

$Sim(Q, d_1) = 0.62 * 0.92 + 0 + 0 + 0.62 * 1.22 + 0)$

$= 1.33$ //

$Sim(Q, d_2) = 0.77$

$d_3 = 1.20$

$d_4 = 0.58$

$d_5 = 1.12$

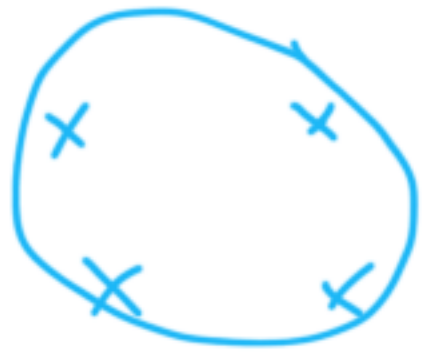
ranking

$d_1 > d_3 > d_5 > d_2 > d_4$

Term Discrimination Value

This concept assumes that terms which make documents more distinguishable from each other should have higher importance.

Example :



After assigning a good discriminator

After assigning a bad term (a term which is not a good indicator)

documents become more similar to each other

How to calculate TDVs

1. Using similarity values
2. " cover coefficient concept